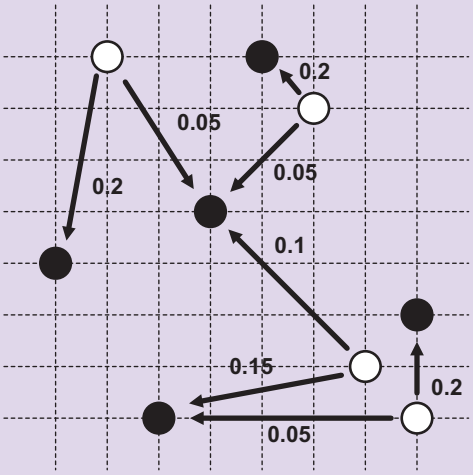
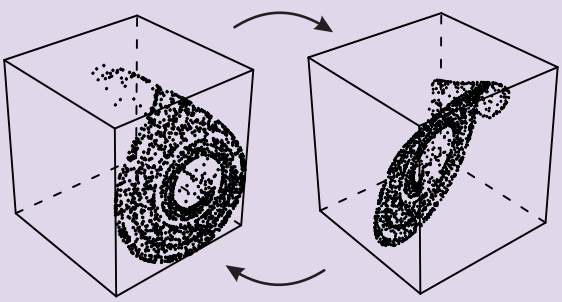
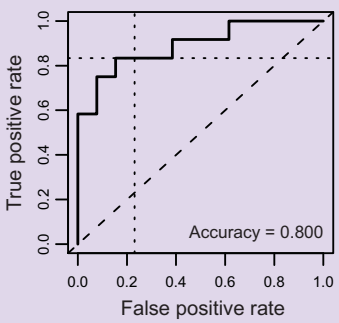


Distance-based analysis of dynamical systems and time series by optimal transport

Michael Muskulus





Distance-based analysis of dynamical systems and time series by optimal transport

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 11 Februari
klokke 11.15 uur

door

Michael Muskulus

geboren te Sorengo, Switzerland
in 1974

Promotiecommissie

Promotor:

prof. dr. S.M. Verduyn Lunel

Overige leden:

dr. S.C. Hille
prof. dr. J.J. Meulman
prof. dr. P.J. Sterk (Academisch Medisch Centrum, Universiteit van Amsterdam)
prof. dr. P. Stevenhagen
prof. dr. S.J. van Strien (University of Warwick)

Distance-based analysis of dynamical systems and time series by optimal transport

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

Muskulus, Michael, 1974–
Distance-based analysis of dynamical systems and time series by optimal transport

AMS 2000 Subj. class. code: 37M25, 37M10, 92C50, 92C55, 62H30
NUR: 919
ISBN: 978-90-5335-254-0

Printed by Ridderprint Offsetdrukkerij B.V., Ridderkerk, The Netherlands

Cover: Michael Muskulus

This work was partially supported by the Netherlands Organization for Scientific Research (NWO) under grant nr. 635.100.006.

After I got my PhD, my mother took great relish in introducing me as, "This is my son. He's a doctor, but not the kind that helps people".

Randy Pausch

Für Frank & Ingrid

And to the most beautiful neuroscientist in the world
Sanne, thank you for our adventures in the past, in the present, and in the future

Contents

List of boxes

Prologue

Of course this limits me to being there in my being only in so far as I think that I am in my thought; just how far I actually think this concerns only myself and if I say it, interests no one.

Jacques Lacan[1]

The scientific endeavour has grown enormously in recent decades, with many new scientific journals being set up to accomodate the continuing flood of research papers. It is a sobering fact that many of these articles are never read at all, or more precisely, are never being cited in a research context (Meho, 2007). Some of the reasons for this development are probably the rising popularity of quantitative evaluations of the research output of scientists, measured in the number of publications produced, by university administrations and research agencies.

In my experience, the majority of journal articles belong to three distinct categories. Depending on whether to locate the subject of the research on either the methodological or the experimental side, many scientific publications describe minor modifications of existing methods or, on the other hand, report empirical findings obtained under minor modifications of experimental protocol. These modifications are necessary to guarantee the status of a truly innovative work or, put differently, to avoid the vice of double publication, but apart from a limited set of experts working on the same subject the value of these findings is often questionable to a broader scientific public. More frightening is the thought that most of these findings might actually be false: Methodological improvements do not always merit the effort to realize them in practice, and empirical findings might be purely coincidental (Ioannidis, 2005). Even in mathematics, where rigorous proofs can be had, and which consequently does not belong under the general heading of science, technical improvements do sometimes not lead to further our understanding of the subject matter (e.g., as in the computer-assisted proof of the four-colour theorem by Appel and Haken). The third major category of publications are of a secondary nature and

[1] On "'cogito ergo sum' ubi cogito, ibi sum" [Where I think 'I think, therefore I am', there I am].

consist of editorials, commentaries, letters and review articles. These do not present new findings, and although of occasional interest to readers most of these, disregarding the latter, never find themselves being cited. Review articles are often highly valued, however, since they not only lend themselves as introductions to a specific research trend, but offer advice and insight that goes beyond mere exposition, and in the best case combine relevant publications that would otherwise go unnoticed in a focussed way.

Regarding the above, it is my opinion that there is much value in reviewing and combining seemingly unrelated fields of research and their tools, which might be one way to define the ubiquitious term *interdisciplinary research*. Indeed, I think that many important methods and ideas do already exist in the large scientific literature, but experience seems to confirm that these are often less well known in other areas of science where they might be favourably used. This transfer of knowledge across boundaries of scientific disciplines is usually not easy. The scientific conservationism exposed by Thomas Kuhn in 1962 seems not to have diminished over time, and it is still difficult to convince scientists from other disciplines about the value of techniques they are not already familiar with. To overcome such scepticism it is necessary to concentrate on essential and proven methods, and to exhibit their advantages (and disadvantages) in as clear a way as possible.

In this thesis I have therefore tried to use well known tools from a variety of distinct (sub-) disciplines in statistics, physics and the theory of dynamical systems to derive new and nontrivial insights in various fields of application, by combining long established and robust ideas in a general framework. To be convincing applications, this involved a lot of time implementing these methods as actually useable computer code, closing quite a few gaps in existing software and collecting the necessary tools in one central place. The text of this thesis follows the same idea of making essentially all of the necessary methods understandable and accessible, even to non-experts. As the reader might imagine, a lot of ideas and results that did not fit into this presentation have been left out (but many of these are discussed cursorily in notes), and, to a mathematician, quite a frightening amount of redundancy might have crept into the text.

I hope the reader will appreciate this effort.

Michael Muskulus
Leiden, January 2010

Chapter 1

General Introduction

Scientific discovery consists in the interpretation for our own convenience of a system of existence which has been made with no eye to our convenience at all. One of the chief duties of a mathematician in acting as an advisor to scientists is to discourage them from expecting too much of mathematicians.

Norbert Wiener

1.1 Distance-based analysis

This section gives a concise overview of the methodology of distance-based analysis and the underlying ideas without undue details.

Systems and measurement devices

The setting is the following: We are given a number of *systems* $(S_1, S_2, \dots)$ that we want to study. Conceptually, we are not concerned with the actual nature of the systems; but we need to specify a way in which to obtain objective (quantitative) information about them. This is achieved by specifying one or more *measuring devices* $(D_1, D_2, \dots)$ that map each system (possibly at a specific point in time or with various other influential factors fixed), arising from a class $\mathcal{S}$ of measurable systems, into a set of numerical measurements: $D_i : \mathcal{S} \to M$. In the simplest case these measurements will be univariate (a single number) or multivariate (a "vector" of numbers). Let us illustrate this with an example. Consider a time series, i.e., a set of numbers $(x_1, x_2, \dots, x_n)$ sampled at n discrete points in time. This can be considered to be a single measurement of a certain system, e.g., it might represent temperature measurements at noon on consecutive days at a certain point in space, and would be naturally represented by the vector $x = (x_1, x_2, \dots, x_n)^t \in \mathbb{R}^n = M$, where the superscript t denotes the transpose. Other measurements might include some kind of processing and could be more involved. For example, the mean $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ of the time series can be considered a single measurement.

Generalized measurements

This framework is very general and allows for much flexibility. With a small generalization we can even accommodate "ideal" (not physically realizable) measurements. For example, we might consider the probability distribution of temperature values, i.e., the function $F(x)$ that tells us how likely it is to find a temperature $T \leq x$ on an otherwise unspecified, and therefore random, day. This distribution could only be obtained by an infinitely long time series (and there might still be problems with its definition without some further stationarity assumptions), but we can always approximate it by an estimate obtained from a finite time series. Whether and when at all such an estimate would be sensible is not the point here, but rather that the class of measurement devices should be allowed to also contain potentially infinite objects, namely, *probability measures*. This is a natural enough extension, since each deterministic measurement $x \in M$ can be identified with a probability measure over M where outcomes other than x have zero probability. Let us denote the class of probability measures defined over a common space M by $P(M)$. We think of these as *generalized measurements* of our systems.

The space of measurements

To compare quantitatively two probability measures from $P(M)$, we will require that the space M already comes with a notion of distance, i.e., is a *metric space*. For example, measurements that result in numerical quantities, such that $M \subseteq \mathbb{R}^n$ for some finite $n \in \mathbb{N}$, are usually metric. Only if M is a proper subset of some $\mathbb{R}^n$ there might be problems, but we are usually allowed to assume that, at least potentially, the range M covers the totality of $\mathbb{R}^n$. Note that the interpretation and usefulness of this metric depends on the measurement apparatus. This expresses the important fact that we only have access to a system through a measurement device D. Properties that are not measured by D can obviously not be restituted later from such measurements. This seems a serious limitation at first, and in a certain way it is: Since it is only possible to analyze measurements, we need to assume that the measurement device captures all the information necessary for the analysis task at hand. It is a fact, however, that this limitation is a *principal* problem that is inherent to all experiments and data analysis and cannot be overcome by *any* method. We are therefore justified, indeed almost compelled, to *identify* the measurements in M with the systems themselves. In practice this is rarely problematic, especially if a lot of data is available (e.g., a long time series recording) that can be assumed to characterize all interesting aspects of a system sufficiently. The main challenge lies in the way we deal with this data and extract meaningful insights from it.

Distances between measurements

In *distance-based analysis* we calculate an abstract distance $d : P(M) \times P(M) \to [0, \infty)$ between each pair of (generalized measurements of) systems. To ensure sensible behaviour in a multivariate setting, we require the function d to be a proper distance: It should be a nonnegative number that is zero between two identical systems, $d(P, P) = 0$ (*reflexivity*). Ideally it should only be zero for two identicals systems, such that $d(P_1, P_2) > 0$ if $P_1 \neq P_2$, which together with the previous property is called *positive definiteness*. This stronger property is not truly needed for most of our applications, but conceptually important. Another property that we require of d is *symmetry*, $d(P_1, P_2) = d(P_2, P_1)$. Finally, the measure d should be minimal in a certain sense. There should be no "short-cuts" possible between two or more systems that result in a shorter distance than the one measured directly between two systems. This property is guaranteed when d fulfills the *triangle inequality*, $d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2)$ for all systems.

Although there are many candidates for such a distance measure, we prefer the class of optimal transportation distances, also called *Wasserstein distances* in the mathematical literature (Villani, 2003). These quantify the amount of "work" that is needed to transform one probability measure P_1 into a second measure P_2. Detailed definitions will be given later; let us note here that the Wasserstein distances are, remarkably, true distances that fulfill all metric properties.

The first two steps in the distance-based analysis are therefore:

1. To measure some properties of systems in the form of a probability measure over some numerical space (usually some Euclidean space $\mathbb{R}^n$).
2. To quantify distances between all pairs of systems under consideration by calculating a Wasserstein distance. This results in a matrix of mutual distances, which will then be analyzed further.

Reconstruction of distances as point configurations

Although there exist quite a few statistical methods to deal with distance matrices, mostly originating from problems in ecology (Legendre and Legendre, 1998), it is very advantageous to represent the distances as actual points in some space with which we are familiar. We will use standard Euclidean space $\mathbb{R}^k$ for this purpose, and furthermore require that $k \ll m$. The latter implies a reduction to a low-dimensional subspace that preserves the most interesting properties of the systems under study. The coordinates of each system in this space $\mathbb{R}^k$ will therefore be derived by the methods of multidimensional scaling (Borg and Groenen, 2005). This is similar to principal component analysis and results in a representation where the first coordinate describes ("explains") the largest variation in the distances, the second coordinate describes the largest remaining variation in the distances, orthogonal to the first, and so on.

Reconstruction in a Euclidean space

Here the reader might wonder about an important issue. It is not obvious that the distances measured can be represented at all in a Euclidean space. For example, geodetic distances (where the shortest line between two points is not necessarily a straight line) are not obvious to realize in a Euclidean space. Phase distributions are the canonical example here: A phase is a number between 0 and 2π where we identify 2π with 0. Topologically, the space of all phases is a circle, and there are always two paths between each pair of distinct phases. The geodetic distance between two phases φ_1 and φ_2 is the shorter value of these, either $|\varphi_1 - \varphi_2|$ (not crossing 0) or $2\pi - |\varphi_1 - \varphi_2|$ (crossing 0). To represent a set of phases as points in a Euclidean space, where distances are measured by straight lines, usually introduces errors, i.e., misrepresentations of the distances. On the other hand, if the dimensionality k of the ambient space is chosen high enough (e.g., $k = N - 1$ for distances between N points) distances can always be represented perfectly by points in an $\mathbb{R}^k$. Using a smaller dimension $k' < k$ will usually introduce errors, but it is hoped that these stay reasonably small for a much smaller value $k' \ll k$.

The reasons we advocate the use of Euclidean space, even though it might not be optimal for certain kinds of distances, are threefold. A technical reason is that the reconstruction of Euclidean coordinates from distances is straightforward, whereas representations in other kinds of spaces are much more involved — and bring their own problems in terms of computational efficiency, convergence of algorithms, and so on. Another advantage of Euclidean space is the great familiarity we have with this kind of representation, which is therefore very useful in data exploration and the search for patterns. On top of this, we argue that it is not necessary to obtain a perfect representation anyway. As in the case of mapping the earth, it is often not necessary to consult a globe (an almost distortion-free model of the spatial relationship between points on the earth's surface), but various two-dimensional projections (Mercator, Albers projection, equirectangular projection) suffer for most applications, even though these each of these introduce specific misrepresentations and errors.

In practice, we will therefore use multidimensional scaling to obtain a low-dimensional Euclidean representation of a given distance matrix. Various diagnostic measures allow us to assess the misrepresentation error introduced, and to obtain a viable compromise with regard to the dimensionality used. This representation of the original systems in an abstract Euclidean space is called the *functional* or *behaviour* representation, since it illustrates the relationships between the systems measured and is obtained from their function or behaviour (as defined by the measurement device).

Classification in functional space

This functional space supplies us with coordinates that are then amenable to statistical analysis by the usual tools of multivariate analysis (Härdle and Simar, 2003; Webb, 2002). We will primarily be interested in the classification of different subtypes of systems, and the main tool we will employ is linear discriminant analysis (LDA). Although there exist many more involved methods, LDA is easy to implement, well understood both theoretically and in practice, and relatively robust: even if its assumptions (normal distribution for the subgroups in functional space, with equal variance) are not met, it is usually the method of choice for small sample sizes (number of systems studied), since further degrees of freedom (e.g., in quadratic discriminant analysis which discards with the equal variance assumption) need large datasets to improve upon LDA. The same Caveat also applies to truly nonparametric methods such as kernel density or nearest-neighbour classification, and these methods will not be used in our applications (where sample sizes are on the order of 10 to 100 typically).

As will be seen later, this rather simple sounding setup (distances between probability measures, reconstruction of coordinates in a functional space, classification of systems by normal models) allows for a surprisingly effective classification of complex systems, improving on currently established methods in many cases. Moreover, this approach is completely modular. Each of these steps can be individually refined, and we will indeed describe some of these possibilities in more detail later.

1.2 Reader's guide

Interdisciplinary research is faced with two dilemmas, both of which are mirrored in this thesis. The first is concerned with the *relevance* of findings. Ideally, problems from one discipline, when transferred into another domain, are more easily solved, and result in nontrivial insights in the original setting. Moreover, questions in the former should also lead to interesting problems and insights in the second domain. In practice, this balance is seldomly achieved and usually one domain benefits the most. Here we will be mostly concerned with two domains: the application domain (medicine, neuroscience), and the mathematical domain (analysis, statistics). We have tried to balance the text such that readers from both domains will find the thesis interesting. However, as the actual application of distance-based methods can be quite involved, the text is biased toward the applications, and the mathematical side of the story is not fully developed.

The second dilemma is concerned with the necessary level of *exposition*. Since this thesis should be both readable by mathematicians and non-mathematicians alike, compromises had to be made on both sides. On the one hand, we will not state mathematical results in their most elegant or general form, but discuss a setting that

is natural for most applications and refer to further results in notes at the end of the thesis (starting at page 229). On the other hand, we will also discuss some subtle mathematical problems and actually prove a few theorems.

Most of these technical details have been delegated to two appendices that can be consulted later or in parallel with the rest of the thesis, and there are therefore three entry points to this thesis:

- The reader can start reading with Chapter 2, which gives a detailed overview of distance-based analysis (without too many distracting technicalities),
- he/she can directly skip to the applications (starting with Chapter 3),
- or he/she can first read appendices A and B for a more mathematical exposition, and then continue with Chapter 2.

In the rest of this section we will briefly describe the contents of the remainder of this thesis.

Chapter 2: Dynamical systems and time series

Dynamical systems as a general framework for time-varying phenomena are discussed in Chapter 2. Optimal transportation distances for dynamical systems are defined, and their usefulness is discussed by way of examples. The main theoretical problem is the dependence of the distances on the measurement apparatus (projection from the space of systems to the metric space of measurements). In dynamical systems theory this is avoided by focussing on properties that are invariant under diffeomorphisms, i.e., smooth changes of coordinates. This solution is not available when metric properties are important. On the other hand, in practice one often faces the situation where a number of systems are measured by one and the same measurement apparatus, avoiding this difficulty. Other ways to aleviate this problem are discussed in Section 2.7.

This chapter is based on:

Muskulus M, Verduyn-Lunel S: 2009 — Wasserstein distances in the analysis of dynamical systems and time series. Technical Report MI-2009-12, Mathematical Institute, Leiden University. Extended journal version submitted.

Chapter 3: Lung diseases

In Chapter 3 we will analyze experimental data that consists of time series containing information on mechanical properties of the lungs. We will see that optimal transportation distances allow to successfully distinguish different lung diseases and might potentially allow to track airway status over the course of time. For completeness, the complementary approach of fluctuation analysis is also discussed.

This chapter is based on:

Muskulus M, Slats AM, Sterk PJ, Verduyn-Lunel S — Fluctuations and determinism of respiratory impedance in asthma and chronic obstructive pulmonary disease. Submitted.

Chapter 4: Structural brain diseases

Chapter 4 considers an application to brain imaging. Instead of tracking dynamical processes in time, the tissue properties of the brain are evaluated at one or more points in time by magnetic resonance (MR) imaging, and the goal is the quantification and detection of brain diseases from the distribution of MR parameters (relaxometry). The classical quantitative approach to quantitative MR imaging is outlined, including a discussion and critique of commonly used histogram analysis methods (Tofts, 2004). Wasserstein distances are then tested in the detection of subtle tissue changes in patients suffering from systemic lupus erythematosus (with respect to healthy controls), and in the detection of Alzheimer's disease.

This chapter is based on:

Luyendijk J, Muskulus M, van der Grond J, Huizinga TWJ, van Buchem MA, Verduyn-Lunel S — Diagnostic application of a new method of magnetization transfer ratio analysis in systemic lupus erythematosus: initial results. In preparation.

and also

Muskulus M, Scheenstra AEH, Braakman N, Dijkstra J, Verduyn-Lunel S, Alia A, de Groot HJM, Reiber JHC: 2009 — Prospects for early detection of Alzheimer's disease from serial MR images in transgenic mouse models. *Current Alzheimer Research* **6**, 503–518.

Chapter 5: Deformation morphometry

For completeness, Chapter 5 is included which discusses the related problem of deformation morphology, i.e., how to find regions in the brain that are deformed with respect to an average brain image.

This chapter is based on:

Muskulus M, Scheenstra AEH, Verduyn-Lunel S: 2009 — A generalization of the Moore-Rayleigh test for testing symmetry of vector data and two-sample problems. Technical Report MI-2009-05, Mathematical Institute, Leiden University. Extended journal version submitted.

Chapter 6: Electrophysiology of the brain

In contrast to the application to imaging data is the electrophysiological approach of Chapter 6, where time series with high temporal resolution are obtained. Even on the sensor niveau, i.e., on the brain surface, optimal transportation distances reveal interesting phenomena (Section 6.5). Ideally, this application would need to validate

the distances by a forward model, or to solve the inverse problem of source localization (e.g., by beamforming) and then compare source-related time series of activation. Unfortunately, many details of these exciting ideas still need to be worked out, and we therefore restrict ourselves to a proof of principles. A different use is discussed in Section 6.4, where Wasserstein distances are used to qualitatively compare distributions of instantaneous phases obtained from magnetoencephalographic recordings by the Hilbert transform. These offer interesting insights into the functional organization of neuronal networks.

This chapter is based on:

Muskulus M. Houweling S, Verduyn-Lunel S, Daffertshofer A: 2009 — Functional similarities and distance properties. *Journal of Neuroscience Methods* **183**, 31–41.

and also

Muskulus M, Verduyn-Lunel S: 2008 — Reconstruction of functional brain networks by Wasserstein distances in a listening task. In: Kakigi R, Yokosawa K, Kurik S (eds): Biomagnetism: Interdisciplinary Research and Exploration. Hokkaido University Press. Sapporo, Japan, pp. 59-61.

Epilogue, Boxes & Notes

In the epilogue, we look back on the experiences obtained with the optimal transportation distances and end with an outlook to the future. This is followed by a number of appendices that have been included for completeness (see below) and then a Notes section where additional topics are discussed, references to additional or complementary works are given, and other details are discussed that would disrupt the flow of the main text. To increase readability, scattered throughout the text are also *boxes* that highlight and summarize important points, see Box 1 for an example.

Appendix A: Distances

Appendix A recalls the basic facts and properties of distances. After introducing metric spaces as the abstract mathematical object in which a notion of distance is defined, the question of embeddability in Euclidean spaces is discussed, culminatig in the classical solution of Theorem 4. This is followed by an exposition of multidimensional scaling, which is the main tool that allows for the reconstruction of a metric space from its distances, when these are influenced by (small) errors and numerical inaccuracies. As discussed in Section 1.1, this reconstruction will be the vantage point for statistical analysis. Multivariate methods can be successfully applied in the reconstructed space, and we provide the interested reader with details about diagnostic measures, linear classification methods, cross-validation, and permutation tests for distance matrices.

Box 1. Additional typographic elements

- Summaries of main points and results can be found in boxes such as these, scattered throughout the text.
- Pointers to further topics and additional references can be found in the Notes at the end of the thesis.

Appendix B: Optimal transportation distances

Appendix B recalls facts about the optimal transportation distances that are used throughout the rest of the thesis. These distances are motivated by considering a few other possible distances and their shortcomings, and introduced in a general setting that shows the elegant theoretical foundation. For the application in practice this will be considerably restricted, and the theory will attain a discrete and more combinatorial flavour.

Additional appendices in the electronic version

In the electronic version, additional appendices complete the presentation, in which the two software packages are documented that were developed for the computations in this thesis.

1.3 Major results & discoveries

Due to the many topics touched upon in this thesis, it seems worthwile to stress the main innovations and results in a central place. Let us begin with the major general achievements:

- This thesis provides the main tools for the multivariate analysis of complex systems by distances. Although these are all more or less well-known techniques in their relative disciplines, this is the first time that they are *combined* in such a way.
- *Extensive software* has been written that allows for a streamlined application of the methods introduced and discussed here.
- All of the necessary background information is available *in one place* in this thesis, with extensive references also covering further, advanced topics.

Minor innovations that are introduced, but not actually worked out in detail, include the following:

- Wasserstein distances be used in the numerical analysis of dynamical systems to visualize their response to changes in parameters. This numerical *bifurcation analysis* is quantitative.
- The Wasserstein distances can be defined *relative to transformations* of the data, e.g., a translation or rotation. This offers an alternative way to solve the Procrustes problem, and quantifies differences in the shape of distributions that are invariant with respect to location or orientation.
- Wasserstein distances can always be interpolated along geodesics between two distributions. An iterative stochastic algorithm then allows to find approximations to centroids by these bivariate interpolations only. Thereby, it becomes possible to determine *characteristic representations* even for very complex kinds of data.

With regard to actual applications, the following are main results discussed in the text:

- Lung diseases are assessed routinely by the forced oscillation technique, but mostly only time averages are used. It is shown that there is potentially much more information contained in the dynamics of these signals, and that this information allows to discriminate between healthy and diseased lungs with very high accuracy.
- There is indeed evidence for power-law-like scaling behaviour of the fluctuations of respiratory impedance. This has been conjectured previously, but due to methodological problems the evidence in previous work has to be judged unreliable and inconclusive. Our findings provide evidence by state-of-the-art maximum-likelihood estimation that supports the power-law hypothesis..
- Systemic lupus erythematosus is a neurodegenerative disease that affects the brain. It has previously been shown that it significantly alters the distribution of magnetization transfer ratios in the brain, which is evaluated in so-called "histogram analysis" by quantifying changes in the location and height of its mode. The distance-based analysis improves on this and allows for better classification of individual patients.
- The distance-based analysis of distributions of relative phases, obtained from magnetoencephalographic signals in a bimanual coordination task, revealed and quantified interactions (crosstalk) between motor areas in a robust way. From the functional representation of these distances it now becomes possible to formulate hypotheses regarding the underlying neural networks and to set-up mathematical models.

Chapter 2

Dynamical systems and time series

Abstract

A new approach based on Wasserstein distances, which are numerical costs of an optimal transportation problem, allows to analyze nonlinear phenomena in a robust manner. The long-term behavior is reconstructed from time series, resulting in a probability distribution over phase space. Each pair of probability distributions is then assigned a numerical distance that quantifies the differences in their dynamical properties. From the totality of all these distances a low-dimensional representation in a Euclidean space is derived, in wich the time series can be classified and statistically analyzed. This representation shows the functional relationships between the dynamical systems under study. It allows to assess synchronization properties and also offers a new way of numerical bifurcation analysis.

The statistical techniques for this distance-based analysis of dynamical systems are presented, filling a gap in the literature, and their application is discussed in a few examples of datasets arising in physiology and neuroscience, and in the well-known Hénon system.

2.1 Introduction

Linear time series analysis is a well-developed technique with elegant theoretical underpinnings (Brockwell and Davis, 1998), but most real-world systems are decidedly nonlinear, which manifests itself in a much larger spectrum of possible dynamical behaviors. Possibilities include intermittency, bursting activity and a sensitive dependence on initial conditions, the latter being one of the hallmarks of so-called *chaotic* behavior. Prediction tasks are therefore much more difficult in nonlinear systems. For example, consider a time series

$$x = (x_1, x_2, \ldots, x_N) \tag{2.1}$$

of length N, generated by a dynamical system. In linear systems prediction is relatively straightforward by minimizing the total error made over some time interval of length $n < N$. Assume

$$x^* = (x_1^*, x_2^*, \ldots, x_N^*) \tag{2.2}$$

is a synthetic time series generated by a parametric model for the system under study. Optimizing the model parameters such that the error functional

$$d_n(x, x^*) \stackrel{!}{=} \sum_{i=1}^{n} ||x_i - x_i^*|| \tag{2.3}$$

is minimized usually results in an adequate fit that allows prediction on short to medium timescales. The parametric model then captures essential features of the (linear) dynamical system under study. In contrast to this, in a nonlinear system (and also in systems influenced by a stochastic process), such a model is usually useless. Already after a few time steps, the values x_i^* (for $i > n$) often show large deviations from the corresponding values x_i, and the parametric model does usually *not* capture the dynamics properly.

The focus in nonlinear time series analysis lies therefore not on predicting single trajectories, but on estimating the totality of possible states a system can attain and their statistical properties, i.e., how often the system can be expected to be in a particular state. Of particular importance hereby is the long-term behavior of the system, the so-called *attractor*, which can roughly be defined as the set of all *recurrent* states of the system (for a discussion of different notions of recurrence, see (Alongi and Nelson, 2007); formal definitions of attractors are given in (Milnor, 1985; Ruelle, 1981)). More precisely, the notion of an *invariant measure* captures the statistical properties of a dynamical system. This is a probability distribution over phase space that is invariant under the dynamics. In other words, if an ensemble of systems is taken with initial states randomly distributed according to the invariant measure, after evolving the systems for some common time, although the individual systems will be in quite different states than before, the distribution of systems in phase space does not change (in the limit of an infinite number of such systems). By necessity, invariant measures are concentrated on attractors, as recurrent behavior is a prerequisite for invariance.

Changes in long-term dynamical behavior can then be detected by comparing properties of the long-term behavior (or its invariant measure) (Hively et al., 1999; Diks et al., 1996). Unfortunately, many of these methods are based on the assumption that the dynamics is given by a *deterministic* (and possibly chaotic) process, and this usually unverifiable assumption can lead to doubts about the validity of the analysis (Rapp et al., 1993). Moreover, commonly used measures such as Hausdorff dimension and Lyapunov exponents are notoriously difficult to estimate. For this reason, Murray and Moeckel introduced the so-called *transportation distance* between attractors, which is a single number that expresses how closely the long-term behavior of two dynamical systems resembles each other (Moeckel and Murray, 1997). In contrast to general divergences (Frigyik et al., 2008; Ali and Silvey, 1966), for example the Kullback-Leibler divergence, mutual information or the Kolmogorov-Smirnov

statistic, this has the added avantage that it is a true *distance* on the space of (reconstructed) dynamical systems: It is reflexive, symmetric, and fulfills the triangle inequality, and is therefore a very natural concept to measure similarity of dynamical systems and their time series. In particular, this allows to compare more than two dynamical systems with each other in a sensible way.

The transportation distance is based on a convex optimalization problem that optimally matches two invariant measures, minimizing a cost functional. Mathematically, it is an example of a Wasserstein distance between probability measures (Villani, 2003). Although computationally involved, Wasserstein distances are much more robust than, for example, Hausdorff distance. Furthermore, these distances have interesting theoretical features, for example interpolation properties that allow to reconstruct dynamical behaviors *in between* two invariant measures.

Unfortunately, since their introduction in (Moeckel and Murray, 1997), the concept of transportation distance has received little attention. The reasons are probably that (i) its computation is involved, and (ii) to many researchers it did not seem clear how to further analyze the distances after they had been calculated. In this article we address the second point: After introducing general Wasserstein distances between dynamical systems (Section 2.2), and discussing implementation issues (Section 2.3), we show how such distances can be analyzed statistically (Section 2.4), which allows interesting insights into the global structure of dynamical systems. In particular, dynamical systems can be classified by properties of the *shape* of their invariant measures, that are quantified by these distances. Also, changes in behaviour when one or more parameters of the system are changed (i.e., bifurcations) can be studied from this new perspective.

Methods based on matrix eigendecompositions allow to represent and visualize these distances in a low-dimensional space that represents all possible dynamical behaviors of the systems under study (Section 2.4.2). In this *behavior space*, the totality of dynamical systems can be studied by the methods of multivariate statistical analysis. In particular, the statistical significance of separation of classes of systems in this space can be assessed by permutation methods (Section 2.4.5), and discriminant analysis allows to classify the time series by their dynamical features (Section 2.4.3).

We demonstrate the feasibility of our approach by two examples. First, we study the behavior of the Wasserstein distances with respect to sample size, parameter changes and the influence of noise in the well-known Hénon map (Section 2.5). Interactions between two systems are also discussed, where the distances allow to estimate coupling strength, i.e., our method also allows to assess (generalized) synchronization between dynamical systems (Section 2.5.4).

Secondly, we discuss an application to a dataset of tidal breathing records (Section 2.6), a subset of data previously published in (Slats et al., 2007), where we discriminate between patients suffering from asthma and those suffering from chronic obstructive pulmonary disease (COPD). Moreover, by the same methodology it is

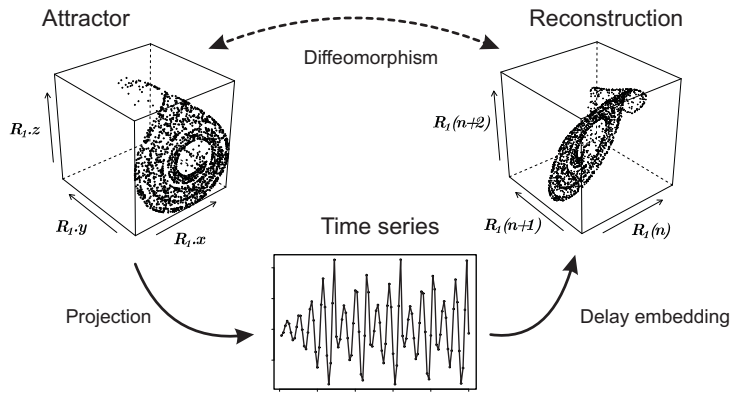


Figure 2.1: The methodology of attractor reconstruction via delay embeddings. The true attractor is projected into a time series by some measurement function, from which an image of the attractor can be formed by delay reconstruction, up to some diffeomorphism.

possible to trace changes in the dynamical state of the disease in one subject over the course of time.

These two distinct examples show the wide applicability of the concept of Wasserstein distances in nonlinear time series analysis. For completeness, we also include Section 2.7 where we discuss a further generalization of the Wasserstein distances that addresses a particularly interesting issue in nonlinear time series analysis, and that contains new ideas for future theoretical work.

2.2 Wasserstein distances

A dynamical system is implicitly given by the information contained in repeated measurements, and delay vector reconstruction allows to represent its trajectories in a Euclidean space (Packard et al., 1980; Takens, 1981; Stark, 2000). Given a time series

$$x = (x_1, \ldots, x_N) \tag{2.4}$$

of N measurements of a single observable X, a dynamical system is reconstructed by mapping each consecutive block

$$x_{[i]} = (x_i, x_{i+q}, \ldots, x_{i+(k-1)q}) \tag{2.5}$$

of k values, sampled at discrete time intervals q, into a single point $x_{[i]}$ in a Euclidean *reconstruction space* $\Omega = \mathbb{R}^k$. The intuitive idea is that the information contained in

the block $x_{[i]}$ fully describes the state of the (deterministic) system at time i, albeit in an implicit fashion. From a statistical point of view, the reconstructed points capture higher-order (i.e., not only between pairs of values as in linear time series analysis) correlations in the time series. If the embedding dimension k is large enough, and some simple genericity assumptions are fulfilled (Takens, 1981), the resulting distribution of points is indeed an *embedding* of the true attractor (in the limit of infinite time series), i.e., its topological and differential structure is identical to the attractor's, up to a smooth change of coordinates. The result that shows this is the following:

Theorem 1 (Takens (1981))**.** Let M be a compact manifold of dimension m. For pairs (X, y), X a smooth (i.e., C^2) vector field and y a smooth function on M, it is a generic property that $\Phi_{X,y} : M \to \mathbb{R}^{2m+1}$, defined by

$$\Phi_{X,y}(z) = (y(z), y(\varphi_1(z)), \ldots, y(\varphi_{2m}(z)))$$

is an embedding, where φ_t is the flow of X.

In our notation, $k = 2m + 1$ is the reconstruction dimension, and we allow for a general delay $q > 0$ instead of $q = 1$ as in Taken's original theorem. The two genericity assumption needed are the following: (i) If $X(x) = 0$ then all eigenvalues of $(\mathrm{d}\varphi_1)_X : T_X(M) \to T_X(M)$ are different and different from 1 (simple hyperbolicity), and (ii) that no periodic solution of X has integer period less or equal than k (resolvability). In practice, not only does one usually make these assumptions, they are almost always justified. In the sequel, we therefore assume that the reconstruction from Eq. 2.5 results in (the finite approximation of) an embedding.

This methodology of attractor reconstruction by delay embedding is illustrated in Figure 2.1. Even in the case of systems influenced by noise this reconstruction is possible (Stark et al., 1997). Likewise, attractors can also be reconstructed from multivariate time series, where more than one scalar variable is measured (Cao et al., 1998), but for simplicity of exposition we mainly consider the scalar case here.

The optimal value of the lag q can be estimated from the data (Fraser and Swinney, 1986) and similar tests exist for the embedding dimension k (Abarbanel et al., 1993; Kantz and Schreiber, 2004). The result of the embedding process is a discrete trajectory in phase space $\Omega = \mathbb{R}^k$ and this trajectory is interpreted as a probability measure μ on (the Borel σ-algebra of) Ω, where

$$\mu[A] = \frac{1}{N'} \sum_{i=1}^{N'} \delta_{x_{[i]}}[A], \quad A \subseteq \Omega, \tag{2.6}$$

is the time average of the characteristic function of the points in phase space visited; here $\delta_{x_{[i]}}$ is the Dirac measure of the block $x_{[i]}$ and $N' = N - (k-1)q$ is the length of the reconstructed series. In the limit $N' \to \infty$ the measure μ is invariant under the dynamics. Assuming that the system is subject to small random perturbations

leads to the uniqueness of the invariant measure under mild assumptions (Lasota and Mackey, 1997), which is then called the *natural invariant measure*. Its support contains an attractor in the sense of Ruelle (Ruelle, 1981). If a dynamical model is available, subdivision methods allow to approximate the attractor and its natural measure with arbitrary precision (Dellnitz and Junge, 1999); in the case of finite time series this measure is approximated by the available data.

In the following we assume that the reconstruction process has been performed, so let $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ now denote vectors in Ω. To compare the long-term behavior of dynamical systems quantitatively, we employ the Wasserstein distances of their natural invariant measures. Given two probability measures μ and ν on Ω, the Wasserstein distance $W(\mu, \nu)$ is defined as the solution of an optimal transportation problem in the sense of Kantorovich (Kantorovich, 1942b,a; Villani, 2003). The cost per unit mass is given by a distance function on phase space Ω. Only the case of Euclidean distance, also called L_2 distance,

$$d_2(x, y) = ||x - y||_2 = \left(\sum_{i=1}^{k} |x_i - y_i|^2 \right)^{1/2} \tag{2.7}$$

is considered here, since it is the natural choice, being rotationally invariant. Other distances are possible, though, and Moeckel and Murray (1997) use L_1 ("Manhattan") distance throughout. Although all distances are topologically equivalent in Euclidean space, distinct distances emphasize different aspects of the statistical properties (i.e., of the shape) of the invariant measures. In a sequel to this paper, we will discuss various properties and merits of the different distances.

The functional to be optimized is the total cost

$$C[\pi] = \int_{\Omega \times \Omega} ||x - y||_2 \; \mathrm{d}\pi[x, y], \tag{2.8}$$

over the set $\Pi(\mu, \nu)$ of all probability measures on the product $\Omega \times \Omega$ with prescribed marginals μ and ν, such that

$$\int_{\Omega} \mathrm{d}\pi[U, y] = \mu[U], \qquad \int_{\Omega} \mathrm{d}\pi[x, V] = \nu[V] \tag{2.9}$$

for all measurable $U, V \subset \Omega$ and all $\pi \in \Pi(\mu, \nu)$. Each measure $\pi \in \Pi(\mu, \nu)$ is interpreted as a transportation plan that specifies how much probability mass $\pi[x, y]$ is transferred from each location $x \in \Omega$ to each location $y \in \Omega$, incurring a contribution $d_2(x, y) \cdot \mathrm{d}\pi[x, y]$ to the total cost. The cost of an optimal transportation plan is called the *Wasserstein distance* between the measures μ and ν and is denoted by

$$W(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} ||x - y||_2 \; \mathrm{d}\pi[x, y]. \tag{2.10}$$

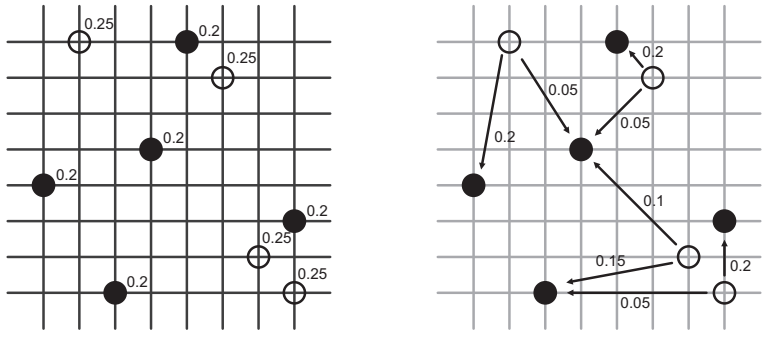


Figure 2.2: Example optimal transportation problem in the discrete case. Open circles correspond to the first measure, filled circles correspond to the second measure. For simplicity, the points are distributed on a grid with unit spacing. Left panel: Initial configuration. Numbers indicate probability mass at each point. Right panel: An optimal transportation plan with Wasserstein distance $W \approx 3.122$. The numbers next to the arrows indicate how much probability mass is transported from the first measure to the second measure.

Such problems arise in a number of applications in image analysis (Haker et al., 2004), shape matching (Gangbo and McCann, 2000) and inverse modeling in physics (Frisch et al., 2002). The measure theoretic formalism allows a unified treatment, but for finite time series the natural measure corresponds to a finite sum of Dirac measures. In this case the optimal transportation problem reduces to a convex optimalization problem between two *weighted point sets* and can be calculated by standard methods (see Section 2.3). In Figure 2.2, we show an example with the Dirac measures distributed on a regular grid.

Note that the specific Wasserstein distance we consider here is often called the Earth Mover's distance in the image analysis community (Rubner et al., 2000) and the Kantorovich-Rubinstein distance in the mathematical literature (Villani, 2003). This distance is preferred over the theoretically better understood squared Wasserstein distance (see (Villani, 2003) again), since it is more robust with respect to its statistical properties (confer the discussion in (Mielke and Berry, 2007)).

Remark 1. It is only possible to compare the long term dynamics of dynamical systems that occupy the same (reconstructed) phase space. This is not a problem in practice, when we compare classes of comparable dynamical systems, e.g., study the same system under parameter changes. This issue is further considered in Section 2.7.

2.3 Implementation

2.3.1 Calculation of Wasserstein distances

We assume that the reconstruction of the invariant measures has been performed, utilizing Theorem 1, resulting in discrete approximations of the invariant measures. As remarked before, the optimal transportation problem in the discrete case reduces to a transportation problem of weighted point sets (for possible approaches in the continuous case, which is an active area of research, see (Benamou et al., 2002; Haker et al., 2004)). Let the discrete measures be given by

$$\mu = \sum_{i=1}^{n_1} \alpha_i \delta_{x_i}, \qquad \nu = \sum_{j=1}^{n_2} \beta_j \delta_{y_j}, \tag{2.11}$$

where the *supplies* $\alpha_i \in (0, 1]$ and the *demands* $\beta_j \in (0, 1]$ are normalized such that $\sum_i \alpha_i = \sum_j \beta_j = 1$. The left panel of Figure 2.2 shows an example of two such measures (on a regular grid).

Any measure in $\Pi(\mu, \nu)$ can then be represented as a nonnegative matrix f_{ij} that is *feasible*, which is to say that it fulfills the source and sink conditions

$$\sum_j f_{ij} = \alpha_i, \quad i = 1, 2, \ldots, n_1, \quad \text{and} \tag{2.12}$$

$$\sum_i f_{ij} = \beta_j, \quad j = 1, 2, \ldots, n_2. \tag{2.13}$$

These are the discrete analogs of the respective conditions on the marginals in Eq. 2.9.

In this case the optimal transportation problem reduces to a special case of a minimum cost flow problem, the so-called transportation problem (Bertsekas, 1991; Balakrishnan, 1995):

$$W(\mu, \nu) = \min \sum_{ij} f_{ij} c_{ij}, \tag{2.14}$$

over all feasible flows f_{ij}, where $c_{ij} = ||x_i - y_j||_2$.

In principle, a general linear programming solver can be used to find the solution, but the special structure allows more efficient algorithms[1]. Indeed, the transportation problem can be solved in polynomial time by a network simplex algorithm (Schrijver, 1998; Balakrishnan, 1995). An actual implementation can be found in (Löbel, 1996). It is this algorithm that we have used in the examples in this paper.

Remark 2. Alternatively, relaxation methods can be used, for example, the *Auction algorithm* developed by Dimitri Bertsekas (Bertsekas and Castanon, 1989): Starting

[1] In the two-dimensional case, the transportation problem can be solved effectively in linear time, as already noted by Kantorovich. This is a consequence of the so-called Monge property of the distance matrix (Burkard et al., 1996).

from an initial condition, the total cost of the problem is successively reduced by a converging bidding process. Its main advantage is its ability to restart the problem from an approximate solution. Large numbers of similar transportation problems can be efficiently solved thereby. However, its implementation is non-trivial, so for most problems the algorithm of (Löbel, 1996) should be the first choice[2].

2.3.2 Bootstrapping and binning

Even with state-of-the-art algorithmic implementations, the computational cost of the calculation of Wasserstein distances remains a concern. A practical solution is to resample smaller subseries from the reconstructed trajectory and to estimate the Wasserstein distances multiple times, bootstrapping its expected value (Davison and Hinkley, 1997). Not only does this ease the computational burden tremendously (most algorithms for the transportation problem have at least a quadratic dependence on sample size), but it also supplies a quantitative measure of accuracy in the form of the bootstrapped *self-distances* $W(\mu, \mu)$ (see the discussion in Section 2.5.1), and introduces a further level of robustness (as the original time series are finite, we have to consider them as approximations of the true dynamical behavior anyway). This is the preferred method for most problems, and we discuss its properties in Section 2.5.

For completeness, we also mention that the reconstructed points can be clustered or binned prior to the calculation, as utilized in (Moeckel and Murray, 1997), for example. Since, by the Kantorovich-Rubinstein theorem, the distance function is based on a metric, we have that $W(\mu, \nu)$ depends only on the difference of μ and ν (see (Villani, 2003)). Therefore, if a measurement point $x \in \Omega$ of weight $\mu[x]$ is moved to a different location $x + \xi$, where $\xi \in \Omega$ is a displacement vector, the total cost changes by at most $||\xi|| \cdot \mu[x]$. This also shows that the Wasserstein distances are robust against the influence of (additive) noise, with the expected maximal error bounded by the standard deviation of the noise. Likewise, binning with regular bins of diameter $b \in \mathbb{R}$ introduces an error of at most $\sqrt{k} \cdot b$ to the total cost.

2.3.3 Incomplete distance information

In the following, we always assume that all pair-wise Wasserstein distances have been calculated for a set of dynamical systems under considerations. Since the number of distances grows quadratically with respect to the number of systems, in practice one might want to reduce the number of computations by only computing a fraction of the distance matrix. This point is discussed in (Borg and Groenen, 2005, Section 6.2), where it is shown that under low noise levels even in the absence of

[2] An implementation as a package for the statistical computing environment R (http://www.r-project.org/) is available from the author's homepage.

80% of the distances (randomly chosen) the remaining distances contain enough information for excellent recovery when a modification of the method of Section 2.4.2 is applied (Spence and Domoney, 1974). In the case of noisy time series, the reconstruction of dynamical behavior is an area of active research. At the moment, the recently published method of (Singer, 2008) should be the preferred approach.

2.3.4 Violations of distance properties

In practice, if the invariant measures are bootstrapped because of computational complexity considerations, violations of the distance properties can arise. These are due to the finite number of points sampled, and only the triangle inequality is potentially affected. For completeness, we will also discuss violations of reflexivity here.

The triangle inequality is violated if

$$W(\mu,\nu) > W(\mu,\eta) + W(\eta,\nu) \tag{2.15}$$

for some triple of points μ, ν, η from the set of invariant measures that are considered. For a finite number of points such a violation can always be corrected. If it occurs, let

$$c = \max_{\mu,\nu,\eta \in P(\Omega)} W(\mu,\nu) - W(\mu,\eta) - W(\eta,\nu) \geq 0 \tag{2.16}$$

be the maximal violation of the triangle inequality, where $P(\Omega)$ denotes the set of the dynamical systems considered. Adding c to all distances, $W(\mu,\nu) \mapsto W(\mu,\nu) + c$, corrects the violation. The value of c can be found in time of order $O(N^2 \log N)$ by sorting the sums of distances on the right-hand side of Eq. 2.15. The effect of adding such a constant is more pronounced for the smaller distances, which get stretched more than the larger ones. The constant c is a very interesting measure in itself (Laub et al. (2006); also see the discussion of this point in (Borg and Groenen, 2005)).

Violations of reflexivity arise, for example, when one estimates the self-distances $W(\mu,\mu)$ under bootstrapping. Of course, these can be simply corrected by setting $W(\mu,\mu) \mapsto 0$; nevertheless, the estimation of $W(\mu,\mu)$ under bootstrapping allows one to assess the simulation error. It seems a reasonable assumption that each distance is perturbed by a normal distribution with mean zero and standard deviation $\sigma(\mu,\nu)$ that depends on the two measures. However, since distances can be only nonnegative, for the self-distances and measures whose true distance is smaller than $\sigma(\mu,\nu)$ this has to be modified. A simple model for the errors is then

$$W(\mu,\nu) = W_0(\mu,\nu) + |\epsilon(\mu,\nu)|, \qquad (W_0(\mu,\nu) < \sigma(\mu,\nu)), \tag{2.17}$$

$$W(\mu,\nu) = W_0(\mu,\nu) + \epsilon(\mu,\nu), \qquad (W_0(\mu,\nu) > \sigma(\mu,\nu)), \tag{2.18}$$

where $W_0(\mu,\nu)$ is the theoretical distance for infinite time series and sample size and $\epsilon(\mu,\nu) \sim \mathcal{N}(0,\sigma^2(\mu,\nu))$. Of course, since $W_0(\mu,\nu)$ is not known, in practice it is

difficult to tell whether a small distance signifies almost identical systems, i.e., the *resolution* of the Wasserstein distances is on the order of $\epsilon(\mu, \nu)$. The standard choice then is to leave all distances $W(\mu, \nu)$ with $\mu \neq \nu$ unchanged, lest they violate the triangle inequality. However, depending on the application, one might make different choices with regard to this (see Section 2.5.4). In Section 2.5 we show numerical evidence that the assumption of normality is approximately fulfilled in practice.

2.4 Analysis

In Figure 2.3 we show the steps in the analysis of dynamical systems by Wasserstein distances. From an underlying dynamical system, measurements are obtained in the form of time series. From these, discrete approximations of the invariant measures are reconstructed by standard delay embedding. After calculating the Wasserstein distances for all pairs of such probability measures, one can store the information in a distance matrix. In the following we discuss how to proceed with the analysis of the information contained in this matrix.

2.4.1 Distance matrices

The statistical analysis of distance matrices is a well developed topic in multivariate analysis (Härdle and Simar, 2003; Borg and Groenen, 2005). Important applications arise in ecology (Legendre and Legendre, 1998), psychology, and in the statistical analysis of shape (Small, 1996). Far from being comprehensive, we give a short overview of some techniques that are particularly useful in the analysis of Wasserstein distances.

Throughout we assume that the distance information is presented in the form of a single matrix M whose entries $M_{ij} = W(\mu_i, \mu_j)$ represent the distance between two dynamical systems (which are calculated from their invariant measures μ_i and μ_j, as discussed before). The actual distance used is left unspecified. In the later examples (Section 2.5-2.6) we employ the Kantorovich-Rubinstein distance (Eq. 2.10), but the class of Wasserstein distances contains other interesting distances (e.g. total variation) that test distinct properties of the invariant measures. The interesting problem of how to combine the information obtained from various distance measures into a generalized distance matrix is beyond the scope of the present paper. In any case, we first need to consider how a single distance matrix can be analyzed.

2.4.2 Reconstruction by multidimensional scaling

Multidimensional scaling (MDS) is the generic name for a number of techniques that model distance data as points in a geometric (usually Euclidean) space. In the application to dynamical systems, each point in this space represents a single dynamical

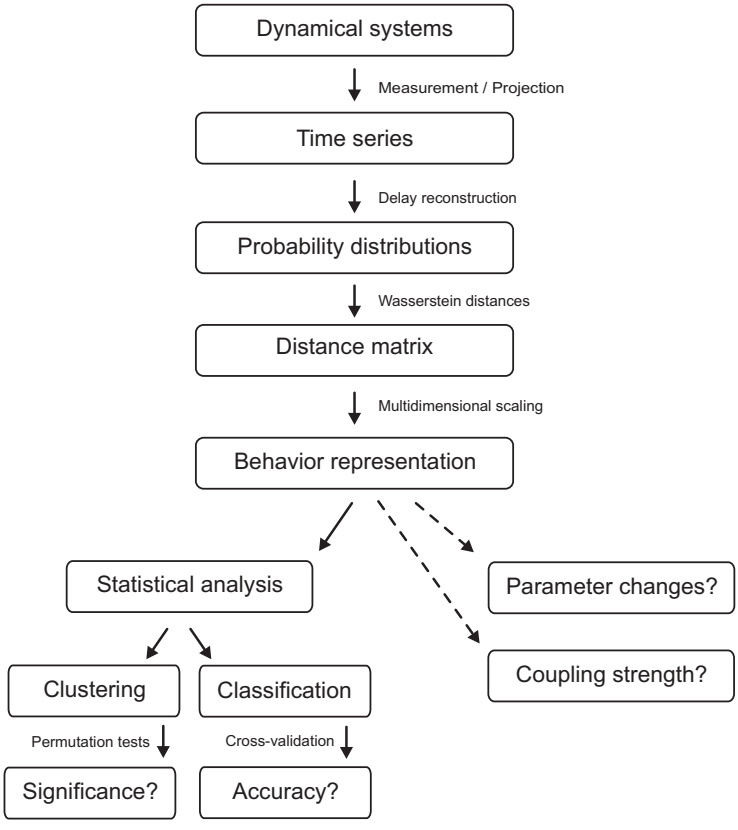


Figure 2.3: Outline of the methodology of distance-based analysis of dynamical systems and time series. The solid arrows indicate the main flow of steps. The broken arrows indicate optional steps that are applicable in particular situations. Note that we do not discuss clustering methods in this paper, as they are generally well-known.

system and the space can be interpreted as the space of (the totality of) their possible dynamical behavior. We therefore call it the *behavior space*. It should not be confused with the k-dimensional reconstruction spaces of each single dynamical system, which are only used in the calculations of the Wasserstein distances. As the behavior space represents possible dynamical behavior, its dimension is not directly related to the dimensionality of the dynamical systems under consideration, but rather reflects the structure of the dynamical variability inherent in the totality of systems studied.

Classical (also called metric) MDS is similar to principal component analysis (PCA) and has been pioneered by Torgerson and Gower (see (Borg and Groenen, 2005) for references). Although there are many modern developments and variations

(a few of which are discussed in Section 2.8), we only focus on classical MDS. Let us assume *a priori* that the distances M_{ij} are the distances between n points (representing n dynamical systems) in a m-dimensional Euclidean space, for some choice of $m \leq n$. Denote the coordinates of the i-th point by $x_{i1}, x_{i2}, \ldots, x_{im}$. In the following, we want to determine the n-by-m matrix $X = x_{ij}$ of the totality of these coordinates from the distances in M_{ij}.

The squared distances $(M_{ij})^2$ can be expanded as

$$(M_{ij})^2 = \sum_{a=1}^{m} \left(x_{ia}^2 + x_{ja}^2 - 2x_{ia}x_{ja} \right), \tag{2.19}$$

which results in the matrix equation

$$D^2 = c1_n' + 1_n c' - 2XX'. \tag{2.20}$$

Here D^2 represents the matrix with elements $D_{ij}^2 = (M_{ij})^2$, the vector $c = (c_1, \ldots, c_n)'$ consists of the norms $c_i = \sum_{a=1}^{m} x_{ia}^2$, and 1_n is an n-by-1 vector of ones. The matrix transpose is indicated by $'$.

Reversing this identity, the scalar product matrix $B = XX'$ is given by

$$B = -\frac{1}{2} J D^2 J, \tag{2.21}$$

where $J = I - \frac{1}{n} 1_n 1_n'$ is the *centering matrix*, and I denotes the n-by-n identity matrix. The operation in Eq. 2.21 is called *double centering* and has been thouroughly studied (Critchley, 1988). It is often called "an application of the law of cosines" in the literature. Note the use of squared distances; this is necessary since the "distances" $x_{i\cdot} - x_{j\cdot}$ are unknown, as the (absolute) distances $M_{ij} = |x_{i\cdot} - x_{j\cdot}|$ contain no information on the sign.

To find the classical MDS coordinates from B, we factor B by its eigendecomposition (singular value decomposition):

$$B = Q\Lambda Q' = (Q\Lambda^{1/2})(Q\Lambda^{1/2})' = XX'. \tag{2.22}$$

Here $\Lambda^{1/2}$ is the matrix square root of Λ; this exists as Λ is a diagonal matrix, with the eigenvalues of B on its diagonal.

In general, the dimension m is not known in advance, and has to be considered a parameter. Let the eigenvalues of B be ordered by decreasing size (by permuting the relevant matrices, if necessary). Denote by Q_m the matrix of the first m columns of Q; these correspond to the first m eigenvalues of B, in decreasing order. The coordinate matrix of classical MDS is then given by

$$X := Q_m \Lambda_m^{1/2}. \tag{2.23}$$

The distances in M can now be represented as points in a *Euclidean* space if X is real, or equivalently, if the first m eigenvalues of B are nonnegative (Young and Householder, 1938; Havel et al., 1983). In that case, the coordinates in X are found up to a rotation. Moreover, the reconstructed coordinates are a principal axes solution, i.e., the coordinates of a m'-dimensional reconstruction, where $m' < m$, correspond to the first m' coordinates of an m-dimensional reconstruction, which allows a nested analysis. Since this is PCA of scalar products, it has been called principal coordinate analysis by Gower. However, there is a subtle difference: The centering operation usually has the effect that the first principal component (representing a baseline/mean) has been removed (Heiser and Meulman, 1983).

The optimal maximal dimensionality m of the reconstruction can be determined by considering the *strain*,

$$S = ||XX' - B||^2 = \sum_{ij} |(XX')_{ij} - B_{ij}|^2. \tag{2.24}$$

The strain quantifies the error made by projecting the distances to the m-dimensional subspace, and decreases monotonously as the reconstruction dimension m is increased, as long as no negative eigenvalues are encountered under the m eigenvalues used in the reconstruction. However, the speed of decrease varies with the dimensionality. A rapid fall in the beginning usually turns into a much slower decrease above a certain dimensionality m^*, a so-called *elbow phenomenon* (see Panel C in Figure 2.15 for an example). The dimension m^* so obtained is the usual, optimal choice for m, representing a compromise between parsimony and resolution similar to the Akaike information criterion. Of course, depending on the actual use of the behavior space representation, there might be more appropriate ways of determining the optimal dimension.

Note that the primary use of the MDS reconstruction is *dimension reduction*. This is particularly useful in *exploratory data analysis*, i.e., as a first step in a comprehensive analysis where the emphasis is on the detection of interesting features, and in the *visualization* of distance information. In the example sections, we use a number of two-dimensional reconstructions of the behavior space for visualization purposes (as more than two dimensions are obviously difficult to assess visually).

A different application is the discrimination of lung diseases by their dynamical properties (Section 2.6). In this example, we determine the optimal dimension of the behavior space by cross-validation of the accuracy of linear discrimant analysis. We now turn to a discussion of this technique. Before that, however, let us stress the main principle underlying the distance-based analysis of dynamical systems.

Principle 1. The reconstructed behavior space, i.e., the MDS coordinates derived from a distance matrix, is the object at which all (statistical) analysis starts.

Following this principle, in the following sections on statistical analysis we only

consider points in behavior space and do not consider distance matrices anymore.

2.4.3 Classification and discriminant analysis

In applications, an important problem is the classification of time series, see Section 2.6 where we use time series of respiratory measurements to discriminate between two lung diseases. Again, we only discuss the standard approach, *linear discriminant analysis* (LDA) or Fisher discriminant analysis, for two classes.

Assume a number of points $x_i \in \mathbb{R}^m$ are given, where $1 \leq i \leq n$. Consider a partition of the index set $I = (1, \ldots, n)$ into the indices I_1 belonging to the first class, and the remaining indices $I_2 = I \setminus I_1$. The weighted class means (also called centroids) are

$$c_1 = \frac{1}{n_1} \sum_{i \in I_1} x_i, \qquad c_2 = \frac{1}{n_2} \sum_{i \in I_2} x_i, \tag{2.25}$$

with corresponding intra-class variances

$$\Sigma_1^2 = \sum_{i \in I_1} (x_i - c_1)(x_i - c_1)', \qquad \Sigma_2^2 = \sum_{i \in I_2} (x_i - c_2)(x_i - c_2)'. \tag{2.26}$$

The overall mean is

$$\bar{x} = \frac{1}{n} \sum_i x_i = \frac{n_1 c_1 + n_2 c_2}{n}. \tag{2.27}$$

The goal of LDA is to find a vector $w \in \mathbb{R}^m$ that maximizes the *generalized Rayleigh quotient*

$$J(w) = \frac{w'(c_1 - c_2)(c_1 - c_2)'w}{w'(\Sigma_1^2 + \Sigma_2^2)w}, \tag{2.28}$$

i.e., the difference in means divided by the sum of variances, all of which are projected onto the direction of w. The motivation for this is that the optimal direction maximizes the separation (or inter-class scatter) of the means, scaled by the variances in that direction (the corresponding sum of intra-class scatter), and which can, in some sense, be considered the signal-to-noise ratio of the data.

The direction w is easily found by a spectral technique (Shawe-Taylor and Cristianini, 2004), and the method is implemented in standard software packages (for example, see (Maindonald and Braun, 2003)). Points are then classified by their nearest neighbour in the projection onto the direction of ω. Application of LDA to point coordinates in behavior space allows to classify dynamical systems.

Note that it is not possible to apply LDA directly on distance matrices since these are collinear, and the results therefore cannot be trusted (Næs and Mevik, 2000). This is the main reason behind Principle 1.

2.4.4 Cross-validation

It is well known from work in machine learning that *resubstitution accuracy*, i.e., predictive accuracy on the data used to derive a model, inevitably improves as the prediction model becomes more complex. In the case of LDA in behavior space, increasing the dimensionality m of the behavior space inevitably improves the accuracy of classification (as long as no negative eigenvalues are encountered). However, this does not usually tell us much about the accuracy obtained when faced with the classification of an additional data item of unknown class.

The usual solution to assess predictive accuracy in a useful way is to partition the available data into a training and a test set of about the same size. After setting up the discrimination method on the former, its accuracy is then tested on the latter. However, for small datasets this is usually not feasible, so we recommend the use of cross-validation. In leave-one-out cross-validation, the i-th data point is removed from the n points available, the discriminant function is set up, and the i-th point classified, for all possible values of $i \leq n$. The average accuracy of all these classifications is the (leave-one-out) *cross-validated predictive accuracy* of the classification.

Cross-validation of LDA in behavior space seems straightforward: first the behavior space is constructed by the classical MDS solution, then the classification of points in this space is cross-validated. Note however that a (often significant) bias is introduced, if the MDS reconstruction makes use of the distance information of each point that is left out in the cross-validation step. Ideally, when classifying the i-th point as an "unknown data item" we would like to construct behavior space from a submatrix of the distance matrix, with the i-th row and column removed, classifying the i-th point in this space. For simplicity, let $i = n$, such that the coordinates of the last point need to be found in the behavior space defined by the first $n - 1$ points. A solution to this problem has been recently given in (Trosset and Priebe, 2008), following earlier work of (Anderson and Robinson, 2003).

The main idea is to apply double centering to D^2 with respect to the centroid of the first $n - 1$ points only. Instead of deriving the scalar product matrix by the usual double centering (Eq. 2.21), the scalar product matrix B is then computed as

$$B = -\frac{1}{2}\left(I - \frac{1}{n-1}1_{n-1}1'_{n-1}\right) D^2 \left(I - \frac{1}{n-1}1_{n-1}1'_{n-1}\right), \tag{2.29}$$

where $1'_{n-1}$ is used instead of $1'_n$. Denote by b the *fallible* scalar products of the cross-validated item with the others, and by β its squared norm. The coordinates $y \in \mathbb{R}^m$ of the last item are then given as the solution of the following nonlinear optimization problem (Trosset and Priebe, 2008):

$$\min_{y \in \mathbb{R}^m} \; (\beta - y'y)^2 + 2\sum_{i=1}^{n}(b_i - x'_i y)^2, \tag{2.30}$$

which can be solved by standard methods. Our implementation uses the Nelder-Mead simplex algorithm (Nelder and Mead, 1965).

2.4.5 Statistical significance by permutation tests

Given a partition of time series into two or more classes, one way to quantify the *separation* between the classes is given by the cross-validated predictive accuracy of the previous section.

More directly, however, from the representation in behavior space we can calculate the ratio of the intra-class average distances by the inter-class average distances. Unfortunately, this single number does not tell us how *significant* (in the statistical sense) the separation is. A solution to this problem is given by the multiple response permutation procedure (MRPP) (Mielke and Berry, 2007), a method that allows to assess the significance of separation of two or more classes in an independent and unbiased way. Its advantage is that it does not require the assumption of normality that is inherent in most multivariate tests of association (see Huberty and Olejnik (2006) for an overview).

Assuming two classes of systems as before, the usual MRPP statistic is given by

$$\delta = \sum_{i=1}^{2} \frac{n_i}{n_1 + n_2} \Delta_i, \tag{2.31}$$

where

$$\Delta_i = \frac{1}{\binom{n_i}{2}} \sum_{k,l \in I_i} M_{kl}, \qquad i = 1, 2. \tag{2.32}$$

is the average distance of the i-th class.

Under the null hypothesis that the classes of dynamical systems arise from the same (unknown) distribution of systems in behavior space, we can reassign their class labels arbitrarily. For each of these $\binom{n_1+n_2}{n_1}$ labelings, the MRPP statistic δ is calculated. The distribution of values of δ under all possible relabelings is (for historical reasons) called the *permutation distribution*. The significance probability (P-value) of this statistical test is given by the fraction of labelings of the permutation distribution with a smaller value of δ than the one obtained by the original class labels. Note that the δ statistic itself is generally not scale-invariant, but that the P-value derived from it can be used to compare the quality of separation across different datasets.

In practice the number of possible labelings to consider is usually too large, so the results in the example sections are based on 10^5 randomly generated labelings, as is common practice in statistics.

2.5 Example: The Hénon system

In this section we demonstrate some basic properties of Wasserstein distances using the well-known Hénon map (Hénon, 1976) to generate the time series. The Hénon map is given by

$$\begin{aligned} x_{n+1} &= 1 + y_n - a x_n^2, \\ y_{n+1} &= b x_n. \end{aligned} \tag{2.33}$$

Throughout, we use simulations of the Hénon system for 5096 time steps. Discarding the first 1000 samples as a transient leaves $N = 4096$ values for the analysis.

2.5.1 Sample size and self-distances

As discussed before, bootstrapping the Wasserstein distances leads to an error which is a combination of simulation error, due to the finite number of bootstraps, plus a statistical error, due to the finite number of points from the invariant measures sampled and the finite length of the time series. Fortunately, the estimation of the self-distances $W(\mu, \mu)$ allows to assess these errors. The left panel of Figure 2.4 shows the self-distances against the sample size used for bootstrapping in a double logarithmic plot. Only the values of the x variable have been used for the reconstruction, with the standard choice of parameters $a = 1.4$ and $b = 0.3$, for which it is known that the Hénon map exhibits chaotic behavior. All distances have been bootstrapped 25 times.

The standard deviation of the bootstrapped distance was lower than the vertical extent of the crosses used in the plot and is therefore not indicated in Figure 2.4. This shows that the simulation error is much smaller than the statistical error, so bootstrapping the Wasserstein distances with the low number of 25 realizations seems sufficient. Compare Figure 2.5 for a typical distribution of the bootstrapped distances for the Hénon system ($N = 25$ and $N = 1000$ realizations).

The lowest line in Figure 2.4 corresponds to a one-dimensional (trivial) embedding. Increasing the embedding dimension leads to the lines above it, with the highest one corresponding to a six-dimensional delay embedding. As expected, the self-distances decrease with increasing sample size. Interestingly, the slope of this decrease is -0.53 ± 0.03 ($R^2 = 0.989$, P-value $4.4 \cdot 10^{-6}$), in the double-logarithmic plot (for embedding dimension $k = 3$, with similar values for the other dimensions), which is consistent with the typical scaling behavior of Gaussian noise. In other words, the error is mainly statistical, which is evidence for the robustness of the Wasserstein distances. This also provides evidence for the hypothesis in Sec. 2.3.4 on the Gaussian nature of the errors. A different value of the slope would suggest that the dynamics of the Hénon map influence the Wasserstein *self*-distances, but

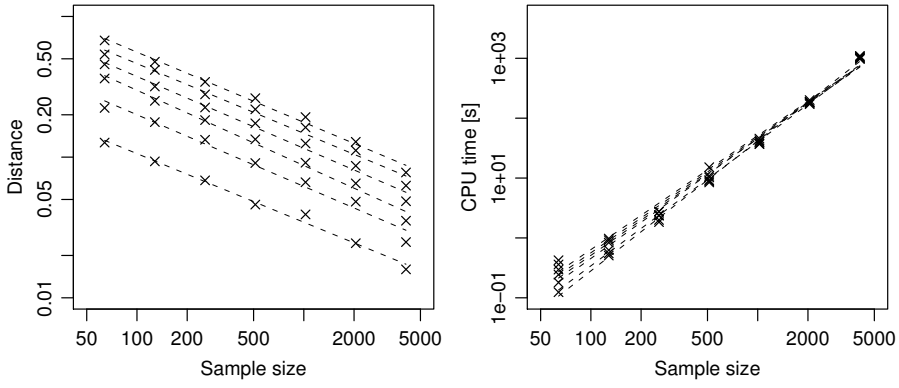


Figure 2.4: Dependence of Wasserstein self-distances on sample size. Left panel: Wasserstein distances for embedding dimensions 1 (lowest curve) to 6 (highest curve). The deviation from the true value of zero is an indication of the statistical error. The slope of the regression lines is roughly $-1/2$, which is the typical scaling behavior of Gaussian noise. Right panel: CPU time needed for these calculations, with a slope of roughly 2, i.e., a quadratic dependence on sample size.

even for small sample sizes no deviation from the square root scaling behavior can be discerned.

The right panel of Figure 2.4 shows CPU time in seconds, on a typical personal computer (AMD Athlon XP 2400+). The exponent in this case is 2.01 ± 0.04 ($R^2 = 0.989$, P-value $< 10^{-16}$), so the typical time complexity of the Wasserstein distances is quadratic with respect to sample size.

From the above we see that self-distances can be used to assess errors in embeddings, and that they can also provide an alternative way to estimate the optimal embedding dimension in nonlinear time series analysis.

2.5.2 Influence of noise

To study the influence of additive noise, normally distributed random variates were added to each point of the time series prior to reconstruction of the invariant measures. The mean of the noise was zero, and the standard deviation a fixed fraction of the standard deviation of the signal over time. Figure 2.6 shows the dependence of the Wasserstein self-distances for different noise levels. In the left panel, the embedding dimension was varied from one (lowest line) to six (highest line), for a fixed sample size $N = 512$ and 25 bootstraps. The effect of noise is higher for larger

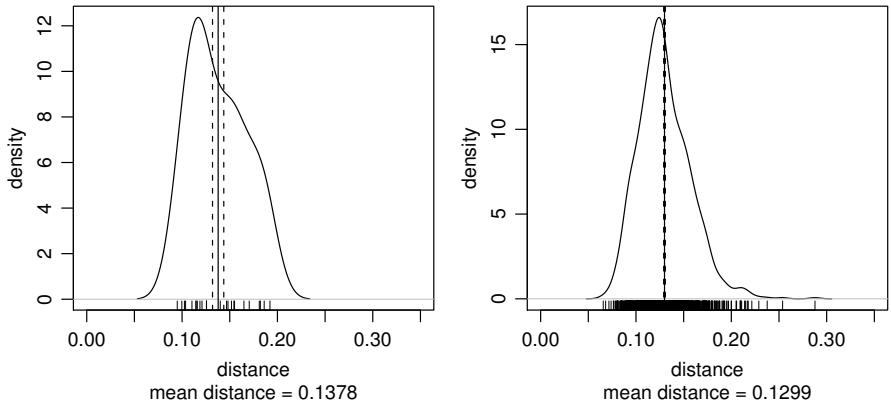


Figure 2.5: Distribution of Wasserstein self-distances under bootstrapping in the Hénon map, for 512 sample points. Left panel: $N = 25$ bootstraps. Right panel: $N = 1000$ bootstraps. Curves are kernel density estimates and the rugs at the bottom indicate the individual values of the distances. Vertical lines show mean distance (solid) and its standard deviation (stippled) over all N realizations.

embedding dimensions, with a linear change in the slope of the regression lines of 0.15 ± 0.01 ($R^2 = 0.99$, P-value $8.0 \cdot 10^{-5}$). This results from the added degrees of freedom in higher dimensions, which account for the linear increase in error. This error can partially be compensated by increasing the sample size, as can be seen in the right panel of Figure 2.6, for the case of a three-dimensional embedding. For $N = 512$ sample points, the slope of the Wasserstein distances is 2.02 ± 0.03 (with similar values for other sample sizes), i.e., the statistical error doubles for noise on the order of the original variability in the signal. This shows the robustness of the Wasserstein distances with respect to noise, since the statistical error is of the order of the signal-to-noise ratio, and not higher.

Moreover, due to this (almost) linear dependence of the Wasserstein distances on the signal-to-noise ratio, it should be possible to estimate the *noise level* of the signal and extrapolate its theoretical noise-free value by estimating the Wasserstein distances under artificially *added* Gaussian noise ("noise titration", see (Poon and Barahona, 2001)) of known standard deviation, for a few distinct noise levels.

2.5.3 Visualizing parameter changes

One of the most interesting aspects of the distance analysis outlined in Section 2.4 is the possibility to visualize changes in dynamical behavior with respect to parameter

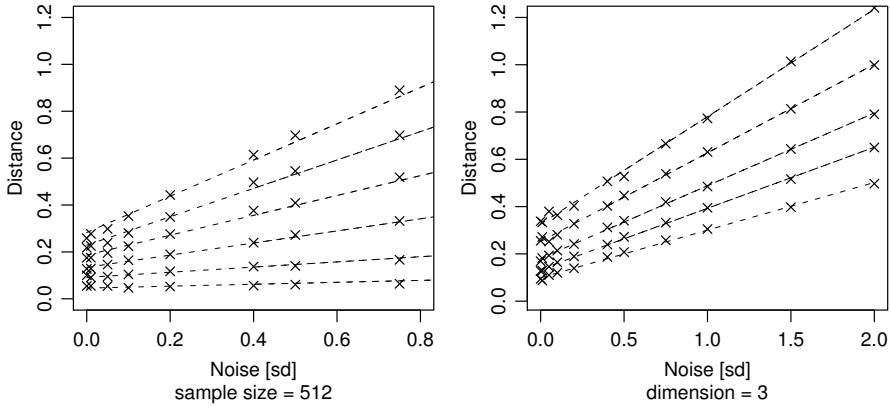


Figure 2.6: Dependence of Wasserstein self-distances on noise. Left panel: Wasserstein distances for embedding dimensions 1 (lowest curve) to 6 (highest curve) and fixed sample size $N = 512$. Right panel: Wasserstein distances for sample sizes $N \in \{64, 128, 256, 512\}$ (from top to bottom) and fixed embedding dimension $k = 3$.

changes, similar to a bifurcation analysis. However, whereas in the usual bifurcation analysis only regions of phase space are identified where the qualitative behavior of a dynamical system changes, in the distance-based analysis of dynamical systems these changes are quantified. This has not only potential applications in numerical bifurcation analysis, but also aids in quickly identifying interesting (for example, atypical) regions of parameter space. We demonstrate this approach again using the Hénon map.

The parameters a, b of the Hénon map were varied, with a ranging from 0.7 to 1.4 in steps of 0.05, and b ranging from 0.02 to 0.3 in steps of 0.02. For simplicity, only parameter values $(a, b) = (a_i, 0.3)$ and $(a, b) = (1.4, b_j)$ were considered, where $a_i = 1.4 - 0.05i$, for $0 \leq i \leq 14$, and $b_j = 0.3 + 0.02j$, for $-14 \leq j \leq 0$. The invariant measures of the x-variable, corresponding to the trivial embedding dimension $k = 1$, are shown in Figure 2.7. Dark areas correspond to large time averages, and light areas to small time averages. On the top of the plots, the indices of the corresponding parameter values are indicated.

Bootstrapping all mutual distances, again by 25 bootstraps with 512 sample points each, the left panel of Figure 2.8 shows a two-dimensional projection of behavior space, i.e., of the Wasserstein distances of the respective dynamical systems. The distinct behavior of these systems, with respect to parameter changes, is clearly discernible. Larger deviations of the parameters from $(a_0, b_0) = (1.4, 0.3)$ result in points that are farther away from the point 0, corresponding to (a_0, b_0). Summariz-

ing, the points are well-separated, although quite a few of their distances are smaller than the mean self-distance 0.091 ± 0.005 (indicated by a circle in the left panel of Figure 2.8). Note that the triangle inequality was not violated, but subtracting more than 0.030 will violate it. Only the self-distances have therefore been adjusted, by setting them to zero.

Theoretically, as the Wasserstein distances are true distances on the space of (reconstructed) dynamical systems, it is clear that the points corresponding to changes in one parameter only lie on a few distinct piecewise-continuous curves in behavior space. At a point where the dynamical system undergoes a bifurcation, i.e., a qualitative change in dynamical behavior occurs, these curves are broken, i.e., a point past a bifurcation has a finite distance in behavior space from a point before the bifurcation. The relatively large distance of point 10 (with parameter $a_{10} = 0.9$) from the points with indices larger than 11 corresponds to the occurence of such a bifurcation, as seen in Figure 2.7.

The right panel of Figure 2.8 shows a two-dimensional reconstruction of the Hénon system on a smaller scale, where the parameters were varied as $a_i = 1.4 - 0.0125i$, for $0 \leq i \leq 14$, and $b_j = 0.3 + 0.005j$, for $-14 \leq j \leq 0$, i.e., for values of a ranging from 1.4 to 1.225, and b ranging from 0.3 to 0.23. Even on this smaller scale, where the mean self-distances were 0.118 ± 0.003, the points are relatively well separated and there are indications of bifurcations. Note that the triangle inequality again holds, with a threshold of 0.070 before it is violated.

2.5.4 Coupling and synchronization

Wasserstein distances also allow to quantify the coupling between two or more dynamical systems, for example, to analyze synchronization phenomena in dynamical systems (Pikovsky et al., 2003). In this section we consider two unidirectionally coupled chaotic Hénon maps similar to the example discussed in (Stam and van Dijk, 2002). The systems are given by the following equations

$$x_{n+1} = 1 + y_n - 1.4x_n^2, \qquad y_{n+1} = 0.3x_n, \tag{2.34}$$

$$u_{n+1} = 1 + v_n - 1.4(Cx_n + (1-C)u_n)u_n \qquad v_{n+1} = Bv_n, \tag{2.35}$$

and we call the (x, y) system the *master* and the (u, v) system the *slave* system. The strength of the coupling is given by the coupling parameter C, which was varied from 0 (uncoupled systems) to 1 (strongly coupled systems) in steps of size 0.05. The parameter B was either $B = 0.3$ (equal systems) or $B = 0.1$ (distinct systems).

Figure 2.9 shows Wasserstein distances between the dynamics reconstructed from the variables x and u, respectively, against coupling strength C, in a separate panel for each of these two distinct cases. As before, the time series consisted of 5096 values of which the first 1000 values were discarded as a transient. Reconstruction was performed in three dimensions and the distances were bootstrapped 25 times, with

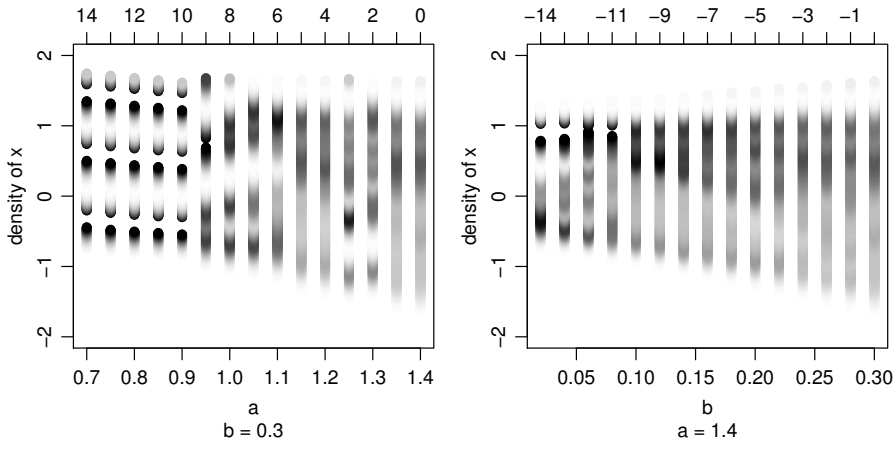


Figure 2.7: Invariant measures of the x-variable in the Hénon system, for different values of the parameters. Left panel: Variation in parameter a, with constant $b = 0.3$. Right panel: Variation in parameter b, with constant $a = 1.4$. See text for details of the parameter values used. Darker shade indicates large time averages, and lighter shade smaller time averages. Top axis of the panels shows indices of the dynamical systems used in Fig. 2.8.

512 samples each. The initial conditions of the two Hénon systems were chosen uniformly from the interval $[0, 1]$. Ten such random initial conditions were considered and are depicted in Figure 2.9 as distinct lines (top). The dots correspond to the mean of the distances over the ten realizations. The variation over the 10 different initial conditions is considerably small, as expected, i.e., the approximations of the invariant measures are considerably close to the true, unique invariant measure, that does not depend on the initial condition. The bottom lines display corrected distances, where the *minimum* of all distances has been subtracted. This seems appropriate in the setting of synchronization analysis, and does not violate the triangle inequality.

A further important feature of the Wasserstein distances can be seen in the left panel of Figure 2.9, where the distances for the two Hénon systems with equal parameters (but distinct, randomly realized initial conditions) are depicted. As the distances are calculated from (approximations of) invariant measures, these equivalent systems are close in behavior space either when (i) they are strongly coupled, but also (ii) when the coupling is minimal. The latter arises from the fact that the invariant measures of the two systems do not depend on the initial condition and are (theoretically) identical here. In between, for increasing coupling strengths the distances initially rise to about the four-fold value of the distance for $C = 0$, and then fall back to values comparable to the uncoupled case, from about $C = 0.7$ on.

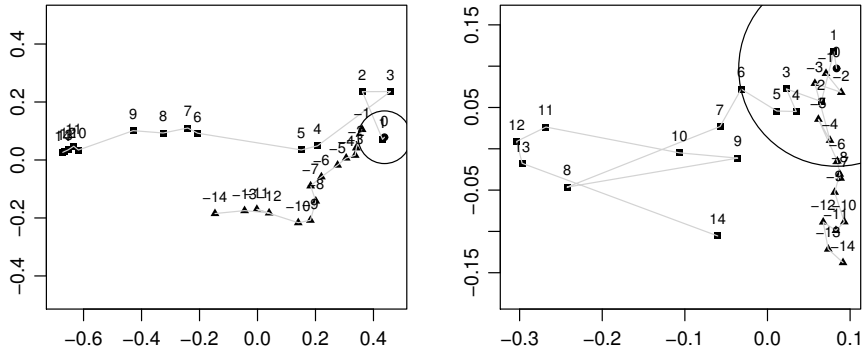


Figure 2.8: Two-dimensional MDS representation of Wasserstein distances for the Hénon system under parameter variation. Left panel: Parameter values as in Fig. 2.7. Right panel: Smaller range of parameter values (see text). Squares correspond to variation in the first parameter, triangles to variation in the second parameter. Numbers next to the symbols correspond to the indices of the dynamical systems introduced in the top axes of Fig. 2.7. The circles around the points corresponding to $a = 1.4$, $b = 0.3$ have radius 0.091 and 0.118, which are the mean self-distances.

If one interprets the distance between two systems as a measure of "synchronization" between them, this leads to the paradox, that in some cases (when C is less than roughly 0.5 here) an increased distance, usually indicative of "less synchronization", can actually be caused by an *increase* in coupling strength. Of course, this is just a variant of the usual fallacy that arises when one falsely assumes that statistical correlation between two variables does imply an underlying causal connection. This illustrates that one has to be very careful when drawing conclusions from synchronization measures (not only the one considered here) in practice.

The right panel of Figure 2.9 shows the case of two unequal Hénon systems, where the initial distances ($C = 0$) are positive and eventually decrease for stronger coupling. Interestingly, also in this case one sees the phenomenon that increasing coupling first results in a rise of the distances, that only decrease after a certain threshold in coupling is crossed. This can be interpreted as follows: Weak forcing by the master system does not force the behavior of the slave system to be closer to the forcing dynamics, rather the nonlinear slave system offers some "resistance" to the forcing (similar to the phenomenon of *compensation* in physiology). Only when the coupling strength is large enough to overcome this resistance does the slave dynamics become more similar to the master's (*decompensation*).

Figure 2.10 illustrates this phenomenon in behavior space, reconstructed by multidimensional scaling from the distances between the dynamics in the u-variables (the slave systems) only. The left panel, for equal systems, shows a closed curve, i.e., the dynamics of the slave systems is similar for both small and large coupling strengths. The right panel, for unequal systems, shows the occurence of the compensation/decompensation phenomenon in the curves of the right panel of Figure 2.9. Namely, the dynamics of the initially uncoupled slave system (point 1) settles for large coupling strengths at a different behavior (point 21). However, for small coupling strengths the behavior is perturbed *away* from this (points 2-6). If the coupling is increased further, a rapid transition (points 6-9) occurs. Note that this plot contains more information than Figure 2.9, as the information from all mutual distances (of the slave systems) is used, in contrast to the single distances between the master and slave dynamics depicted in the former.

Finally, Figure 2.11 shows the dependence of the Wasserstein distances between the master and slave systems for different bootstrap sizes. As expected, the (uncorrected) distances become lower when increasing the sample size. Interestingly, when correcting the distances, the distances become larger. This means that increasing the sample size increases the range of the (corrected) distances, i.e., their sensitivity.

2.5.5 Summary

By studying the behavior of the Wasserstein distances in the Hénon system, the following points have been observed (see also Box 2):

- In practice, when estimating distances by bootstrapping, the simulation error is almost normally distributed, due to the robustness of the Wasserstein distances. This justifies the use of statistical techniques with implicit assumptions of normality. It should also be possible to estimate the amount of inherent noise in the signals by artifical "noise titration".
- Due to the metric properties of the Wasserstein distances, numerical bifurcation analysis becomes possible. The distances between systems with varying parameter settings reflect the changes in their invariant measures and can help to pinpoint and track bifurcations. Hard bifurcations, e.g., when a stable periodic orbit becomes unstable, should result in detectable jumps in the distances.
- The Wasserstein distances can measure synchronization between two dynamical systems. However, being symmetric, they cannot provide information on the directionality of coupling. In general, one has to be careful when using similarities between dynamical systems as a measure of interaction strength, as independent systems with the same recurrent behavior will seem to be strongly coupled.

Box 2. Wasserstein distances of dynamical systems

- Wasserstein distances of dynamical systems are based on approximations of the invariant measure of the dynamics. This necessitates that only systems defined on the same phase space, i.e., determined by the same measurement process, can be compared.
- Due to the computational complexity of their calculation, Wasserstein distances usually need to be approximated by resampling estimates. Simulation error under such bootstrapping is almost normally distributed.
- Numerical bifurcation analysis is possible, with hard bifurcations resulting in visible jumps in the distances and the reconstructed point configurations.
- Wasserstein distances can quantify synchronization between two dynamical systems. However, independent systems with the same qualitative behavior will seem to be strongly coupled; a problem, that is common to most other synchronization measures.

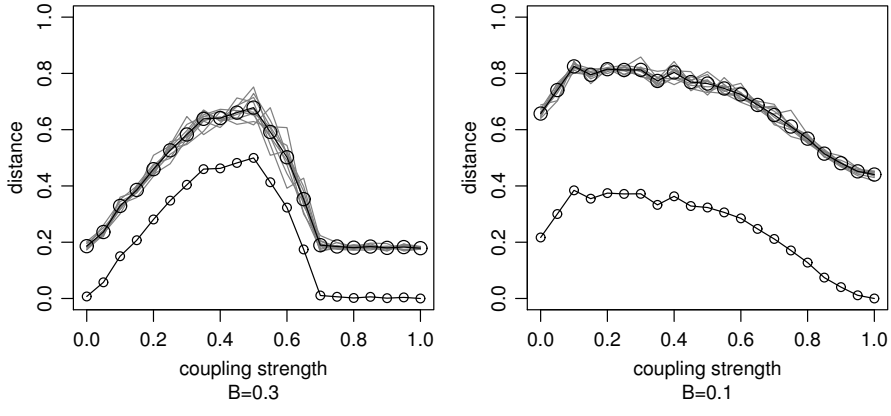


Figure 2.9: Distances for coupled Hénon systems (see text for details of the coupling). Coupling strength varied from $C = 0.0$ to $C = 1.0$ in steps of 0.05. Left panel: Equal Hénon systems ($B = 0.3$). Right panel: Distinct Hénon systems ($B = 0.1$). Top curves are uncorrected distances, the lower curves are corrected by subtracting the minimum distance encountered. Only the mean curve is depicted at the bottom.

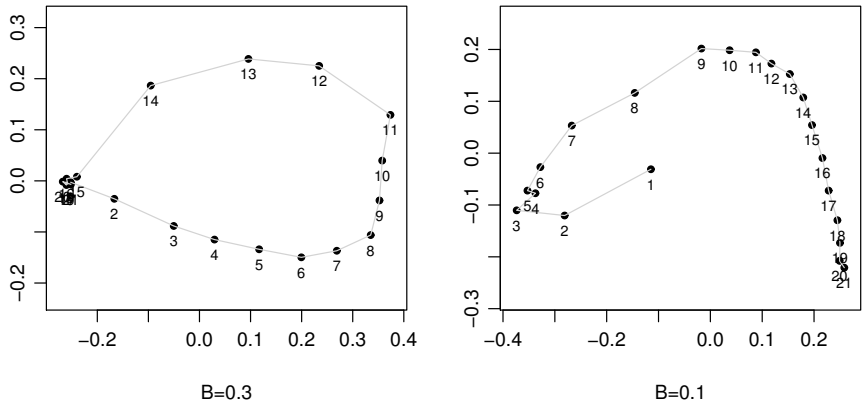


Figure 2.10: Two-dimensional MDS representation of Wasserstein distances for coupled Hénon systems. Coupling strength varied as in Figure 2.9. Left panel: Equal Hénon systems ($B = 0.3$). Right panel: Distinct Hénon systems ($B = 0.1$). The numbers next to the dots indicate the coupling strength, with larger number representing coupling strengths closer to 1.0. The points in the left panel constitute a closed curve, since the dynamics of strongly coupled equal Hénon systems is very similar to that of uncoupled equal Hénon systems. The points in the right panel show a different pattern, since the dynamical behavior of two distinct uncoupled Hénon systems is not similar, confer Fig. 2.9.

2.6 Example: Lung diseases[‡]

An interesting concept to connect dynamical systems and physiological processes is the notion of a *dynamical disease*, which was defined in a seminal paper (Mackey and Milton, 1987) as a change in the qualitative dynamics of a physiological control system when one or more parameters are changed (also see (Glass and Mackey, 1988; Beuter et al., 2003)). This allows to apply the methods of nonlinear science in a clinical context as well, and particularly the Wasserstein distances (Muskulus and Verduyn-Lunel, 2008a).

2.6.1 Background

Both asthma and the condition known as chronic obstructive pulmonary disease (COPD) are obstructive lung diseases that affect a large number of people worldwide, with increasing numbers expected in the future. In the early stages they show

[‡] This dataset is treated in a more sophisticated way in Chapter 3; here it is used to illustrate the methods for a real-world example.

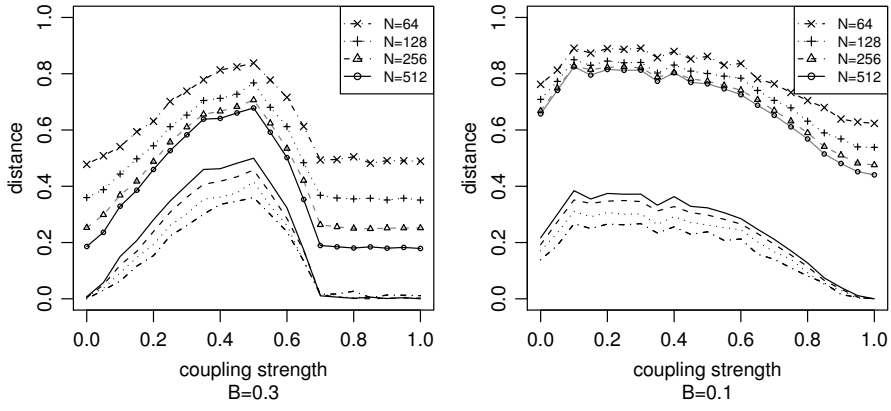


Figure 2.11: Distances for coupled Hénon systems for different bootstrap sizes. Coupling strength varied from $C = 0.0$ to $C = 1.0$ in steps of 0.05. Left panel: Equal Hénon systems ($B = 0.3$). Right panel: Distinct Hénon systems ($B = 0.1$).

similar symptoms, rendering correct diagnosis difficult. As different treatments are needed, this is of considerable concern.

An important diagnostical tool is the forced oscillation technique (FOT), as it allows to assess lung function non-invasively and with comparatively little effort (Oostveen et al., 2003). By superimposing a range of pressure oscillations on the ambient air and analyzing the response of the airway systems, a number of parameters can be estimated that describe the mechanical properties of airway tissue. In particular, for each forcing frequency ω, transfer impedance $Z(\omega)$ can be measured. This is a complex quantity consisting of two independent variables. The real part of $Z(\omega)$ represents airway *resistance* $R(\omega)$, and its imaginary part quantifies airway *reactance* $X(\omega)$, i.e., the elasticity of the lung tissue. Both parameters are available as time series, discretely sampled during a short period of tidal breathing. The dynamics of $R(\omega)$ and $X(\omega)$ are influenced by the breathing process, anatomical factors and various possible artifacts (deviations from normal breathing, movements of the epiglottis, etc.). Clinicians usually only use the mean values $\bar{R}(\omega)$ and $\bar{X}(\omega)$ of these parameters, averaged over the measurement period, but clearly there is a lot more (dynamical) information contained in these time series. Figure 2.12 shows example time series of these fluctuations for two patients, with the mean values indicated as horizontal lines.

It is well known that asthma usually results in increased values in both mean $R(\omega)$ and mean $X(\omega)$ compared to COPD (Lutchen et al., 1998). However, the values given in the literature are group means, and the parameters can fluctuate largely in

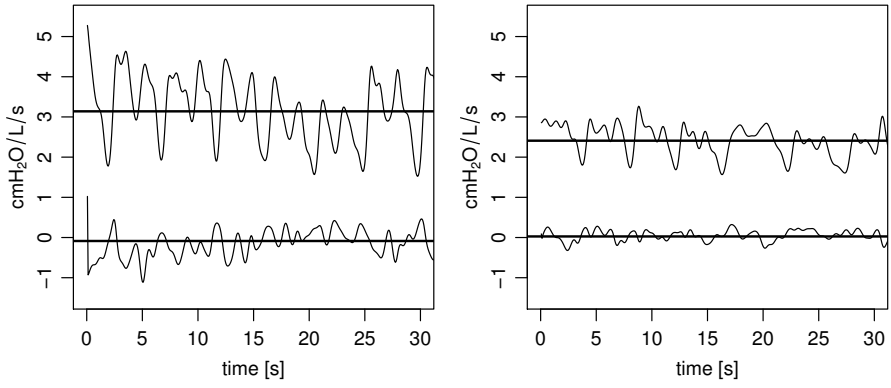


Figure 2.12: Example time series of respiratory resistance $R(8)$ (upper curves) and respiratory reactance $X(8)$ (lower curves) by forced oscillation technique during thirty seconds of tidal breathing. Left panel: A patient with mild asthma. Right panel: A patient with mild to severe chronic obstructive pulmonary disease. The horizontal lines indicate the mean values used routinely in clinical assessment.

individuals from the same group, usually with considerable overlap between the groups.

Figure 2.13 shows the distribution of mean resistance (left panel) and mean reactance (right panel) in a study conducted by A. M. Slats et. al (Slats et al., 2007), measured at 8 Hz forcing frequency. The solid curves show kernel density estimates of the distribution of mean values in the group of patients that suffer from mild, persistent asthma ($N_1 = 13$). The dashed curves show kernel density estimates in the group of patients suffering from mild to moderate COPD ($N_2 = 12$). Both resistances and reactances have been measured over a 1 minute interval of tidal breathing, repeated 12 times in the course of a few weeks. Ripples on top (asthma) and at the bottom of the plot (COPD) indicate the individual values of mean $R(8)$ and $X(8)$ per patient, and considerable overlap between the two classes of patients can be discerned. Note that, for these patients with mild asthma, the resistance values are actually lower (on the average) than the ones for the COPD group.

2.6.2 Discrimination by Wasserstein distances

The main motivation for the application of Wasserstein distances to this dataset is the assumption that the two lung diseases affect the temporal *dynamics* of transfer impedance in distinct ways, and not only its mean value. Considering asthma and

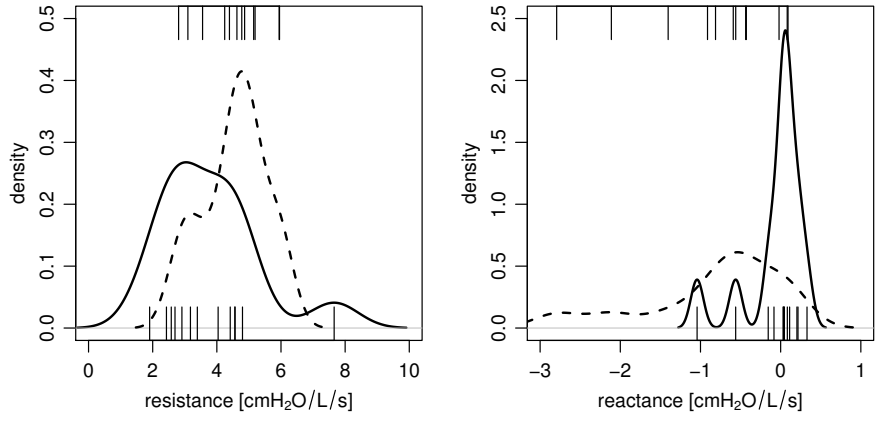


Figure 2.13: The distribution of time-averaged resistances $R(8)$ and reactances $X(8)$ in the dataset studied. The solid curves are kernel density estimates for the asthma group ($N_1 = 13$), the dashed curves show the corresponding estimates for the COPD group ($N_2 = 12$). Ripples at the bottom (asthma group) and top (COPD group) indicate the individual values of mean $R(8)$ and $X(8)$, respectively.

COPD as dynamical diseases, we assume an underlying dynamical systems with different parameters for the different diseases. Although these parameters are not accessible, it is then possible to discriminate the two diseases, with the Wasserstein distances quantifying the differences in the shape of their dynamics.

For simplicity, we only consider a two-dimensional reconstructing here, where the time series of $R(8)$ and $X(8)$ were combined into a series of two-dimensional vectors with trivial embedding dimension $k = 1$, trivial lag $q = 1$, and a length of about 12000 values (recorded at 16 Hz, the Nyquist frequency for the 8 Hz forced oscillation, concatenating all 12 measurements into one long series per patient). A more elaborated analysis will be presented elsewhere. Here we consider the distribution of these points in $\Omega = \mathbb{R}^2$ an approximation of the invariant measure of the underlying dynamical system.

The results for the squared sum of *differences*

$$d_{ij} = \left((\bar{X}_i(8) - \bar{X}_j(8))^2 + (\bar{R}_i(8) - \bar{R}_j(8))^2\right)^{1/2} \tag{2.36}$$

in means (*not* the Wasserstein distances), are shown in Figure 2.14. Panel A on the left shows a two-dimensional reconstruction of their behavior space by metric MDS. The strain plot in Panel B suggests an optimal reconstruction occurs in two dimensions, and indeed the classification confirms this. Although the maximal accuracy

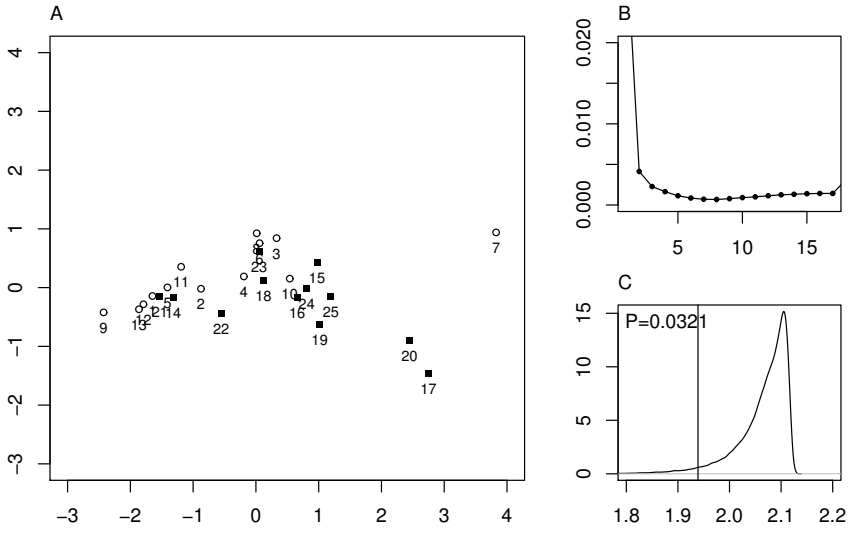


Figure 2.14: Results for distances in means (see text for details). Panel A: Two dimensional MDS reconstruction for patients suffering from asthma (open circles) and COPD (filled squares). The patient number, arbitrary assigned for comparison purposes, is shown below the symbols. Panel B: Strain values against reconstruction dimension. Panel C: MRPP statistic for the two classes. The value of δ for the labeling in panel A is indicated by the vertical line. The P-value is shown in the upper left corner.

of classification is 0.88 in a 11-dimensional reconstruction (i.e., 88 percent of the patients could be correctly classified), this drops to 0.72 in two dimensions when cross-validated. The separation of the two classes is significant at the 0.033 level, as indicated by the MRPP statistic in Panel C.

For comparison, the results for the Wasserstein distances W of normalized and centred data (to make the two parameters $R(8)$ and $X(8)$ comparable) are shown in Figure 2.15. These distances were bootstrapped 9 times for 250 sample points each. Here the separation of classes is much more pronounced, significant at the 0.0003 level. The classification is even perfect in a 12-dimensional reconstruction, with a maximal accuracy of 0.88 in a 9-dimensional reconstruction when cross-validated. Although the information about the means and their variance has been removed, the classification by Wasserstein distances is actually *better*. From this we conclude that the dynamical information contained in the fluctuations of respiratory impedance contains valuable clinical information. Note that these distances respect the triangle inequality (with a mean self-distance of about 0.25).

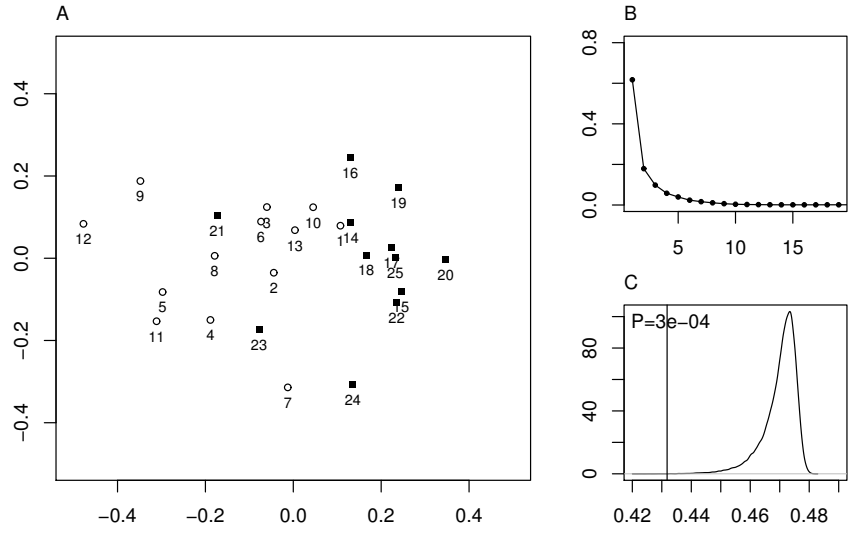


Figure 2.15: Results for Wasserstein distances W of normalized and centred data. Representation as in Fig. 2.14.

We have compared the results for the Wasserstein distances with a classification based on the *differences* in means (Eq. 2.36) employing the same distance-based methodology. If classical tests are used to classify these patients by their mean impedance values (e.g. in a tree-based classification), the classification results are even worse (not shown) than the ones we compare with the results obtained by the Wasserstein distances.

The above results show that the Wasserstein distances are able to capture differences of shape of the long-term behavior of real-world time series. Even for the trivial embedding shown and a low number of bootstrap samples, the Wasserstein distances allow to classify a large proportion of the lung diseases correctly. In fact, these are the best known classification results (at single FOT frequency) of these two lung diseases known to us. As the data have been centred before their calculation, the information about their mean values, which is usually used for classification, has been removed, so the classification is achieved by subtle dynamical differences instead.

2.7 Generalized Wasserstein distances

In this section we discuss a further generalization of the Wasserstein distances that addresses a particularly interesting issue in nonlinear time series analysis. We approach this problem from two sides.

Consider the following general problem: When comparing measurements of the same modality, but taken at different times or with different equipment, the question of comparability turns up. This also happens in applications to physiological data, where large variations can occur due to anatomical differences, when we want to compare data across subjects.

Detecting a change in the dynamics between two measurements is possible by simply centering and normalizing both time series before the analysis, but since it is a priori unknown whether differences in amplitude (standard deviation) and location (mean) of the time series are due to uncontrolled effects (noise, drift in the measurement apparatus, etc.) or due to a change in the dynamics, i.e., in the signal, this invariably leads to a loss in discriminating power. From the statistical point of view, although the empirical mean and variance are unbiased, consistent and effective estimators, they are not robust against outliers and non-stationary behavior. An interesting idea then is to transform the data in a data-driven way to partially account for such effects. By extending the notion of Wasserstein distance this can be done in a robust manner.

A second motivation for this comes from theoretical considerations. As remarked before, reconstruction by delay embedding results in an image of the attractor *up to a smooth change of coordinates*. This diffeomorphic change of coordinates is not accessible, as the underlying dynamical system is usually not known, only its projection by some measurement function (compare Fig. 2.1). In principle, only invariant, for example differential topological properties, can therefore be compared reliably between dynamical systems. Examples of such invariants include the number of fixed points or Lyapunov coefficients. In practice, however, one wants to use metric information to *quantitatively* compare dynamical systems on a much finer scale, as has also been done in this article.

Comparing a number of dynamical systems that are reconstructed in essentially the same way, i.e., by the same measurement function, it can be argued that the embeddings in reconstruction space, with its metric structure, *can* be compared, as essentially the same quantities (finite-difference approximations of derivatives, see (Packard et al., 1980)) are assessed. Nevertheless, it seems desirable to lessen the dependence of the Wasserstein distances on the particular embedding that is used.

In the following sections we discuss two complementary approaches to this problem: (i) Generalized Wasserstein distances (Section 2.7.1-2.7.7), and (ii) Nonmetric multidimensional scaling (Section 2.8).

2.7.1 Translation invariance

The first approach to the problem outlined before is to define *Wasserstein distances with respect to a class of global geometric transformations* as the minima of Wasserstein distances, optimized over the set of all possible transformations from a given class. From the statistician's viewpoint, this is similar to fitting a parametric transformation to data and then subjecting the transformed data to a distance analysis, and in the terminology of functional data analysis it can be considered a *registration* of the data (Ramsay and Silverman, 1997).

Considering a translation $\tau \in \Omega$, let μ_τ be the image of the measure μ under the transformation $x \mapsto x - \tau$.

Definition 1. The *Wasserstein distance with respect to translations* is given by

$$W^t(\mu, \nu) = \inf_{\tau \in \Omega} W(\mu_\tau, \nu) = \inf_{\tau \in \Omega} W(\mu, \nu_{-\tau}). \tag{2.37}$$

The following shows that this is indeed well-defined:

Proposition 1. $W^t(\mu, \nu)$ is a distance on the space of probability measures.

Proof. We have to check the three properties of a distance. Reflexivity $W^t(\mu, \mu) = 0$ and symmetry $W^t(\mu, \nu) = W^t(\nu, \mu)$ are obvious from the definition and the corresponding properties of the Wasserstein distance W. For the triangle inequality, consider three measures μ, ν and ρ. Assume that the distance $W^t(\mu, \rho)$ is realized by some translation τ_1, and that $W^t(\rho, \nu)$ is realized for some translation τ_2. Then

$$\begin{aligned} W^t(\mu, \rho) + W^t(\rho, \nu) &= W(\mu_{\tau_1}, \rho) + W(\rho, \nu_{-\tau_2}) \\ &\geq W(\mu_{\tau_1}, \nu_{-\tau_2}) = W(\mu_{\tau_1 + \tau_2}, \nu) \\ &\geq \inf_{\tau \in \mathbb{R}^k} W(\mu_\tau, \nu) = W^t(\mu, \nu), \end{aligned} \tag{2.38}$$

where we use the triangle inequality for the Wasserstein distances (Clement and Desch, 2008). □

The Wasserstein distance with respect to translations is obviously *invariant* under this class of transformations: If the data is shifted before its calculation, the value of the distance does not change.

Note that the values obtained are usually different from the Wasserstein distances of normalized data, as can be seen in Fig. 2.16, which is based on two realizations of $\mathcal{N}(0, 1)$ random variables. The minimum of the Wasserstein distance is attained in an interval to the left of the point where the empirical means coincide. In particular, the translation for which the minimal distance is realized is not unique. This nonuniqueness (also of transportation plans) is a special feature of the Kantorovich-Rubinstein distances, but for larger sample sizes, and especially in more than one dimension,

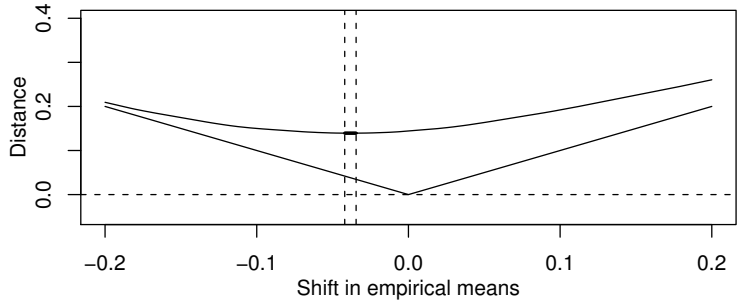


Figure 2.16: Translated Wasserstein distances for two normalized realizations ($n = 50$) of a $\mathcal{N}(0, 1)$ random variable. Lower solid curve: distance between empirical means against translation. Upper solid curve: Wasserstein distance $W(\mu_\tau, \nu)$. The minimum, attained in the indicated interval (dotted lines), is the Wasserstein distance with respect to translations, $W^t(\mu, \nu)$.

the area where the optimum is attained becomes very small, in fact smaller than the typical numerical accuracy. Moreover, the value of $W^t(\mu, \nu)$ itself is clearly unique.

In the example in Figure 2.16 the difference between $W(\mu, \nu)$ and $W^t(\mu, \nu)$ is of a statistical nature and due to the small sample size. For larger sample sizes the two values indeed converge against each other. In general however, i.e., when the measures have non-vanishing higher-order (beyond two) statistical moments, there will be a finite difference between the two values for *all* sample sizes.

To summarize: the optimization involved in the computation of $W^t(\mu, \nu)$ has the two-fold effect of (i) finding a more robust alternative to the center of mass, similar to the geometric median (Fermat-Weber point), and (ii) because of this, the distance $W^t(\mu, \nu)$ stresses the information on the higher-order moments present in the data, i.e., on the differences in the shape of the two measures involved.

2.7.2 Rigid motions

Considering rotations $\Theta \in \mathrm{SO}(k)$, where $\mathrm{SO}(k)$ is the special orthogonal group of $\mathbb{R}^k$ consisting of all rotation matrices, a center point is needed against which to rotate. For finite point sets, it has been shown by Klein and Veltkamp (Klein and Veltkamp, 2005) that the only sensible choice for the center point is the mean $m(\mu)$,

$$m(\mu) = \int_\Omega ||x||_2 \, \mathrm{d}\mu[x]. \tag{2.39}$$

Accordingly, we define a rotation as the image of μ under the map

$$x \mapsto \Theta(x - m(\mu)) + m(\mu), \tag{2.40}$$

denoted by $\Theta \cdot \mu$. We assume in the following that both measures are centred, such that $m(\mu) = m(\nu) = 0$. Then $W(\Theta \cdot \mu, \nu) = W(\mu, \Theta^{-1} \cdot \nu)$ holds, and the following is well-defined:

Definition 2. The *Wasserstein distance with respect to rigid motions* is

$$W^r(\mu, \nu) = \inf_{\Theta \in \mathrm{SO}(k)} \inf_{\tau \in \Omega} W((\Theta \cdot \mu)_\tau, \nu). \tag{2.41}$$

Note that only one rotation is needed, and that the translation is applied last, as that makes its interpretation easier (the alternative would be to take the infimum over $W(\Theta \cdot \mu_\tau, \nu)$).

Proposition 2. $W^r(\mu, \nu)$ is a distance on the space of probability measures.

Proof. Reflexivity and symmetry are obvious again. For the triangle inequality, consider three measures μ, ν and ρ. Assume that the distance $W^r(\mu, \rho)$ is realized by some translation τ_1 and rotation Θ_1, and that $W^r(\rho, \nu)$ is realized for some translation τ_2 and rotation Θ_2. Then

$$\begin{aligned} W^r(\mu, \rho) + W^r(\rho, \nu) &= W((\Theta_1 \cdot \mu)_{\tau_1}, \rho) + W((\Theta_2 \cdot \rho)_{\tau_2}, \nu) \\ &= W((\Theta_1 \cdot \mu)_{\tau_1}, \rho) + W(\rho, \Theta_2^{-1} \cdot \nu_{-\tau_2}) \\ &\geq W((\Theta_1 \cdot \mu)_{\tau_1}, \Theta_2^{-1} \cdot \nu_{-\tau_2}) \\ &= W((\Theta_2 \Theta_1 \cdot \mu)_{\Theta_2 \tau_1 + \tau_2}, \nu) \\ &\geq \inf_{\Theta \in \mathrm{SO}(k)} \inf_{\tau \in \mathbb{R}^k} W((\Theta \cdot \mu)_\tau, \nu) = W^r(\mu, \nu). \end{aligned} \tag{2.42}$$

□

2.7.3 Dilations and similarity transformations

An important further class of transformations are the *dilations* where $\lambda > 0$ is a scale parameter. Again a center point is needed against which to scale, and the mean $m(\mu)$ is the natural choice (Klein and Veltkamp, 2005). A dilation is the image of μ under the map $x \mapsto \lambda x$, denoted by $\lambda\mu$.

A number of problems are encountered when working with dilations, though, as these transformations do not respect the distance properties in general. For a start, to preserve the symmetry of the Wasserstein distances, we either need to consider $W(\lambda\mu, \lambda^{-1}\nu)$ or $W(\lambda\mu, (1-\lambda)\nu)$. As λ is bounded in the second case, we prefer the latter. Recall that the two measures are centred, such that $m(\mu) = m(\nu) = 0$. We furthermore assume that the measures are normalized, such that their second moments satisfy

$$m_2(\mu) = \left(\int_\Omega ||x - m(\mu)||_2^2 \, \mathrm{d}\mu[x] \right)^{1/2} \stackrel{!}{=} 1. \tag{2.43}$$

This sets a common global scale for these distances and allows to compare them between different datasets (respecting the Caveats discussed at the beginning of Section 2.7).

Definition 3. The *Wasserstein "distance" under similarity transformations* is

$$W^s(\mu,\nu) = \inf_{\lambda>0} \inf_{\Theta\in \mathrm{SO}(k)} \inf_{\tau\in\Omega} W((\lambda\Theta\cdot\mu)_\tau, (1-\lambda)\nu). \tag{2.44}$$

Note that both measures are transformed reciprocally by λ, since otherwise (in case we would define the distance to be $W((\lambda\Theta\cdot\mu)_\tau, \lambda\nu)$, for example) the optimum would be achieved by shrinking both measures to single points, i.e., in the limit as $\lambda \to 0$. The above definition prevents this: if μ is shrunk ($\lambda < 1/2$), the measure ν is expanded (as $1-\lambda > 1/2$), and vice versa. The translation is again applied last, as that makes its interpretation easier.

Unfortunately, it is not clear when $W^s(\mu,\nu)$ is truly a distance, i.e., under which conditions the triangle inequality holds. In general, therefore, one has to be careful when using the "distance" W^s. In Section 2.3.4 we have discussed how these violations of metric properties can be corrected. Because of this, Eq. 2.44 presents us still with a potentially useful notion of distance, and as W^s might be interesting for certain applications, we include W^s when we talk about the *(generalized) Wasserstein distances*.

2.7.4 Weighted coordinates

Although one elegant property of the delay vector construction is the fact that each coordinate has the same statistical distribution (disregarding effects due to the finiteness of the underlying time series), there are many applications where two or more scalar time series are available, indeed necessary, for a reconstruction. The simplest way to accomodate this is by assigning distinct coordinates of the delay vectors to different time series. For example, if we are given two time series

$$x^{(1)} = (x_1^{(1)}, \ldots, x_N^{(1)}), \quad x^{(2)} = (x_1^{(2)}, \ldots, x_N^{(2)}) \tag{2.45}$$

of N measurements, in the simplest case the underlying dynamical system is reconstructed by mapping each consecutive block

$$\begin{aligned} x_{[i]} = &(x_i^{(1)}, x_{i+q_1}^{(1)}, \ldots, x_{i+(k_1-1)q_1}^{(1)}, \\ &(x_i^{(2)}, x_{i+q_2}^{(2)}, \ldots, x_{i+(k_2-1)q_2}^{(2)}) \end{aligned} \tag{2.46}$$

to a single point in $\Omega = \mathbb{R}^{k_1+k_2}$ (see (Cao et al., 1998) for generalizations and more advanced techniques).

Here the question of comparability turns up again. The usual solution is to normalize all time series involved (as has been done in Section 2.6), but again we can alternatively employ the Wasserstein distances in order to achieve this in a robust way. Let us consider the generalization of the usual Euclidean distance to a *weighted Euclidean distance*,

$$d_\alpha(x,y) = \left(\sum_{i=1}^{k} \alpha_i |x_i - y_i|^2 \right)^{1/2}. \tag{2.47}$$

Here $\alpha \in \mathbb{R}^k_+$ is a vector of positive weights, normalized such that $||\alpha||_1 = \sum_{i=1}^n \alpha_i = k$.

Definition 4. Given a Wasserstein distance $W^*(\mu,\nu;d)$ (possibly with respect to some class of transformations) between two measures μ and ν over $\mathbb{R}^k$, with a Euclidean distance function d, the *weighted Wasserstein distance (with respect to the same class of transformations)* is

$$W_\alpha(\mu,\nu;d) = \inf_{\substack{\alpha\geq 0,\\ ||\alpha||_1 = k}} W(\mu,\nu;d_\alpha). \tag{2.48}$$

Restricting the weights further, such that α_i is constant for all coordinates arising from the same original time series, leads to a useful notion of distance. In the above example of two time series this means the following requirement:

$$\alpha_i = \begin{cases} \alpha_1 & \text{if } \ 1 \leq i \leq k_1 \\ \alpha_2 & \text{if } \ k_1 < i \leq k_1 + k_2, \end{cases} \tag{2.49}$$

with the obvious generalization to more than two time series.

Note that again it is not clear under which conditions the triangle inequality holds for weighted Wasserstein distances, but for the same reasons as in Section 2.7.3 this does not pose a cause for much concern.

2.7.5 Residuals of Wasserstein distances

In the previous sections we have seen a few examples of classes of transformation for which Wasserstein distances can be optimized. There are obviously many more, but the advantage of the three classes considered above (translations, rigid motions and similarity transformations) is that they are successively allow more freedom. Their respective Wasserstein distances thus form a natural hierarchy:

Proposition 3. The generalized Wasserstein distances satisfy

$$W^s(\mu,\nu) \leq W^r(\mu,\nu) \leq W^t(\mu,\nu) \leq W(\mu,\nu). \tag{2.50}$$

An easy calculation, similar to the one for the discrete case (Rubner et al., 2000), furthermore shows that untransformed Wasserstein distances are bounded from below by the distance in mean,

$$W(\mu,\nu) \geq ||m(\mu) - m(\nu)||_2, \tag{2.51}$$

confer Fig. 2.16.

Eq. 2.51 suggests that we center the measures for which the Wasserstein distances are calculated. The definition of rotations and dilations suggests that we also normalize the measures. The monotonicity property of the Wasserstein distances for the classes of transformations in Eq. 2.50 then ensures that the following is well-defined:

Definition 5. Given two normalized and centred measures μ and ν on the same probability space, the *residual of the Wasserstein distance with respect to translations* is

$$R^t(\mu,\nu) = W(\mu,\nu) - W^t(\mu,\nu). \tag{2.52}$$

The *residual of the Wasserstein distance with respect to rigid motions* is

$$R^r(\mu,\nu) = W^t(\mu,\nu) - W^r(\mu,\nu). \tag{2.53}$$

The *residual of the Wasserstein distance with respect to similarities* is

$$R^s(\mu,\nu) = W^r(\mu,\nu) - W^s(\mu,\nu). \tag{2.54}$$

Again, these *residuals* are usually not distances. Nevertheless, due to the nonlinearity inherent in the definition of the generalized Wasserstein distances, these residuals quantify differences in higher order moments of probability measures, i.e., in their shape. However, contrary to moment or multipole expansions, each distance in the sequence of (residual) distances

$$(W(\mu,\nu), R^t(\mu,\nu), R^r(\mu,\nu), R^s(\mu,\nu)) \tag{2.55}$$

measures a complex interplay of all higher order moments.

2.7.6 Optimization of generalized cost

Optimizing the Wasserstein distances over a class of transformations is straightforward. The Nelder-Mead simplex algorithm (Nelder and Mead, 1965) is a simple, but reliable algorithm that only uses function evaluations and does not need gradient information. We have found that it works reasonably well in practice. If more control is required, in the distance case (e.g. when considering translations only) it is

possible to show that the generalized Wasserstein distances fulfill a Lipshitz condition, and global Lipshitz optimalization (Mladineo, 1986) is then an interesting, but slower, alternative.

To parametrize rotations in k-dimensional space, we use the Lie algebra $\mathrm{so}(k)$ of the group of rotations $\mathrm{SO}(k)$. This is the algebra consisting of all k-by-k skew-symmetric matrices, which is described by $k(k-1)/2$ independent parameters. Exponentiation results in a parametrization of the rotations, i.e., if $A \in \mathrm{so}(k)$ then $\exp(A) \in \mathrm{SO}(k)$, and the image of $\mathrm{so}(k)$ under exponentiation is precisely the group of all rotations (since $\mathrm{SO}(k)$ is connected and compact, see for example (Frankel, 1997)). The function $\exp(A)$ is of course the matrix exponential of A (consult (Moler and Loan, 1978) for implementation issues).

2.7.7 Example: The Hénon system

Continuing the example of Section 2.5.1, where the Hénon map was discussed in terms of self-distances, the results for the generalized Wasserstein distances are shown in Figure 2.17. These distances have been optimized with respect to rigid motions, where a maximum of 3000 distance evaluations was imposed in the Nelder-Mead algorithm (which was never exceeded), stopping the iterations if there was no improvement of relative size 10^{-6}. Initial rotation parameters in the Lie algebra $\mathrm{so}(k)$ were set to 1.0 to avoid the simplex search to end up in a possible local minimum around the trivial rotation. Of course, in a more sophisticated application, the optimization should actually be performed a number of times with different initial parameters.

The slope of these distances is much lower than for the distances in Figure 2.4, and levels out for larger embedding dimensions. This is an indication that the generalized Wasserstein distances do not suffer from statistical error as much as the untransformed distances do. In fact, for embedding dimensions larger than two, the generalized distances have comparable values; this is an indication that the attractor of the system has been properly unfolded. Increasing the embedding dimension beyond three does not significantly improve the quality of the embedding, as quantified by these self-distances. Note that these distances are calculated for normalized data, such that they cannot be directly compared with the untransformed distances in Figure 2.4.

2.8 Nonmetric multidimensional scaling

As a second approach to the problem addressed in the beginning of Section 2.7, we mention the use of *nonmetric* (or ordinal) multidimensional scaling. An exposition of this technique can be found in (Borg and Groenen, 2005, Chapter 9). It is complementary to the approach of Section 2.7. Instead of transforming the data to achieve more

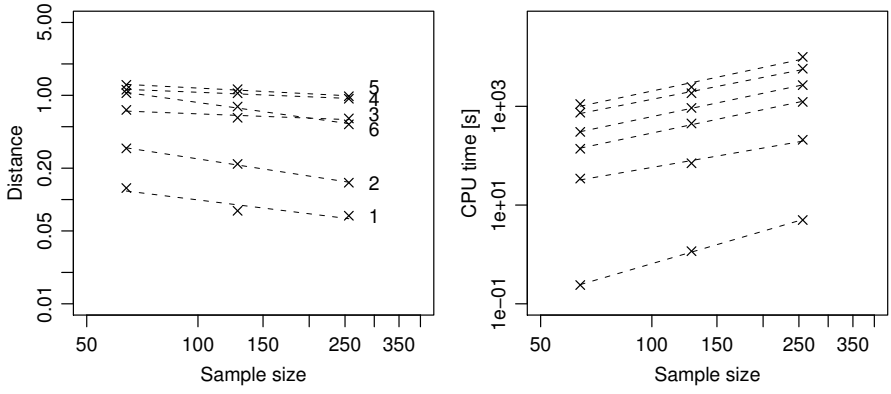


Figure 2.17: Dependence of estimated self-distances on sample size. Left panel: Generalized Wasserstein distances (with respect to rigid motions) for embedding dimensions 1 to 6. The embedding dimension has been indicated on the right side of the regression lines. Right panel: CPU time needed for these calculations, with a slope of roughly $3/2$ (for dimensions greater than two).

natural distances, the basic idea in this approach is that we *transform the distances*. The admissible transformations are those that preserve the (rank-) order of the distances. Thereby, the impact of the particular metric structure of phase space, and consequently of the delay embedding, is reduced considerably. On the other hand, topological properties of the systems in behavior space (i.e., the "relative closeness" of their dynamical behaviors) is preserved by such transformations. The simplest transformation is the rank transformation, in which the totality of $\frac{1}{2}n(n+1)$ distances (the entries of an n-by-n distance matrix) are sorted according to their size, and replaced by the corresponding rank numbers from $\{1, 2, \ldots, \frac{1}{2}n(n+1)\}$. Unfortunately, the rank transformation does in general not respect the triangle inequality, and does not result in a suitable reconstruction of behavior space.

We see that the prize to pay for this generalization is the complexity of reconstruction of behavior space, which cannot be calculated by way of a simple matrix decomposition as in metric MDS. Instead, one needs to use an iterated optimization algorithm, that tries to minimize a given error functional. Instead of strain (Eq. 2.24), this is usually taken to be the *stress*,

$$\sigma(X) = \sum_{i<j} (||X_i - X_j|| - \delta_{ij})^2, \tag{2.56}$$

where δ_{ij} are the transformed dissimilarities of the measured distances M_{ij}, and

the X_i are the coordinates of the i-th system in behavior space. Then one can use *monotone regression* (Kruskal, 1964) to iteratively move the reconstructed points X, minimizing the stress value (see Meulman et al. (1983) for a possible approach).

Not only is this computationally much more involved, but two main problems are encountered: (i) it is possible for the optimization algorithm to return a local minimum of stress, and (ii) degenerate solutions can exist. Nevertheless, this is a viable alternative to Wasserstein distances under transformations. Depending on the situation, one or the other approach, or even a combination of the two, should be considered.

2.9 Conclusions

We have discussed Wasserstein distances in the context of dynamical systems and time series, with a focus on the statistical analysis of the resulting distances. This point has been neglected in the literature so far, which probably explains why the Wasserstein (or transportation) distances are not as well known as they deserve to be. Possible applications of the distance-based analysis of dynamical systems include the classification and discrimination of time series, the detection and quantification of (generalized) synchronization and the visualization and quantification of parameter changes and bifurcations. More generally, the behavior space introduced allows to apply the tools of multivariate statistical analysis, and to test statistical hypotheses about dynamical systems.

Due to their elegant definition and natural properties, e.g., their interpolation properties (see Villani (2003)), Wasserstein distances are very interesting from the theoretical side. However, their estimation in practice is time-consuming and one usually needs to resort to various approximations, e.g. the bootstrapping of the distances. By way of a few examples of synthetic and real-world data sets we have shown that this is quite feasible. This should convince the reader of the utility of the Wasserstein distances and, more generally, of the distance-based analysis of dynamical systems (different applications, e.g., in texture analysis, are hinted at in (Muskulus, Scheenstra, Braakman, Dijkstra, Verduyn-Lunel, Alia, de Groot and Reiber, 2009); applications in electrophysiology, both for sensor recordings and for phase distributions, are discussed in (Muskulus, Houweling, Verduyn-Lunel and Daffertshofer, 2009)). The algorithms used to derive the results of this paper are available from the first author's homepage.

Various questions remain, however. In particular, one would like to (i) understand better how to lessen the dependence of the Wasserstein distances on the particular embedding used, a point that was introduced and discussed in Section 2.7, and (ii) address the point how various distinct Wasserstein (and possibly other) distances can be combined in an analysis of dynamical systems or time series (see the seminal article (Xu et al., 1992) for first steps in this directions). In this article we

have exclusively used the Kantorovich-Rubinstein (or Earth Mover's) distance, but the class of Wasserstein distances encompasses other distances (e.g. total variation) that test different properties of the shape of the invariant measures under study. Combining more than one distance measure should improve the analysis, e.g., the classification of dynamical systems. The details of this last point are postponed to future work.

Applications

Chapter 3

Lung diseases

Medicine is not only a science; it is also an art. It does not consist of compounding pills and plasters; it deals with the very processes of life, which must be understood before they may be guided.

Paracelsus

In this chapter we apply the distance-based analysis to experimental time series obtained from patients that suffer from two lung diseases, as well as healthy controls. Section 3.1 offers background information on the respiratory system. Section 3.2 introduces the forced oscillation technique that was used to obtain the time series. Section 3.3 describes the data in more detail. Section 3.4 is a digression in which two methods of fluctuation analysis are introduced, power-law and detrended fluctuation analysis. Section 3.5 discusses the nonlinear analysis techniques used, including (sample) entropy and Wasserstein distances. Experimental results are presented in Section 3.6 and discussed in Section 3.7, where also clinical implications and future directions are outlined.

3.1 Respiration

Human respiration is a complex phenomenon that is influenced and controlled by diverse factors. Physically, respiration is simply the movement of air through the airways due to differences between pleural pressure and the pressure of the surrounding air, which are created by movements of the pleura and the ribs. The geometry of the airways is intricate, however: already between the opening of the mouth and the main trachea the volume is quite variable, and the air needs to pass the pharynx, epiglottis and larynx before beginning its voyage through the lungs. There, the main trachea branches many times into successively smaller generations of bronchi and bronchioles until reaching the alveoli through the acini. This hierarchical branching greatly increases the surface of the lung. Although consisting of a finite number of levels (usually ~ 25), it is not uncommon to consider the branching of the airways a prime example of self-similarity in the physical world, and fractal descriptions of the lung offer explanations of its efficiency (Weibel, 1963, 1991).

In the alveoli, diffusion of gases removes carbon dioxide from venuous blood and transports oxygen across the respiratory membrane into the capillaries. This transport is modulated by cardiac status and posture, causing local inhomogeneities

in the ventilation-perfusion ratio. The upper part of the lung usually suffers from a moderate physiologic deadspace due to increased hydrostatic pressure, and the lower part of the lung usually exhibits too little ventilation, leading to a moderate physiologic shunt (Guyton and Hall, 2006).

The rate of respiration is regulated in the central nervous system (CNS). The primary respiratory rhythm is generated from bursting inspiratory neuronal action potentials in the brain stem that are subsequently modulated and filtered. Since the hemoglobin-oxygen system buffers the amount of oxygen delivered to tissues, the respiratory drive is mainly regulated by carbon dioxide chemoreceptors; oxygen receptors in the peripheral chemoreceptor system play a role when sufficient arterial oxygen levels cannot be sustained. Interestingly, under exercise, when oxygen demand increases to a multiple of normal values, the main adaptation of respiratory drive seems to be caused by anticipatory signals from muscles, and the chemical receptor-loops are only used for fine control.

Respiration can and has been described on roughly four distinct levels. First of all, there is the *mechanical* act of breathing, i.e., the geometric and mechanical properties of the channels and openings through which the air passes. Secondly, the actual gas exchange by diffusion is a chemical transport phenomenon. Thirdly, the total cardiorespiratory system influences respiration through heart rate, blood pressure and the ensuing dynamic shunting phenomena, that are related to body posture and other systemic properties. Lastly, this system is driven centrally by a complicated regulatory system that is influenced not only by physiological factors and various signalling systems, but also by cognitive state and environmental influences. At each of these levels of description there exist mathematical models that try to capture the essential properties of the observed phenomena, and also integrative approaches that try to model interactions between the various levels of description.

Here we will be mainly interested in the mechanical properties of the airways, which are, however, modulated by all the above mentioned influences. In fact, our prime interest is to use easily measured mechanical properties as markers of more complex changes in the underlying physiological control systems. In particular, we will focus on breathing with a diseased lung.

Box 3. The main questions

- To what extent are mechanical properties of breathing changed in diseased lungs?
- How can this be used to assess airways and disease status?

Subject populations		
A	Asthma	
C	COPD	
H	Healthy	
Measurements		
R_{rs}	Respiratory resistance	$R_{\mathrm{rs}} = \mathrm{Re}\, Z_{\mathrm{rs}}$
X_{rs}	Respiratory reactance	$X_{\mathrm{rs}} = \mathrm{Im}\, Z_{\mathrm{rs}}$
Z^*_{rs}	Complex respiratory impedance	$Z^*_{\mathrm{rs}} = R_{\mathrm{rs}} + iX_{\mathrm{rs}}$
Z_{rs}	Respiratory impedance (gain)	$Z_{\mathrm{rs}} = \lvert R_{\mathrm{rs}} + iX_{\mathrm{rs}} \rvert$
Z_{var}	Squared residuals of Z_{rs}	$Z_{\mathrm{var}}(i) = (Z_{\mathrm{rs}}(i) - \bar{Z}_{\mathrm{rs}})^2$
$\ln Z_{\mathrm{rs}}$	Natural logarithm of Z_{rs}	$\ln Z_{\mathrm{rs}} = \log(Z_{\mathrm{rs}})$
$\ln Z_{\mathrm{rs}}\mathrm{SD}$	Standard deviation of $\ln Z_{\mathrm{rs}}$	$\ln Z_{\mathrm{rs}}\mathrm{SD} = \mathrm{SD}(\log(Z_{\mathrm{rs}}))$

Table 3.1: Abbreviations used in this chapter.

Subscript	
ao	airways
cw	chest wall
eq	equivalent (to the value of a lumped, single-compartment model)
in	input (forcing at the mouth)
pl	pleural
rs	respiratory system
tr	transfer (forcing at the chest wall)

Table 3.2: Subscripts used in this chapter.

3.2 The forced oscillation technique

Th forced oscillation technique (FOT) measures the mechanical properties of lung tissue noninvasively and continuously. A pressure oscillation is superimposed on the air, resulting in a longitudinal pressure wave travelling through lung tissue and back, during which its amplitude and phase are modulated relative to the mechanical properties of the respiratory system. These properties are expressed in quantities characteristic of fluid dynamics (Herman, 2007):

- *Resistance* is the pressure difference ΔP needed to cause a given flow rate $Q = \dot{V}$,

$$R = \frac{\Delta P}{Q}, \tag{3.1}$$

  and is usually measured in units of $\mathrm{cmH_2O\,s/L}$.

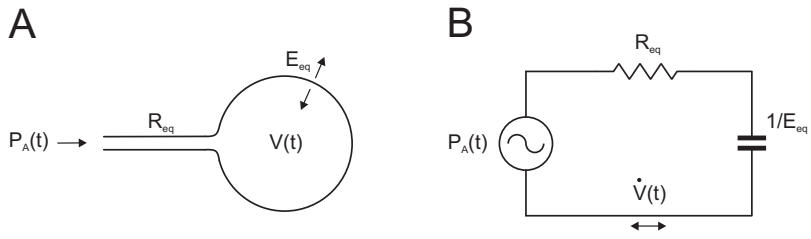


Figure 3.1: Simple compartment model fitted in single frequency FOT. A: Mechanical model of a single airway with resistance R_{eq} and elasticity E_{eq}. B: Equivalent electrical circuit.

- *Complicance* is the change in volume caused by pressure changes in an elastic airway,
$$C = \frac{\Delta V}{\Delta P}, \tag{3.2}$$
and is measured in units of $\mathrm{L/cmH_2O}$. Its inverse, $E = 1/C$ is called *elastance* and is a measure of the airways's rigidity.
- *Inertance* is the change in pressure caused by a change in flow rate,
$$L = \frac{\Delta P}{\Delta Q}, \tag{3.3}$$
and usually given in units of $\mathrm{cmH_2O\,s^2/L}$.

By using a small amplitude oscillation (typically about $\pm 1\mathrm{cmH_2O}$), the respiratory system can be considered a linear time-invariant (LTI) system, and its total frequency response (transfer function) at frequency f is

$$Z^*_{\mathrm{rs}}(f) = \frac{P(f)}{Q(f)}, \tag{3.4}$$

where $P(f)$ and $Q(f)$ are the Fourier transforms of the pressure and flow signal, respectively. Although respiratory impedance is a complex quantity, it is common in the literature to refer to its magnitude by the same name. To avoid confusion, we denote respiratory impedance by Z^*_{rs} and reserve $Z_{\mathrm{rs}} = |Z^*_{\mathrm{rs}}|$ for its magnitude. The expression *respiratory impedance* will in the following refer to Z_{rs}. The real and imaginary parts of the *complex respiratory impedance* Z^*_{rs} are called *resistance* R_{rs} and *reactance* X_{rs},

$$Z^*_{\mathrm{rs}}(f) = \mathrm{Re}\, Z^*_{\mathrm{rs}}(f) + i\,\mathrm{Im}\, Z^*_{\mathrm{rs}}(f) = R_{\mathrm{rs}}(f) + iX_{\mathrm{rs}}(f), \tag{3.5}$$

and can alternatively be expressed by a real-valued *gain* Z_{rs} and *phase angle* φ_{rs},

$$R_{\mathrm{rs}}(f) + jX_{\mathrm{rs}}(f) = Z_{\mathrm{rs}}(f)e^{i\varphi_{\mathrm{rs}}(f)}. \tag{3.6}$$

Assuming negligible inertance, they represent equivalent mechanical resistance and elastance of a single compartment model under periodic forcing (Figure 3.1), where $R_{\mathrm{eq}}(f) = R_{\mathrm{rs}}(f)$ and $E_{\mathrm{eq}}(f) = -X_{\mathrm{rs}}(f)/(2\pi f)$. Under this abstraction, the respiratory system is described by the first-order model

$$P_A(t) - P_0 = R_{eq}\dot{V}(t) + E_{eq}V(t) \tag{3.7}$$

with baseline pressure P_0 and harmonic forcing $P_A(t) = A\sin 2\pi ft$.

In practice, two main protocols are used to generate the forcing. Input forcing at the mouth[1] results in measurement of *input impedance* $Z^*_{\mathrm{in}}(f) = P_{\mathrm{ao}}(f)/\dot{V}_{\mathrm{ao}}(f)$. Input forcing at the chest results in the measurement of *transfer impedance* $Z^*_{\mathrm{tr}}(f) = P_{\mathrm{cw}}(f)/\dot{V}_{\mathrm{ao}}(f)$. The latter has been shown to be more reliable in separating different airway and tissue components and is also less sensitive to upper airway shunting (Lutchen et al., 1998), but more difficult to measure, as total body plethysmography is needed. Input impedance, on the other hand, is easy to measure and well-tolerated, rendering it the most viable for routine assessments of lung function.

Until recently, only average values of Z^*_{rs} were used. Real-time tracking of single frequency FOT signals is possible, however, either through a recursive least-squares algorithm in the time domain (Avanzolini et al., 1997), or through the use of windowed Fourier transforms. In the latter approach, one usually assumes that the true pressure p_i and flow q_i, sampled at finite time points ($i = 1, 2, \ldots,$N), are subject to white noise errors ϵ_i^P, ϵ_i^Q in the frequency-domain,

$$P_i(f) = q_i(f)Z_{\mathrm{rs}}(f) + \epsilon_i^P \tag{3.8}$$

$$Q_i(f) = p_i(f)/Z_{\mathrm{rs}}(f) + \epsilon_i^Q. \tag{3.9}$$

Estimation of $p_i(f)$ and $q_i(f)$ from the observed spectral components $P_i(f)$ and $Q_i(f)$ then becomes possible by a total least squares approach (Slats et al., 2007, online supplement). The gain and phase angle are thereby estimated from the windowed spectral power estimates $\hat{S}^2_Q, \hat{S}^2_P$ and the cross-power estimate $\hat{S}_{PQ}$ as

$$Z_{\mathrm{rs}} = \sqrt{\hat{S}^2_{\mathrm{P}}(f)/\hat{S}^2_Q(f)} \tag{3.10}$$

$$\varphi_{\mathrm{rs}} = \arg \hat{S}_{PQ}(f). \tag{3.11}$$

During breathing, respiratory impedance is modulated in a complex way. Within-breath measurements show a marked bi-phasic pattern that is the result of volume

[1] The subscript "ao" refers to airway-opening, confer Table 3.2.

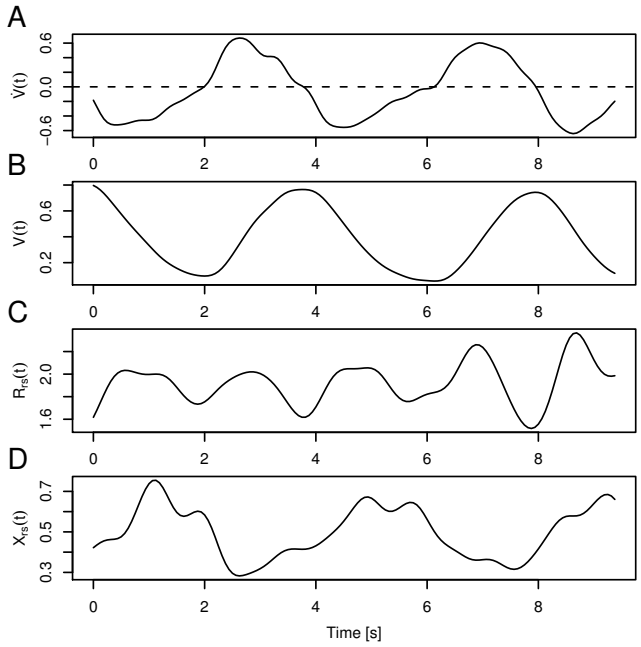


Figure 3.2: Typical forced oscillation signal from input impedance measurements in a healthy subject. A: Flow (positive: expiration, negative: inspiration). B: Volume. C: Estimated respiratory system resistance. D: Estimated respiratory system reactance.

Box 4. Real-time tracking of single-frequency FOT signals

The following important points should be kept in mind about the TLS approach:

- The Fourier transforms are estimated over (maximally overlapping) windows of a characteristic finite length, introducing temporal correlations in the estimates of mechanical lung properties.
- Spontaneous breathing interferes with impedance estimation through higher harmonics (McCall et al., 1957; Delavault et al., 1980; Daróczy and Hantos, 1982), although this influence is usually negligible at high enough forcing frequencies (> 4Hz).

and flow dependence (Davidson et al., 1986a; van der Putten et al., 1993), with slightly different behavior for inspiratory and expiratory phases (Oostveen et al., 1986), which is partially attributed for by interference with the larynx and glottis, but also hints at hysteresis in the respiratory system (Vincent et al., 1970). Figure 3.2 shows an example. The impedance signal also depends on posture, sympathetic tone (Butler et al., 1960), ventilatory inhomogeneities (Gillis and Lutchen, 1999) and

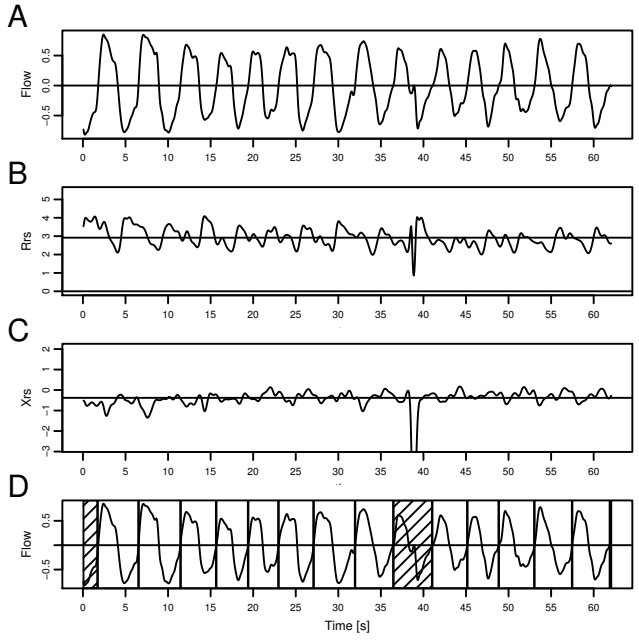


Figure 3.3: Artifact detection in respiratory impedance signals. A: Flow signal sampled at 16 Hz. B, C: Estimated respiratory resistance R_{rs} and reactance X_{rs}. Possible expiratory flow limitation is visible around $T = 38s$, resulting in large $R_{\mathrm{rs}}/X_{\mathrm{rs}}$ values at minimal flow that lie outside the depicted confidence region (here: X_{rs}). D: Flow signal after artifact detection and preprocessing. Distinct respiratory cycles are separated by zero-crossings of flow, and the shaded regions are rejected as incomplete or unreliable.

airway calibre (Peslin et al., 1992).

A further problem is the existence of various kinds of artifacts. Figure 3.3 shows the occurence of a common artifact (expiratory flow limitation) which is probably caused by closure of the glottis, i.e., swallowing or coughing, but might also have a disease-specific origin.

3.3 Asthma and COPD

Asthma and chronic obstructive pulmonary disease (COPD) are two of the most common chronic lung diseases, affecting millions of people worldwide (Global Initiative for Asthma, 2009; Global Initiative for Chronic Obstructive Pulmonary Disease, 2009). These numbers are expected to rise significantly in the years to come, further increasing the global burden. Although there is much known about the

pathophysiology, etiology and genetic epidemiology of these diseases, there are still many intriguing and not-well understood aspects. In particular, the *overlap problem* consists in the fact that the two diseases can be difficult to be correctly identified in clinical practice (Guerra, 2005), since they share many common features, and can even co-exist in the same patient.

3.3.1 Materials: FOT time series

Time series were obtained from the baseline part of a previous study (Slats et al., 2007) in which 13 asthma patients, 12 COPD patients and 10 healthy controls participated. The asthmatic patients were characterized by GINA guidelines (Global Initiative for Asthma, 2009) as mild and intermittent asthma (step I and II), were all non-smokers or ex-smokers with less than five pack years exposure, and had a history of episodic wheezing or chest tightness. Baseline forced expiratory volume in 1s (FEV_1) was more than 70% of predicted and the provocative concentration of methacholine for a 20% fall in FEV_1 (PC_{20}) was less than 8 $\mathrm{mg/mL}$. All asthma patients were atopic, as determined by one or more positive skin prick tests against 10 common aeroallergens.

The COPD patients were diagnosed with mild to moderate COPD (type I and II) according to GOLD guidelines (Global Initiative for Chronic Obstructive Pulmonary Disease, 2009) and were all smokers or ex-smokers with more than ten pack years exposure that had a history of chronic cough or dyspnea. Their $\mathrm{FEV}_1/\mathrm{FVC}$ ratio was less than 70% predicted post-bronchodilator, and the reversibility of FEV_1 by salbutamol was less than 12% of predicted.

All patients were clinically stable, used β_2-agonists on demand only, and had no history of respiratory tract infection or other relevant diseases up to two weeks prior to the study. None of the asthma or COPD patients had used inhaled or oral corticosteroids up to three months prior to the study.

The healty controls had no history of respiratory symptoms and were non-smokers or ex-smokers with less than five pack years exposure. Baseline FEV_1 was more than 80% of predicted and PC_{20} methacholine was more than 16 $\mathrm{mg/mL}$. They also showed no positive reaction to the skin prick test.

A forced oscillation device (Woolcock Institute, Australia) with a fixed oscillation frequency of 8 Hz and an amplitude of $\pm 1 \mathrm{cmH_2O}$ was used, after being calibrated with tubes of known resistance. Subjects breathed through an antibacterial filter with a resistance of $0.2\,\mathrm{cmH_2O\,s/L}$. Respiratory flow was measured by a Fleisch pneumotachograph (diameter 50 mm, Vitalograph Ltd, Maids Moreton, UK) and differential pressure was measured by a $\pm 2.5\,\mathrm{cmH_2O}$ solid-state transducer (Sursense DCAL4; Honeywell Sensing and Control, Milpitas, USA). Mouth pressure was measured using a similar transducer with a higher range ($\pm 12.5\,\mathrm{cmH_2O}$). Analog pressure and flow signals were digitized at 400 Hz.

Pressure and flow time series were transformed to the time-frequency domain by a maximal overlap discrete Fourier transform that acts as a band-pass filter for the frequency 8 Hz (filter width 100 samples, i.e., 0.25 s characteristic time). Time- and frequency-dependent complex respiratory impedance Z^*_{rs} was then estimated by the TLS fit (3.10-3.11), which is equivalent to maximum likelihood estimation.

In each subject measurements were repeated three times during 60 s of tidal breathing on four distinct days, during the course of a few weeks, yielding 12 time series in total. Before further analysis the impedance and accompanying pressure and flow signals were downsampled to 16 Hz, i.e., the Nyquist frequency for the applied pressure oscillation of 8 Hz.

3.3.2 Artifact removal

Artifacts were removed automatically by a custom-written algorithm. First, zero-crossings of the flow were identified that separated respiratory half-cycles, from which full respiratory cycles were constructed, concatenating consecutive expiratory and inspiratory half-cycles (in that order). Each respiratory cycle was then considered individually and rejected if one of the following three conditions were fulfilled:

- flow values within $1/5$ SD of zero occured at some time point where also at least one of X_{rs} or R_{rs} lay outside 3 SD from their mean values, being an indication of a flow artifact (flow limitation, glottis closure, coughing, etc).
- negative resistance values occured or the TLS estimation did not converge at some time point.
- values of R_{rs} or X_{rs} occured that lay outside a range of 5 SD from their mean values, being indicative of otherwise unidentified artifacts.

These events occured infrequently and only a few percent of breathing cycles were thereby rejected. The remaining cycles were concatenated to yield a single time series without gaps for each subject. Information on the beginning and end of each cycle was recorded separately.

3.4 Fluctuation analysis

A characteristic feature of asthma and COPD is their fluctuating behaviour, both in clinical symptoms and in the degree of airway obstruction. This behavior cannot be explained by simple models and suggests either a complex, high- or infinite-dimensional dynamical component and/or a strong stochastic component (Frey and Suki, 2008). Daily measurements of peak expiratory flow (PEF), e.g., exhibit long-range correlations over the course of months, indicating the existence of a long-term memory component in the respiratory system (Frey et al., 2005).

Special tools have been developed to analyze time series with regard to such fluctuations, and we consider power-law analysis in Section 3.4.1, and detrended fluctuation analysis in Section 3.4.2. Results obtained by these are given in Section 3.6. These will be compared to previously obtained results in the final discussion, Section 3.7.

3.4.1 Power-law analysis

Power-law probability distributions occur in a wide variety of contexts (Newman, 2005). Although there exist simple mechanisms that can generate power-laws (Reed and Hughes, 2002; Barabási and Albert, 1999), such distributions usually hint at hidden internal structure and complexities in an observed system, e.g., self-organized criticality (Bak et al., 1987). A characteristic of power-law distributions is that there is no preferred size, i.e., that the dynamic range of the observed realizations is unusually large. It is this latter property that motivates the use of power-laws as models for fluctuations in FOT time series.

Assuming a power-law probability density $f(x) \propto x^{\alpha}$ with exponent $\alpha \leq -1$, this density diverges as $x \to 0$, so there must be some lower bound to the power-law behaviour[2]. We denote this bound by $x_{\min}$. Discarding values below $x_{\min}$, we can then normalize the density to obtain

$$f(x) = -(1+\alpha)x_{\min}^{-(1+\alpha)}x^{\alpha}, \quad \text{for} \quad x \geq x_{\min}. \tag{3.12}$$

Traditionally, the density (3.12) is visualized in a double logarithmic plot of $f(x)$ against x, and the exponent α is estimated by a least-squares linear fit in this representation. However, it is now known that this method is potentially unreliable (White et al., 2008); the preferred robust method is to determine the exponent α by its maximum likelihood estimate (MLE),

$$\hat{\alpha} = -1 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{x_i}{x_{\min}}\right)\right]^{-1}. \tag{3.13}$$

The density (3.12) is known as the density of the Pareto distribution, and (3.13) is essentially equivalent to the Hill estimator commonly used in econometrics (Hill, 1975).

The MLE estimate $\hat{\alpha}$ depends on the usually unknown cutoff point $x_{\min}$, and will respond strongly to deviations from power-law behaviour at small values of x. To estimate it, Clauset et al. (2009) recommend to use the value $\hat{x}_{\min}$ for which the raw

[2] In practice, it is not uncommon that the assumption of an additional upper bound further improves the model fit (Newman, 2005), and often seems warranted due to physical limitations. However, as the estimation of both the power-law exponent and the lower/upper bounds become much more involved then (Aban et al., 2006), we limit the discussion here to the simple model discussed in the text.

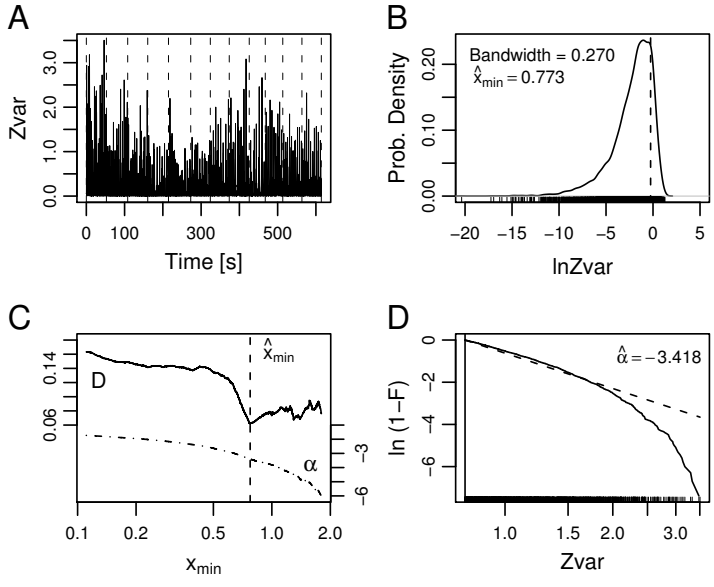


Figure 3.4: Power-law behavior in the distribtion of Z_{rs} fluctuations. A: Time series of fluctuations (Z_{var}) of respiratory impedance; the broken lines indicate the 12 distinct measurements. B: Kernel density estimate of the probability density of $\ln Z_{\mathrm{var}}$. The estimated onset of power-law behavior is indicated (broken vertical line). C: Estimated power-law exponent α and Kolmogorov-Smirnov statistic D. The optimal threshold is located at the minimum of D (broken vertical line). D: Estimated power-law behavior in the right tail of Z_{var} leads to a linear relationship in a double logarithmic plot of the distribution function F. The maximum likelihood estimate of the power-law behavior(with exponent α) is indicated (broken line).

Kolmogorov-Smirnov distance

$$D(x_{\min}) = \max_{x \geq x_{\min}} |F(x; x_{\min}) - S(x; x_{\min})| \tag{3.14}$$

is minimized. Here $F(x; x_{\min})$ denotes the empirical distribution function (only taking into account samples with $x \geq x_{\min}$) and $S(x; x_{\min})$ denotes the model distribution function

$$F(x; x_{\min}) = 1 - x_{\min}^{-(\alpha+1)} x^{\alpha+1}. \tag{3.15}$$

We note that MLE estimation of the lower bound is difficult, if not impossible, as the effective sample size depends on $x_{\min}$. Figure 3.4 shows an example of the

estimation procedure, applied to fluctuations

$$Z_{\mathrm{var}}(i) = \left(Z_{\mathrm{rs}}(i) - \bar{Z}_{\mathrm{rs}}\right)^2 \tag{3.16}$$

of Z_{rs} about its mean.

Given a nonnegative time series, it is always to possible to obtain estimates $\hat{x}_{\min}$ and $\hat{\alpha}$ of power-law behavior. How reliable are these? In other words, how likely is the hypothesis that the underlying distribution really arises from a power-law distribution? A simple, relatively conservative test, is to generate suitable surrogate data, estimate their power-law behavior, and consider the distribution of the Kolmogorov-Smirnov distances obtained for these. The fraction of values of (3.14) that are larger than the one observed for the actual data then results in an (approximate) significance probability for the null-hypothesis that the data arise from a power-law distribution. For large enough significance probabilities the (general) alternative is rejected and the power-law hypothesis is accepted. The surrogate data is constructed as described in Clauset et al. (2009): Each sample point arises either by bootstrapping the empirical distribution of the values $x < x_{\min}$, or is drawn from a power-law distribution with parameters $\alpha = \hat{\alpha}$ and $x_{\min} = \hat{x}_{\min}$. The probability for the first possibility is simply the fraction of sample points smaller than $x_{\min}$. This guarantees an essentially unbiased test that can be assessed, e.g., at the 10 percent significance level[3]. Note that it is not possible to correct for multiple comparisons, since for the general alternative it is not possible to control the family-wise type II error (of falsely accepting the null hypothesis).

3.4.2 Detrended fluctuation analysis

A different assessment of the fluctuations in time series signals is made possible by *detrended fluctuations analysis* (DFA). Invented by Peng and colleagues (Peng et al., 1995), this technique allows to detect long-range correlations and scale-invariant behaviour in time series. The first step in DFA is to integrate the deviations of a signal time series X_i ($i = 1, 2, \ldots, N$) from its mean $\bar{X}$,

$$Y_i = \sum_{j=1}^{i} (X_j - \bar{X}). \tag{3.17}$$

This transforms the (usually bounded) series X_i into an unbounded process Y_i, called the *profile* of X_i. This profile is then divided into $N_s = \lfloor N/s \rfloor$ nonoverlapping segments of length s. Since the length N of the time series is usually not a multiple of

[3] Even though this permutation test can rule out the case where a power-law is not a plausible model for the observed data, it might still be that other distributions (stretched exponential, log-normal) offer a better model. This is a general problem, however, and instead of further investigating this, we will be content here if the experimental evidence does not falsify the power-law assumption.

the scale s, a short part at the end of the profile may remain. In order not to disregard this part of the series, the procedure is repeated starting from the opposite end, leading to a total of $2N$ segments (Kantelhardt et al., 2002). For each such segment the local quadratic trend y is estimated by least-squares regression and subtracted from the data. The squares of the residuals are summed and divided by the length to yield the mean-square error $F^{(2)}(j,s)$ of the j-th segment at scale s,

$$F^{(2)}(j,s) = \frac{1}{s}\sum_{k=1}^{s}\left(Y((j-1)s+k) - y_j(k)\right)^2, \tag{3.18}$$

with quadratic trend y_j subtracted. Formula (3.18) only covers the forward case, the backward case for $j > N_s$ is calculated analogously. The second order fluctuation function is the total root-mean square error,

$$F_2(s) = \left(\frac{1}{2N_s}\sum_{j=1}^{2N_s} F^{(2)}(j,s)\right)^{1/2}. \tag{3.19}$$

The scaling behaviour of $F_2(s)$ is then assessed in a double logarithmic plot for a variety of scales s. In detail, since the smallest scales are biased due to the detrending, the smallest scale considered is usually chosen to be at least $s = 10$. The scale is then successively doubled until s is at most half of the length of the time series.

Power-law behaviour of $F_2(s)$ results in a line in the double logarithmic plot of $F(s)$ against s, which is estimated by weighted linear regression[4]. Weights proportional to the inverse of scale are usually used to account for the fact that the larger scales are estimated from less segments, i.e., with increased invariance. Figure 3.5 shows the procedure applied to (parts of) the impedance time series of a single subject, and Figure 3.6 shows the scaling behaviour found in this series.

The theoretical origin of DFA is the theory of diffusion processes. Assuming independent and identically distributed Gaussian increments $X_{i+1} - X_i$, the profile Y_i will be a trajectory of a random walk, and its variance will increase linearly in the number of time steps. Without detrending, the RMS error (3.19) will then exhibit scaling behaviour,

$$F(s) \propto s^{\alpha}, \tag{3.20}$$

with a characteristic exponent $\alpha = 1/2$. More generally, this relationship holds whenever the increments $X_{i+1} - X_i$ are uncorrelated; in particular, reshuffling the time series randomly will in principle result in such an estimate. On the other hand, long-range correlations in the X_i will lead to superlinear scaling. For example, fractional Brownian motion ("1/f noise") of the profile Y_i is a Gaussian process with zero

[4] Here it is not necessary to use maximum likelihood estimation (compare with the previous section). The number of scales is usually small, and each value $F_2(s)$ is a complex aggregate, so weighted linear regression is actually preferred in this case.

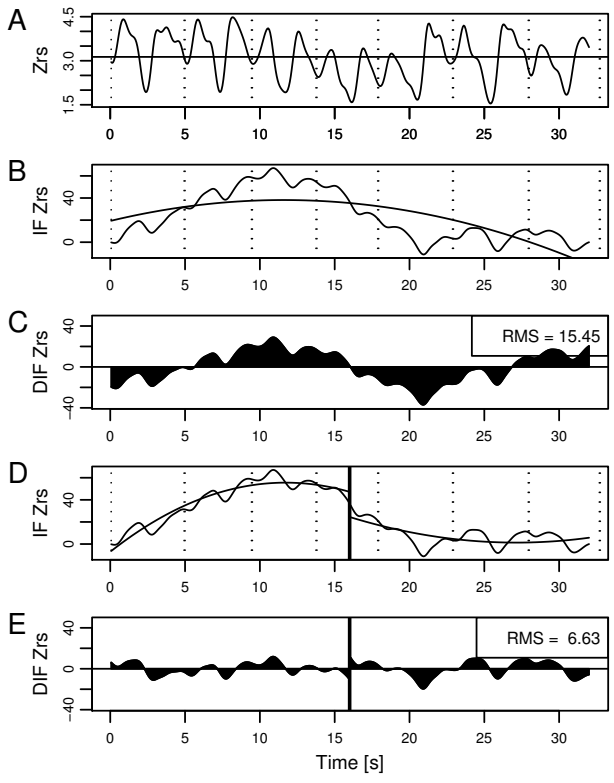


Figure 3.5: Detrended fluctuation analysis. A: Original Z_{rs} time series, only a part of which is shown. B: Integration leads to an unbounded signal, for which a quadratic trend is estimated. C: Subtracting the trend, the residual root-mean squared error (RMS) is calculated from the dark area. D: This process is repeated on a smaller scale, e.g., for each half of the original signal (separated by the vertical line). E: The RMS decreases for smaller scales. If the relation between RMS and scale follows a power-law, self-affine behaviour of the time series is detected and quantified by the DFA exponent (see Fig. 3.6.

mean, stationary increments, variance $\mathbb{E}Y_i^2 = i^{2H}$, and covariance

$$\mathbb{E}[Y_i Y_j] = \frac{1}{2}(i^{2H} + j^{2H} - |i-j|^{2H}), \qquad 0 \leq H \leq 1.$$

The parameter H is called the Hurst exponent, and the increment of fractional Brownian motion, i.e., fractional Gaussian noise, exhibits a DFA scaling exponent $\alpha = H$. Its autocorrelation function falls off with an exponent of $2H - 2$, leading to power-law behavior $P(f) \propto f^{-\beta}$ of the power spectral density with an exponent of $\beta =$

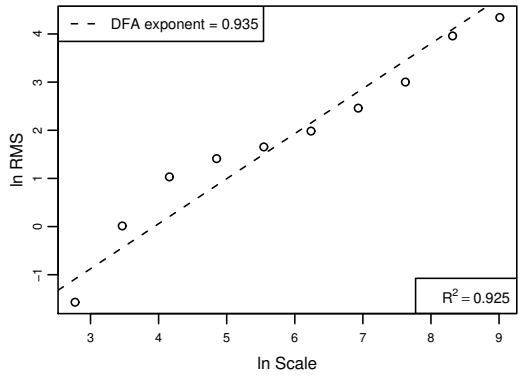


Figure 3.6: Calculation of the scaling exponent α of detrended fluctuation analysis. The relation between root-mean square error (RMS) is plotted against different scales in a double logarithmic plot. A linear least-squares fit with large coefficient of determination R^2 indicates the existence of scaling, here with an exponent (slope) of 0.933, indicative of $1/f$ noise.

$2H - 1$. In case $\alpha > 1/2$ such processes are therefore *long-range dependent*, whereas for $\alpha < 1/2$ they are *anti-correlated*[5].

A direct calculation also shows that (first-order) detrending does not change the scaling relationship (3.19) asymptotically (Taqqu et al., 1995). Its advantage is that detrending allows to account for certain kinds of nonstationarity, e.g., caused by random external influences that introduce weak trends ("drift") into the time series. Different orders of detrending lead to slightly different results, here we focus on second-order detrending, and the method is variously referred to as "DFA2" in the literature.

3.5 Nonlinear analysis

The previous section considered the stochastic properties of the respiratory system. In this section we will approach it from an orthogonal direction, by considering the respiratory system to be influenced by an underlying nonlinear (*deterministic*) dynamical system. In other words, whereas in the previous section our interest was on the properties of stochastic events, here we will consider these simply as "noise" and concentrate on the underlying deterministic component.

As in the previous chapter, from a time series $x = (x_1, x_2, \ldots, x_N)$ we construct

[5] An anti-correlated process has the property that successive values change sign with above chance probability. It's profile covers less distance from the origin (on the average) than Brownian motion.

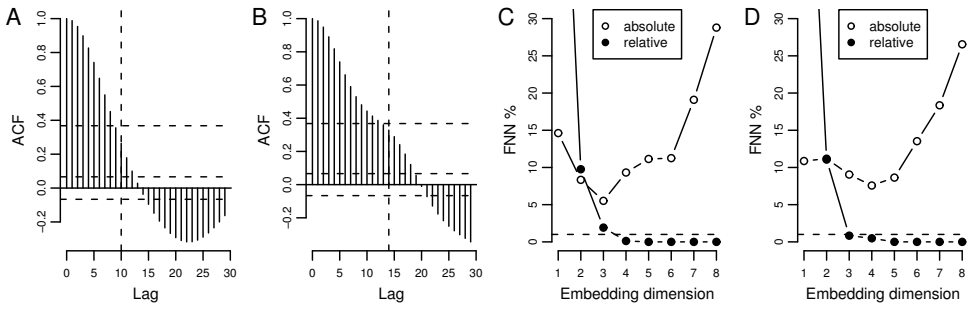


Figure 3.7: Estimation of optimal embedding parameters. A: Autocorrelation function (ACF) of Z_{rs} time series for a subject with asthma. The decorrelation time where the ACD falls off to $1/e$ and confidence intervals about zero are indicated (broken lines). B: Analogous ACF for a subject with COPD. C: Fraction of false nearest neighbours (FNN) for the R_{rs} time series of a subject with asthma. The choice of the optimal embedding dimension involves a compromise between relative and absolute FNN. The relative FNN indicate the number of false-nearest neighbors relative to an increase in embedding dimension and falls off monotonously. Values below 1% (broken line) indicate a proper embedding. This has to be judged against the absolute FNN that quantify the number of false-nearest-neighbors with respect to the diameter of the embedding. Noise in the time series leads to an increase for larger dimensions, and the embedding dimension should be chosen to minimize this influence. In this example the optimal embedding would be trhee-dimensional. D: Analogously for the X_{rs} time series of the same subject. The optimal embedding would be four-dimensional.

k-dimensional delay vectors at lag q,

$$x_{[i]} = (x_i, x_{i+q}, \ldots, x_{i+(k-1)q}), \qquad i = 1, 2, \ldots, N_*, \tag{3.21}$$

where $N_* = N - (k-1)q$. The trajectory $x_{[1]}, x_{[2]}, \ldots, x_{[N_*]}$ in phase space $\mathbb{R}^k$ is used as an approximation of the underlying invariant measure.

Section 3.5.1 discusses how to determine the optimal parameters k and q for this embedding. Section 3.5.2 introduces a measure that quantifies information production in dynamical systems from a given embedding.

3.5.1 Optimal embedding parameters

Although not essential, choosing an optimal time lag in the delay embedding guarantees optimal use of the available information. Such an optimal time lag is conveniently estimated by the decorrelation time of the autocorrelation function (ACF), which is the lag for which the ACF has fallen off to $1/e$, confer Figure 3.7.

The optimal embedding dimension can be estimated by the method of false nearest neighbours (Kennel et al., 1992). Figure 3.7 shows an example for a single measurement. False nearest neighbors are identified by either of two methods. First, a point $x_{[}i]$ is considered to have a *relative* FNN when increasing the embedding dimension increases the distance between the point and its nearest neighbour $x_{[}k(i)]$ by a factor of 10 or more. Secondly, a point $x_{[}i]$ is considered to have an *absolute* FNN when increasing the embedding dimension increases the distance between it and its nearest neighbor $x_{[}k(i)]$ by more than two times the diameter of the phase space embedding, estimated by the standard deviation of the time series. The fraction of relative FNNs usually falls off rapidly, and values below the 1 percent threshold indicate a proper embedding. The fraction of absolute FNNs, however, after a possible initial fall, usually rises strongly for large embedding dimensions. This rise is attributed to noise, whose influence becomes stronger for larger embedding dimensions (the so-called "curse of dimensionality"), and this measure compensates for the effect that distances in higher embedding dimensions automatically increase.

3.5.2 Entropy

Nonlinear systems often exhibit the property of sensitive dependence on initial conditions. This can be interpreted in information theoretic terms as the production of information: If two initial conditions are different but indistinguishable at a certain experimental resolution, they will evolve into distinguishable states after a finite time. The Kolomogorov-Sinai entropy h quantifies the mean rate of information production and is defined by a limit involving shrinking partitions of phase space (Eckmann and Ruelle, 1985). When working with actual data, it has become popular to approximate h by the K_2 entropy of Grassberger and Procaccia (1983), which is a lower bound for h.

To calculate K_2, define the finite correlation sums

$$C_i^k(r) = N_*^{-1} \left\{ \text{number of } j \text{ such that } d(x_{[i]}, x_{[j]}) \leq r \right\} \tag{3.22}$$

$$C^k(r) = N_*^{-1} \sum_i C_i^k(r), \tag{3.23}$$

where $d : \mathbb{R}^k \times \mathbb{R}^k \to [0, \infty)$ is a distance on phase space. The K_2 entropy is then given by

$$K_2 = \frac{1}{\Delta t} \lim_{r \to 0} \lim_{k \to \infty} \lim_{N \to \infty} \log \frac{C^k(r)}{C^{k+1}(r)}. \tag{3.24}$$

In practice, the maximum distance $d(x_{[i]}, x_{[j]}) = \max_{\nu=1,\ldots,k} |x_{i+(\nu-1)q} - x_{j+(\nu-1)q}|$ is used for computational efficiency. To avoid estimating the limits, the finite values

$$\mathrm{ApEn}(k, r, N_*) = \log \frac{C^k(r)}{C^{k+1}(r)} \tag{3.25}$$

		Z_{rs}	$Z_{\mathrm{rs}}\mathrm{SD}$	$\ln Z_{\mathrm{rs}}$	$\ln Z_{\mathrm{rs}}\mathrm{SD}$
Asthmatics	n=13	3.78 ± 1.53	1.07 ± 0.83	1.25 ± 0.37	0.24 ± 0.06
COPD	n=12	4.66 ± 1.18	1.43 ± 0.56	1.48 ± 0.27	0.31 ± 0.11
Controls	n=10	3.31 ± 0.99	1.03 ± 0.82	1.13 ± 0.28	0.26 ± 0.13

Table 3.3: Total averages ($\pm$ SD) for FOT dataset.

can be studied. The use of this familiy of measures was popularized in physiology by Pincus (1991), who recommended to use $k = 2$ and $r = \mathrm{SD}(x)/5$ as benchmark values, under the name of "Approximate Entropy". Richman and Moorman (2000) showed that ApEn was biased due to self-matches, and modified (3.22) to

$$B_i^k(r) = N_*^{-1} \left\{ \text{number of } j \neq i \text{ such that } d(x_{[i]}, x_{[j]}) \leq r \right\}. \tag{3.26}$$

The measure

$$\mathrm{SampEn}(k, r, N_*) = \log \frac{B^k(r)}{B^{k+1}(r)} \tag{3.27}$$

is called "Sample Entropy" and is the negative logarithm of the conditional probability that two sequences within a tolerance r for k time points remain within r of each other at the next time point.

3.6 Results

When comparing the three groups of asthmatics (A), COPD patients (C) and healthy controls (N), instead of only considering significance probabilities of differences on the group level, we were mainly interested in *predictive accuracy* with regard to group membership. This was estimated for (i) the full contrast between all groups, and (ii) the constrast asthma/COPD. For comparison, the worst-case classification accuracies, classifying all subjects as belonging to the largest group, were 0.37 (A/C/N) and 0.52 (A/C). If not stated otherwise, all accuracies reported below are conservative assessments based on leave-one-out cross-validation. Statistical significance was tested at the 1% level, and all tests between two groups of numerical values were Wilcoxon unpaired two-sample tests.

3.6.1 Statistical analysis

The mean values of respiratory impedance (Z_{rs}), resistance (R_{rs}) and reactance (X_{rs}) are shown in Figure 3.8 and summarized in Table 3.3. There was no significant group-wise difference between asthmatics and healthy controls, between COPD and asthma or between diseased subjects (both asthma and COPD) versus healthy controls in mean Z_{rs}, although Z_{rs} was slightly increased ($p = 0.014$) in COPD compared

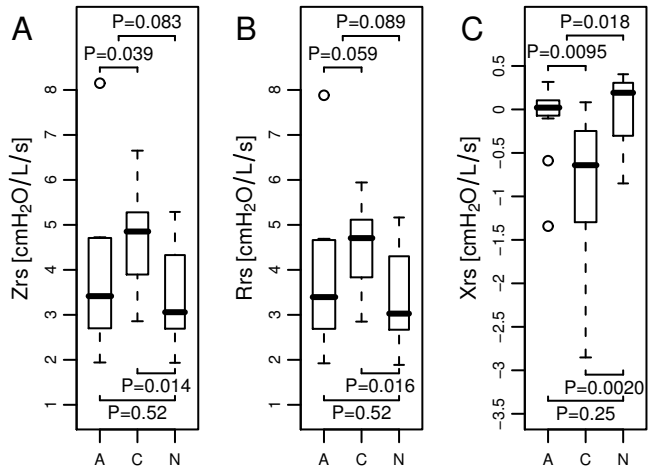


Figure 3.8: Mean values of respiratory impedance Z_{rs}, resistance R_{rs} and reactance X_{rs} in a boxplot that allows group-wise comparison. Significance probabilities (two-sample Wilcoxon test) are indicated. Groups are labeled (A: asthma, C: COPD, N: healthy controls).

to normal subjects and asthmatics ($p = 0.039$). This was attributed to marginally significant decreases in X_{rs} ($p = 0.020/p = 0.095$) and increases in R_{rs} ($p = 0.016/p = 0.059$), allowing for classification of COPD versus healthy subjects by LDA of mean Z_{rs} values with an accuracy of 0.73, and an accuracy of 0.60 in the asthma/COPD contrast, marginally above chance levels.

Since it has been suggested that the distribution of Z_{rs} is better explained by a log-Gaussian distribution than a Gaussian distribution (Diba et al., 2007), Fig. 3.9 depicts mean values of $\ln Z_{\mathrm{rs}}$ and $\ln Z_{\mathrm{rs}}$SD. No significant differences between asthmatics and controls were detected, consistent with the findings of (Diba et al., 2007), but COPD showed marginal increases in $\ln Z_{\mathrm{rs}}$ and $\ln Z_{\mathrm{rs}}$ variability ($p = 0.021$ and $p = 0.043$). Regarding higher moments, there were no significant differences in kurtosis (peakedness) and skewness (asymmetry) of Z_{rs} between the groups either (Fig. 3.10). A marginal decrease in skewness ($p = 0.094$) did achieve an accuracy of 0.73 for the asthma/COPD contrast, however.

Comparable classification in the asthma/COPD contrast was possible when the bivariate distributions of joint mean Rrs and Xrs were considered (Fig. 3.11). This achieved an accuracy of 0.72 (sensitivity 0.58, specificity 0.85 for COPD). The full contrast did not obtain any discrimination above chance levels (accuracy 0.40); in particular, of 10 healthy control subjects only one was correctly identified.

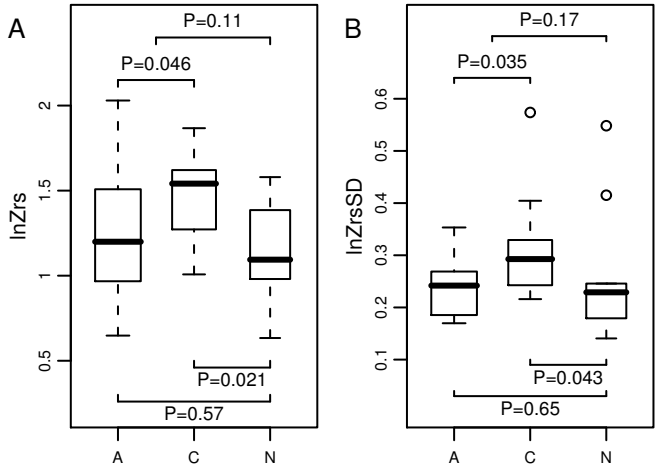


Figure 3.9: Mean values of $\ln Z_{\mathrm{rs}}$ and its standard deviation $\ln Z_{\mathrm{rs}}\mathrm{SD}$ in a group-wise comparison (A: asthma, C: COPD, N: healthy controls). Significance probabilities (two-sample Wilcoxon test) are indicated.

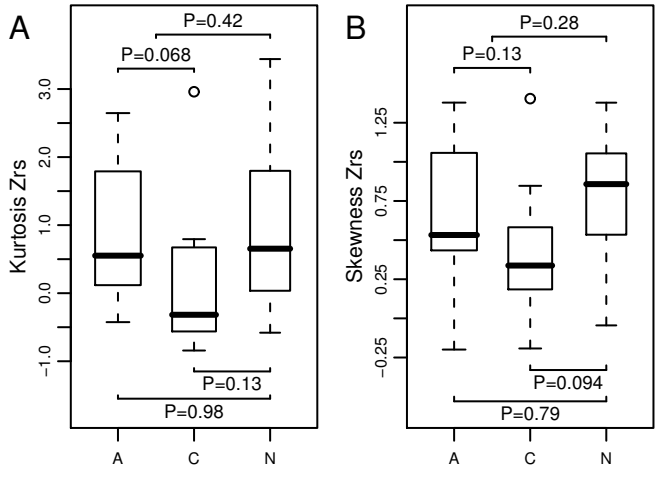


Figure 3.10: Higher moments of the distribution of Z_{rs} values in a group-wise comparison (A: asthma, C: COPD, N: healthy controls). A: Excess kurtosis as a measure of peakedness. B: Skewness as a measure of asymmetry. Significance probabilities (two-sample Wilcoxon test) are indicated.

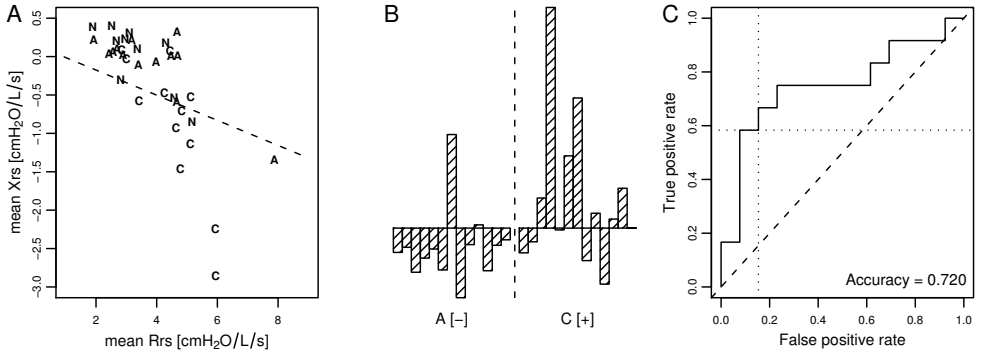


Figure 3.11: Linear discriminant analysis of combined mean resistance (R_{rs}) and reactance (X_{rs}) A: Scatterplot of mean R_{rs} against X_{rs} for all subjects (A: asthma, C: COPD, N: healthy controls). Note that the classification of normal subjects is almost impossible. The decision boundary for the classification of asthma against COPD is indicated (broken line). B: Discriminant scores for all subjects in the asthma/COPD contrast, cross-validated by leave-one-out method. C: Receiver-operator-characteristic for the discrimination of asthma (negatives) against COPD (positives) for these scores. Sensitivity (true positive rate, 0.58) and specificity (1-false positive rate, 0.85) for the optimal threshold are indicated (dotted lines), resulting in an accuracy of 0.72.

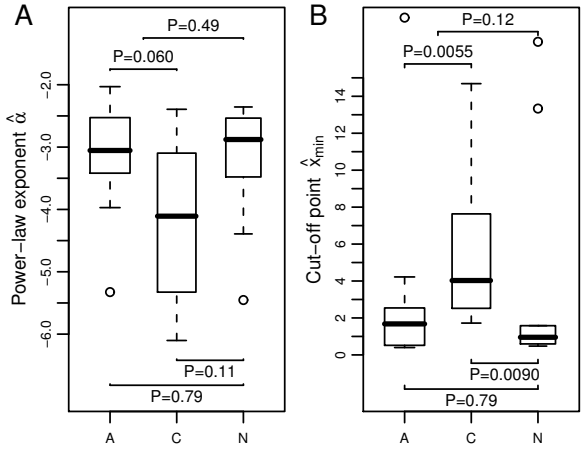


Figure 3.12: Estimated power-law behavior. Exponent (A) and onset threshold (B) in a group-wise comparison (A: asthma, C: COPD, N: healthy controls). Significance probabilities (two-sample Wilcoxon test) are indicated.

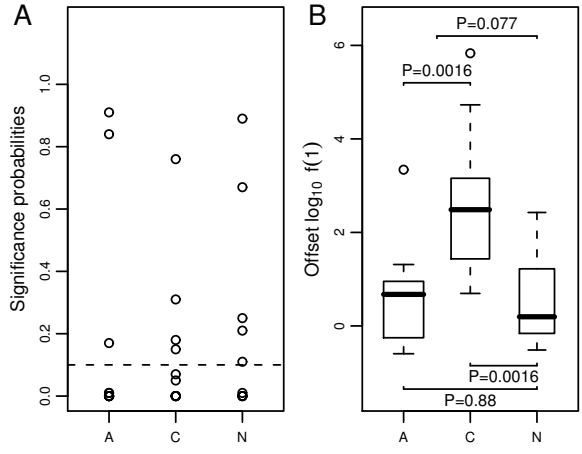


Figure 3.13: Evidence for power-law behavior and estimated intercept in a group-wise comparison. A: Significance probabilities for permutation test (100 bootstraps; A: asthma, C: COPD, N: healthy controls). The null hypothesis of power-law behavior is accepted (0.10 level, broken line) for 12 out of 35 cases, indicating compatibility with the power-law hypothesis. B: Intercept of power-law maximum-likelihood estimate of fluctuations Z_{var}, to compare with the findings of Que et al. (2000).

3.6.2 Variability and fluctuation analysis

We tested the power-law behavior of the Z_{rs} time series, and limited the MLE estimation of α to tails with at least 200 sample points to avoid spurious minima of the Kolmogorov-Smirnov statistic; the estimation was only performed for cutoff points above the 4th decile to speed up the computation. Estimated power-law exponents and thresholds are shown in Figure 3.12. There were no significant differences in exponents between the groups ($p > 0.06$), but in COPD the power-law behavior seemed stronger (smaller exponent α) and the threshold was significantly higher than for the other groups ($p < 0.009$). The latter could be explained by the seemingly larger variability (confer Fig. 3.9) in COPD, and lead to a classification accuracy of 0.68 (asthma/COPD) and 0.73 (COPD/controls). The logarithm of the extrapolated probability density at $x = 1$ showed a marginally significant increase for COPD with respect to the other groups ($p = 0.0016$; Fig. 3.13B), probably caused by the seemingly stronger power-law behavior. However, this only allowed close-to-chance classification. The null hypothesis of power-law behavior was accepted for 12/35 subjects, distributed almost evenly among the three groups (Fig. 3.13A).

Fig. 3.14 shows the scaling exponents and the goodness of fit obtained by DFA for all subjects. There were no significant differences in scaling between the groups, but the exponent was close to one in all cases, which indicates that respiratory impedance

Box 5. Power-law analysis of FOT signals

- Power law analysis is best done by maximum likelihood estimation.
- Validation of presumed power-law behavior is difficult, but significance testing with synthetic surrogate data offers a conservative assessment.
- In the sample dataset, the nullhypothesis of power-law behavior was accepted for 12/35 FOT signals.

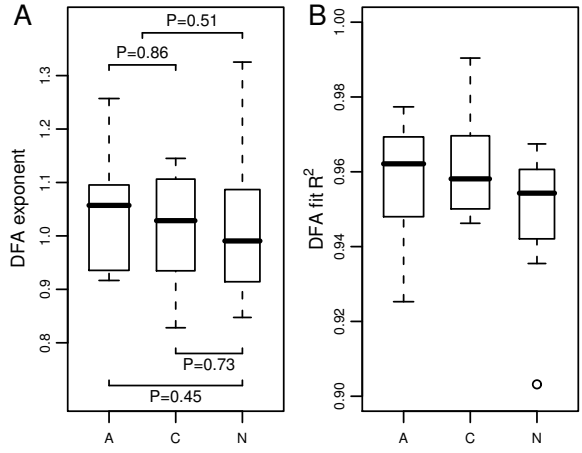


Figure 3.14: Detrended fluctuation analysis of Z_{rs} time series. Scaling exponent (A) and goodness-of-fit (B) DFA scaling exponent (A) and goodness-of-fit (B) in a group-wise comparison (A: asthma, C: COPD, N: healthy controls). Significance probabilities (two-sample Wilcoxon test) are indicated.

fluctuation can be considered an instance of $1/f$ noise, the hallmark of self-organized criticality (Bak et al., 1987) and complex, long-range dependent systems. Independent random fluctuations, e.g., by a white noise process, would result in a scaling exponent of 0.5, and the larger value found suggests a smoother, more correlated structure in respiratory impedance, which is expected due to the influence of the breathing cycle. Note however that the scaling exponent would be close to zero for a purely periodic process, e.g., simple harmonic variations in Z_{rs}.

To elucidate whether the scaling might be caused or influenced by variations in the lengths of the breathing cycles (Peng et al., 2002), we additionally extracted smaller time series of Z_{rs} with only one value per cycle either at the inspiratory endpoint (IEP), or at the expiratory endpoint (EEP), consisting of about 100 values on the average. Submitting these time series to DFA, scaling behavior was still detected

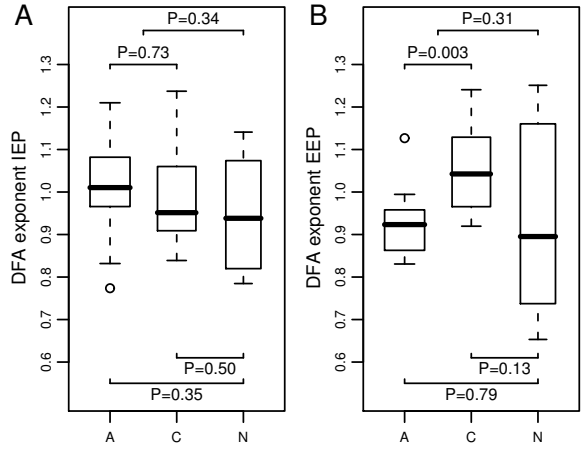


Figure 3.15: Detrended fluctuation analysis of event-related Z_{var} time series. A: Scaling exponents for Z_{rs} at the inspiratory endpoint (IEP) and B: at the expiratory endpoint (EEP) in a group-wise comparison (A: asthma, C: COPD, N: healthy controls). Significance probabilities (two-sample Wilcoxon test) are indicated.

(Fig. 3.15), indicating a more subtle, dynamical cause of the scaling. Interestingly, the EEP fluctuations exhibit a significantly larger exponent in COPD ($p = 0.003$) as in asthma, and allowed to indeed classify with an accuracy of 0.72 (asthma/COPD).

Box 6. Detrended fluctuation analysis of FOT signals

- DFA allows to assess self-similarity and long-range dependence of time series.
- The impedance time series exhibit scaling consistent with long-range dependence ("1/f noise", the hall-mark of complex systems).
- Scaling exponents did not significantly differ between controls and patients suffering from asthma or COPD, but this is due to large variation of the exponents. It seems that exponents might be slightly larger in asthma and in COPD (in that order) than in healthy controls, but longer time series are needed to assess this properly.

3.6.3 Distance-based analysis

Before attempting a dynamical analysis, we quantified differences in the shape of the joint probability distributions of resistance and reactance (Fig. 3.16). The results of the distance-based analysis for these 1+1 dimensional joint probability dis-

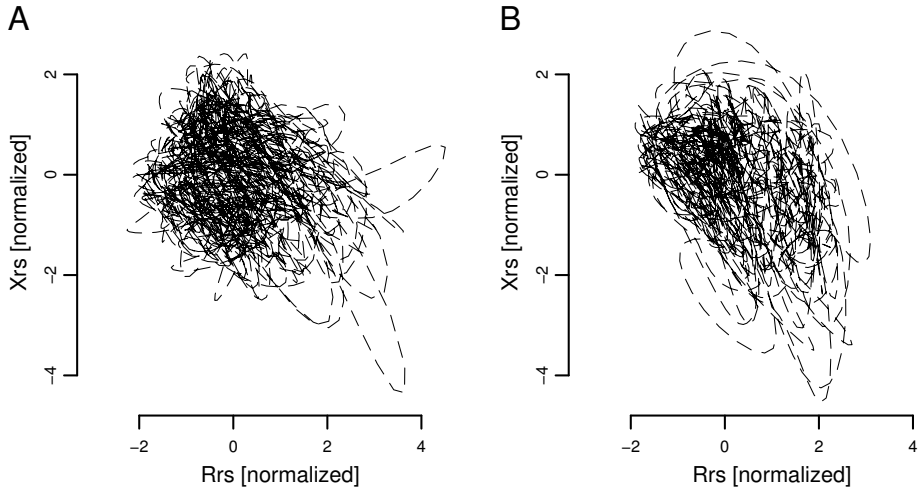


Figure 3.16: Wasserstein distances of mixed resistance (R_{rs}) and reactance (X_{rs}) time series. A: Trajectory of $R_{\mathrm{rs}}/X_{\mathrm{rs}}$ in a 1+1 dimensional embedding for a subject with asthma, normalized to zero mean and unit variance for each component separately. To improve visualization, stippled lines instead of individual sample points are shown. B: Analogous trajectory for a subject with COPD. The Wasserstein distance quantifies the work needed to transform one of these embeddings into the other, and thereby robustly quantifies differences in shape. For this example, the mean Wasserstein distance was 0.412 ± 0.029 SE (bootstrapped 15 times from 512 sample points each).

tributions, where both R_{rs} and X_{rs} were normalized independently, are shown in Fig. 3.17. All distances were bootstrapped 25 times with a sample size of 512 points each. The Wasserstein distances were reconstructed in two dimensions for visualization purposes (Fig. 3.17B), and the eigenvector distribution indicates that this represents the measured distances relatively well (Fig. 3.17C). Consequently, the misrepresentation error (circles in Fig. 3.17B) was relatively small and more or less uniformly distributed among the points. The group structure in this functional space was significantly clustered ($p = 0.002$), but a within-group agreement $A = 0.07$ suggests that only about 7% of the variance among distances is explained by group structure. Indeed, due to the normalization the distributions did not contain any information on mean R_{rs} and X_{rs} and their variability anymore, so only subtle differences in higher-order moments were captured by this approach. Including more reconstruction dimensions, the cross-validated classification accuracies decreased (Fig. 3.17E) and became unstable for dimensions larger than five (probably due to numerical inaccuracies related to very small eigenvalues). LDA in two MDS dimensions classified with accuracy 0.51 in the full contrast, and with accuracy

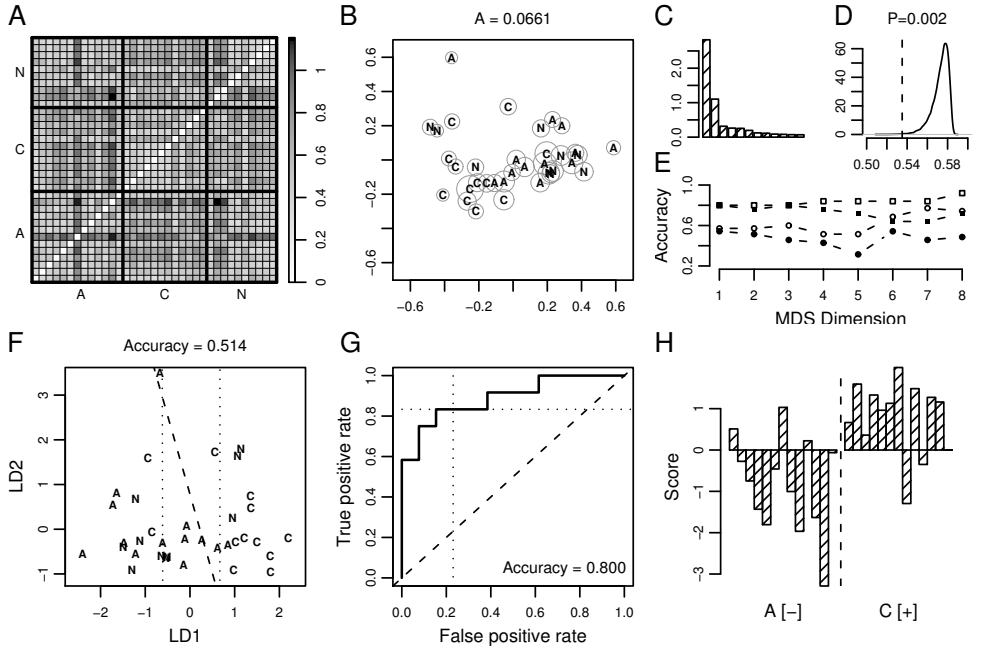


Figure 3.17: Distance-based analysis of normalized probability distributions of combined R_{rs} and X_{rs} (1+1 dimensional, as in Fig. 3.11). A: Distance matrix for all subject-wise comparisons of signals by Wasserstein distances. B: Two-dimensional MDS representation depicting the two major axes of variation. Each subject is represented by a point in this functional space (A: asthma, C: COPD, N: healthy controls), approximating the measured distances as close as possible. Misrepresentation error is indicated (by circles whose area is equal to stress-per-point). C: Eigenvalues of the scalar product matrix obtained from the distances, as a measure of explained variance. D: Distribution of the MRPP statistic δ. The value of δ for the original groups is indicated (broken line). The fraction of permutations to the left of this is the significance probability (P-value) that the distances are not structured with respect to group membership. E: Classification accuracies with respect to the number of MDS dimensions. Circles: full contrast, cross-validated (●) and resubstitution accuracy (○). Squares: asthma/COPD contrast, cross-validated (■) and resubstitution accuracy (□). The resubstitution accuracy rises with increasing dimensionality of reconstruction, but the cross-validated accuracy decreases after an optimal dimension, indicating overfitting. F: Discriminant functions for full classification in a two-dimensional reconstruction. The decision boundaries are indicated (dotted lines: full contrast; broken line: asthma/COPD contrast). G: Receiver-operator-characteristic for the discrimination of asthma (negatives) against COPD (positives) in a one-dimensional reconstruction. Sensitivity (true positive rate, 0.83) and specificity (1-false positive rate, 0.77) for the optimal threshold are indicated (broken lines), resulting in an accuracy of 0.80. H: Corresponding discriminant scores.

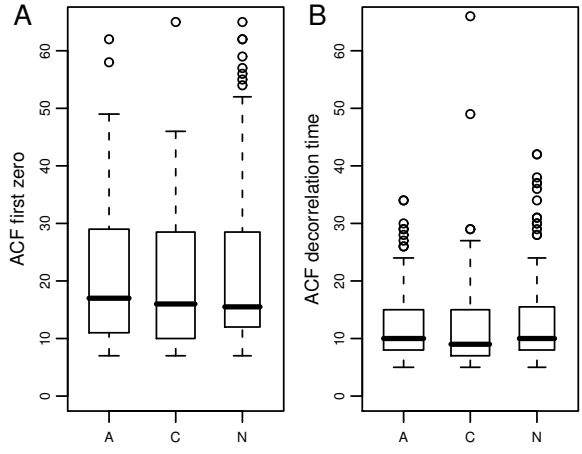


Figure 3.18: Optimal time lag estimation. A: Optimal lags determined by first zero of autocorrelation function (A: asthma, C: COPD, N: healthy controls). B: Optimal lags determined by decorrelation time. A few outliers not shown. Significance probabilities (two-sample Wilcoxon test) are indicated.

0.76 between asthma/COPD (Fig. 3.17F). The best asthma/COPD classification was achieved in just one-dimension, leading to an accuracy of 0.80 with sensitivity 0.83, specificity 0.77 for COPD (Fig. 3.17G-H).

3.6.4 Nonlinear analysis

Assuming the R_{rs} and X_{rs} time series to result from an underlying dynamical system, the proper time lag for delay vector reconstruction was assessed by the decorrelation time of the autocorrelation functions, with mean values of 14 ± 13 SD and 12 ± 9 SD, respectively. Due to the high variability, and since stochastic contributions to the signal might bias these estimates to larger values, the median values of 10 (for R_{rs} and X_{rs} alike) seemed the proper choice, corresponding to 0.625 s as characteristic time scale of the impedance dynamics, i.e., about one-fourth of a breathing cycle. Assessment of false nearest neighbours (FNN) suggested an embedding dimension of three to four (FNN R_{rs}: relative 3.8 ± 0.6 SD, absolute 3.1 ± 1.6 SD; X_{rs}: relative 3.9 ± 0.7 SD, absolute 2.7 ± 1.5 SD) and $m = 3$ was chosen, as balancing the influence of noise seemed more important than improved resolution of the dynamics.

As in the 1+1 dimensional case, we quantified differences between the 3+3 dimensional delay vector distributions of R_{rs} (three delay coordinates) and X_{rs} (the other three coordinates), again normalizing the two to zero mean and unit variance independently. Results are shown in Fig. 3.19. The eigenvector distribution (Fig. 3.19C) suggests that although two dimensions captured most of the variance

Box 7. Embedding parameters used in reconstructing impedance dynamics

- Optimal time lag: 10 samples (median decorrelation time of ACF).
- Optimal embedding dimension: 4 values (minimal absolute FNNs and relative FNNs below 1 percent).

of the distances, quite a few more are needed to represent the distances faithfully. Indeed, for a two-dimensional MDS reconstruction the misrepresentation error was quite large (Fig. 3.19B, compare with Fig. 3.17B). The group structure was still significant ($p = 0.005$; Fig. 3.19D), although the larger value of the significance probability (compared with Fig. 3.17D) indicates an increased level of noise, which is confirmed by the lower within-group agreement $A = 0.03$. The classification accuracies for the full contrast attained their maximum for two dimensions and for the asthma/COPD contrast in six reconstruction dimensions (Fig. 3.19E). The full contrast was difficult to resolve, due to considerable overlap between asthma and controls (Fig. 3.19F), and the accuracy in two dimensions was only 0.51. The asthma/COPD contrast had accuracy 0.88 (sensitivity 1.00, specificity 0.77 for COPD) in six reconstruction dimensions (Fig. 3.19G-H).

3.6.5 Entropy analysis

Figure 3.20 depicts SampEn of the impedance time series for the reference two-dimensional embedding and for a four-dimensional embedding with a time lag of 10 samples. For the former, the COPD group exhibits significantly larger entropy values, but the asthma group does not seem to differ from the control group. Interestingly, in contrast to this the asthma group differs significantly (from the other two groups) in the four-dimensional embedding, with a seemingly lower entropy. However, these differences were quite small: Cross-validated classification in the contrast asthma/COPD by LDA, for example, resulted in an accuracy of 0.59 (2d) and 0.57 (4d), i.e., close to chance classification.

3.7 Discussion

We have attempted to distinguish between asthma, COPD and healthy controls either by assessing fluctuations and scaling behavior, or by robustly comparing probability distributions of the dynamical behavior of R_{rs} and X_{rs}, implicitly assuming an underlying dynamical system.

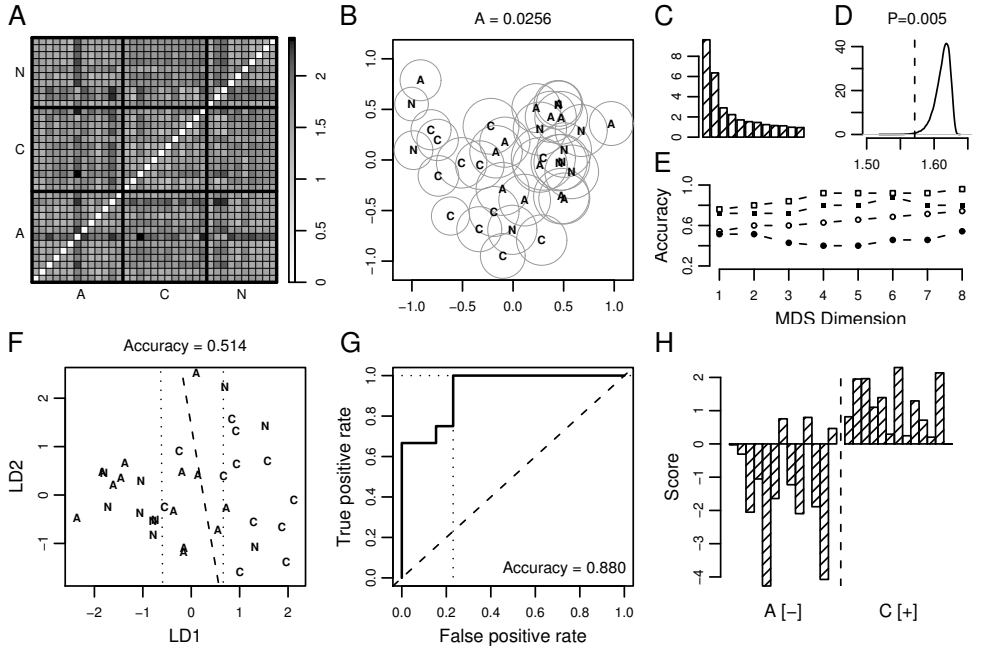


Figure 3.19: Distance-based analysis of normalized probability distributions of combined R_{rs} and X_{rs} in a 3+3 dimensional embedding, compare Fig. 3.17. A: Distance matrix. B: Two-dimensional MDS representation of subjects (A: asthma, C: COPD, N: healthy controls). Note the larger misrepresentation error, due to the increase in dimensionality. C: Eigenvalues of the scalar product matrix. D: Distribution of the MRPP statistic δ. E: Classification accuracies. Circles: full contrast, cross-validated (●) and resubstitution accuracy (○). Squares: asthma/COPD contrast, cross-validated (■) and resubstitution accuracy (□). F: Discriminant functions for full classification in a two-dimensional reconstruction with decision boundaries indicated (dotted lines: full contrast; broken line: asthma/COPD contrast). G: Receiver-operator-characteristic for the discrimination of asthma (negatives) against COPD (positives) in a six-dimensional reconstruction. Sensitivity (true positive rate, 1.00) and specificity (1-false positive rate, 0.77) for the optimal threshold are indicated (broken lines), resulting in an accuracy of 0.88. H: Corresponding discriminant scores.

3.7.1 Main findings

Evidence for the controversial power-law hypothesis was found. Although the power-law hypothesis could not be accepted for all subjects at the 10 percent significance level, this represents a rather conservative test (Clauset et al., 2009), and the fluctuations of 12/35 subjects were consistent with power-law behavior. However, this does not rule out the possibility that the data is still better described by other distri-

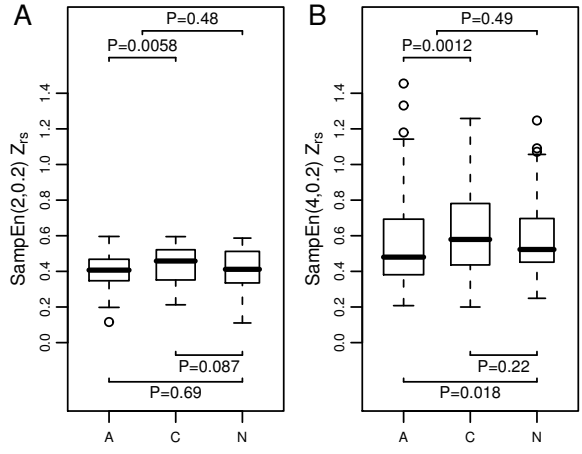


Figure 3.20: Group-wise comparison of sample entropies of Z_{rs} time series (A: asthma, C: COPD, N: healthy controls). Evaluted for each single measurement separately. A: In a two-dimensional embedding with trivial lag. B: In a four-dimensional embedding with lag 10 samples.

Box 8. Sample entropy of FOT signals

- Sample entropy is a simple and efficient method to quantify information production in dynamical systems, even for relatively short time series.
- In a two-dimensional reconstruction, the COPD patients exhibited slighly larger entropies, consistent with larger variability.
- In a four-dimensional reconstruction, the asthma patients exhibited significantly lower entropies, paradoxically suggesting a slightly more stable dynamics.
- However, these effects were relatively small and only allowed for marginal improvements in classification accuracy, compared to the worst-case classifier.

butions. In particular, unphysiologically large values of Z_{rs} cannot *a priori* occur, so there should also be an upper threshold for the assumed power-law behavior. Unfortunately the estimation of such distributions is much more involved and introduces additional sources of uncertainty, so this was not attempted here.

Consistent with earlier findings we did not detect significant changes between power-law exponents. In contrast to Que et al. (2000), we did not detect significant differences in power-law properties between asthmatics and controls. Since the earlier analysis was done with methods now known to be potentially unreliable (confer

(White et al., 2008)), these former findings should be reconsidered.

Detrended fluctuation analysis highlighted the complicated nature of impedance signals, where self-similar scaling behavior was found. Interestingly, the scaling exponents were close to one, indicating that Z_{rs} time series exhibit complex $1/f$ fluctuations in time that are correlated over a wide range of scales. The origin of this $1/f$ noise might be the well-known $1/f$ fluctuations of heart rate (Peng et al., 1995) and it can be hypothesized that these are mediated by the complex interactions in the cardiorespiratory system. However, since the scaling behavior also persisted when only one value per breathing cycle was used, we conclude that this is a subtle effect, i.e., most probably due to dynamical cardio-respiratory coupling, and not simply driven by $1/f$ noise of inter-breath interval variability (Fadel et al., 2004).

Due to large intra-group variance, fluctuation analysis showed no significant differences between the groups of subjects, but there were indications that the scaling exponents might be slightly larger in diseased respiratory systems than in healthy controls, consistent with the idea that diseased dynamics in physiological system are characterized by a decrease in complexity (Goldberger et al., 2002).

The distance-based analysis between probability distributions further evidenced that there exist subtle differences in respiratory properties. Since the R_{rs} and X_{rs} time series were normalized for this analysis, only differences in the shape of the dynamical behavior were thereby quantified. Interestingly, these were sufficiently large to allow robust (cross-validated) classification of 80 percent in the asthma/COPD contrast, which was better than classification based on mean Z_{rs}, $\ln \mathrm{Z_{rs}SD}$, skewness and kurtosis of Z_{rs}, etc., individually. This confirms our hypothesis that the two diseases differentially affect their within-breath dynamics.

Regarding the 3+3 dimensional delay embedding and its Wasserstein distances, these did only improve classification marginally (to 88 percent in the asthma/COPD contrast) with respect to the 1+1 dimensional distributions. In the light of the largely increased noise level (due to the sparseness of delay vectors) this indicates that such delay reconstruction might possibly incorporate additional information that is not present when only using the 1+1 dimensional distributions. However, it seems necessary to reduce the influence of noise considerably before this could be convincingly demonstrated and is left to future studies.

Classification of asthmatics versus healthy controls was problematic for all measures due to large overlap; however, there were indications that some time series should be considered outliers, i.e., either systematically influenced by unidentified artifacts, or exhibiting a unique dynamic relating to uncontrolled covariates.

3.7.2 Clinical implications

The distance-based time series analysis of respiratory impedance led to a correct distinction between patients with asthma and COPD in 80 percent of cases. This means

that the forced oscillation technique can capture discriminative aspects of airway disease from recordings during simple, tidal breathing. The differential diagnosis of asthma and COPD can be a challenge in clinical practice (Guerra, 2005) as it appears that both diseases can exhibit overlapping pathological and physiological features (Gibson and Simpson, 2009). Part of this overlap may be due to real co-existence of both diseases in some patients, whereas in others the current diagnostic techniques apparently fail to pick up the difference (Chang and Mosenifar, 2007; Gibson and Simpson, 2009). Our patients were used as a so-called 'training-set' (Knottnerus and Muris, 2003), thereby being representative of gold-standard patients of either disease. The presently observed discriminative capacity of the dynamic time series analysis is, therefore, promising with regard to differential diagnosis and monitoring of asthma and COPD. The fully non-invasive nature of the measurements, without the requirement of artificial breathing maneuvers, offers great prospect for clinical application in chronic, often elderly patients. However, this still requires validation experiments, in independently recruited patients with an intention-to-diagnose (Knottnerus and Muris, 2003), in order to establish the diagnostic accuracy of the dynamic time series analysis of respiratory impedance in clinical practice.

3.7.3 Further directions

Although the large variability in scaling exponents precludes the reliable discrimination of respiratory diseases in the case of relatively short Z_{rs} signals considered here, it can be hoped that fluctuation analysis might elucidate differential mechanisms in respiratory dynamics if performed with improved input signals. In particular, obtaining impedance signals at a few additional frequencies and partitioning them into different compartmental contributions and mechanical properties, should allow to reduce the influence of upper airway artifacts, thereby enhancing the discriminatory properties. The use of nonlinear compartmental models, e.g., based on Rohrer's equation (Suki, 1993) might improve the analysis even further, even under loss of temporal resolution.

Nonlinear noise reduction algorithms have been quite successful for the removal of artifacts (Kantz and Schreiber, 2004)[ch. 10.3] and could also be investigated. With respect to the detected scaling behavior, the heart-lung interaction could be studied, correlating fluctuations in ECG and respiratory impedance signals. A stochastic description by a Fokker-Planck equation seems another possibility (Friedrich and Peinke, 1997; Nawroth et al., 2007) to separate the stochastic and deterministic contributions to the impedance signal.

LDA, although it assumes normality and equal group-wise variance, was used here instead of more complicated discriminant methods due to its wide spread, ease of interpretation and relative robustness - since fewer parameters need to be estimated than, e.g., in quadratic discriminant analysis, LDA usually outperforms more

sophisticated classification methods. Cross-validated LDA results can be judged to be conservative, however, and more sophisticated methods might lead to improved classification.

3.7.4 Conclusion

Instead of evaluating Z_{rs} signals with respect to the mechanical properties of airways, we have attempted a stochastic and nonlinear analysis. The distance analysis showed that there exist subtle differences in these signals, but the nature of the differential behavior of respiratory impedance is mostly unclear. Self-similar fluctuations were detected in the signals, that hint at a complex modulation of impedance signals which needs further elucidation. The distance analysis has proved particularly useful and detected clustering in functional space, indicating functional changes in respiratory impedance that are characteristic with respect to disease. Reverse-engineering of these patterns is a possibility, since the interpolation properties of Wasserstein distances (Villani, 2003)[ch. 5.1], in combination with nonlinear modeling techniques (Gouesbet and Maquet, 1992; Żółtowski, 2000), principally allow to compute characteristic dynamical models for each group of subjects. This would potentially lead to further insights into how the respiratory system is affected in disease and possibly also allow to assess and track changes in airway caliber over the course of time. This may be used for differential diagnosis and monitoring of asthma and COPD; however, independent validation experiments are required as the next step towards the assessment of diagnostic accuracy (Knottnerus and Muris, 2003).

Chapter 4

Structural brain diseases

My brain: it's my second favorite organ.

Woody Allen

Section 4.1 gives a short overview of magnetic resonance (MR) imaging and its quantitative analysis. The framework for the statistical analysis of MR parameters is introduced in Section 4.2. In Section 4.3 we apply this methodology to data obtained in the case of an important autoimmune disease, systemic lupus erythematosus (SLE). The dataset is analyzed by two established approaches: "histogram analysis" and "multivariate discriminant analysis". Methodological improvements are introduced and discussed; we compare results obtained by the traditional methods with the distance-based analysis of MR parameter distributions and by results obtained when fitting a model distribution to the data. In Section 4.4 we turn to a different brain disease, Alzheimer's disease, where a short application is presented.

4.1 Quantitative MRI

Magnetic resonance (MR) imaging, also known as nuclear magnetic resonance (NMR) imaging, is a medical imaging technique that allows to visualize the internal structure and function of the body. In comparison to computertomography and positron emission tomography, MR imaging does not utilize ionizing radiation, and is therefore an ideal technique for prolonged or repeated measurements. A strong magnetic field is used, and a short radio frequency (RF) electromagnetic field causes subatomic particles with nonzero magnetic moment, e.g., the protons in living tissue, to alter their alignment relative to the field, inducing precession around the longitudinal axis. To be precise, the external magnetic field changes the energy levels of the particles' intrinsic spin-states. Not only does this introduce a basis for the spin eigenstates, which now occur either as longitudinally aligned or counter-aligned with the field, but also introduces an energetic asymmetry, where the aligned state is energetically preferred. This changes the usually random distribution of spin states in thermal equilibrium, leading to a macroscopic net magnetization vector along the magnetic field.

Application of an RF field at the resonant (Larmor) frequency causes particles in the lower energy state to jump to the higher energy state, causing the macroscopic

magnetization vector to rotate away from the longitudinal alignment. After the RF field is turned off, the protons subsequently relax to the alignment preferred in the magnetic field, thereby emitting characteristic electromagnetic radiation. The recovery of longitudinal magnetization is called T_1 or *spin-lattice relaxation* and is macroscopically described by an exponential decay $M(t) = M_\infty + M_1 \exp(-t/T_1)$ with time constant T_1.

T_2 relaxation, or *spin-spin relaxation*, occurs when the spins in the two energetic states exchange energy but do not loose energy to the surrounding lattice. This results in the loss of transverse magnetization, which is again described by a macroscopic exponential process, the so-called free induction decay. Its time constant T_2 is a measure of how long the resonating protons retain coherent phase relationships in their precessing motion. In practice, inhomogeneities in the magnetic field and the electron distribution of the molecules (chemical shifts) render the relaxation time shorter and a specific excitation protocol (spin-echo series) is needed to measure T_2. If the standard excitation protocol (gradient-echo series) is applied, the magnetization suffers from additional losses and is called T_2^*. This measure can further increase contrast for distinct molecules, e.g., venuous blood. Finally, proton-density (PD) weighted scans use an echo sequence that suppresses relaxation times and allows to measure the total amount of available spins.

In the application to biological tissues, consisting of a variety of distinct molecules, the relaxation processes are influenced by the natural quantum mechanical energy levels of the molecules, i.e., their vibrational, rotational and translational energy spectra. Small molecules like water usually move more rapidly and exhibit higher natural frequencies than larger molecules, e.g., proteins. Since the natural frequency of (free) water molecules is much higher than the range of Larmor frequencies used clinically, water has a long T_1 relaxation time relative to other biomolecules. On the other hand, the methemoglobin in blood shortens T_1 times significantly due to dipole-dipole interactions between the paramagnetic iron and the water protons.

Many alternative MR imaging protocols exist that prescribe different sequences of excitation signals and thereby allow to measure additional parameters. Important examples include diffusion weighted and diffusion tensor imaging, functional MRI, and MR spectroscopy. The *magnetic transfer* (MT) parameter is another important parameter that refers to the transfer of longitudinal magnetization from hydrogen nuclei of water that are restricted in their movements, e.g., bound to macromolecules such as proteins and lipids, to nuclei of water that can move freely. Due to interaction effects the T_2 times of bound nuclei are severly shortened and usually not observed in the standard MR protocols. However, the use of a special excitation series allows to saturate the bound nuclei, and during their relaxation magnetization will be transferred to the free nuclei, increasing the T_1 time. The magnetization transfer ratio (MTR) expresses the difference between proton-density in the absence

of saturation (M_0) and in the presence of saturation (M_1), and is defined as

$$\mathrm{MTR} = \frac{M_0 - M_1}{M_0}. \tag{4.1}$$

To acquire images of an extended object, the magnetic field is caused to vary spatially by introducing a field gradient. Different spatial locations then become associated with distinct resonant frequencies, allowing to separate their contributions to the MR signal, which is usually accomplished by Fourier methods. This results in a three-dimensional reconstruction of the MR signal, sampled over a discrete set of voxels (volume-elements). Each voxel is assigned a single numerical value that represents the average of the MR signals measured for that volume. The mixing of distinct tissue contributions is called the *partial volume effect*, and is one of the most problematic issues in the interpretation and analysis of MR images.

Quantitative MR image analysis refers to the analysis of the measured distribution of MR parameter values (one for each voxel) by statistical methods (Tofts, 2004). Instead of the mainly visual analysis by medical experts — which is still the main mode of evaluation of MR images, utilizing the human capacity to detect patterns — quantitative MRI analysis is an objective approach that uses the MR scanner in its original intention, i.e., as a sensitive scientific instrument that results in numerical measurements. The statistical analysis of these measurements, however, is a relatively recent development. A particular problem are the differences in MR scans between different scanners, the strong dependence on the details of the imaging protocols employed, and possible drift, potentially compromising reproducibility even in the same machine. Moreover, focussing on a distinct *region of interest* (ROI) inside the brain is only possible if that region can be identified and delineated with sufficient precision. This necessitates the use of nontrivial image analysis techniques for the segmentation and registration of images that map the observed voxels to a standardized coordinate system, e.g., an average brain atlas.

4.2 Distributional analysis

The basic idea in the distributional analysis of MR images is to consider the imaging parameter a stochastic process. Adopting such a stochastic description allows to apply statistical techniques to quantitatively analyze MR images. The most common approach is to neglect (microscopic) spatial dependence and consider each value of the imaging parameter, of which there is one for each voxel measured, a realization of a single random variable with a common probability distribution. In contrast to other statistical descriptions, e.g., by Markov random fields or texture analysis, this at first seems unfortunate, as information on the spatial origin of parameter values is discarded.

However, there are many advantages to this approach: It is easy to realize and to interpret, and one does not need to make assumptions about spatial properties of the signal. Moreover, focussing on a region of interest (e.g., white or grey matter), it is still possible to account for macroscopic differences in tissue properties, i.e., for the anatomical origin of the signals, without the need for perfect registration of voxels to a common coordinate frame.

The first task in this approach is to estimate the probability distribution of the imaging parameter. Due to the abstraction involved, the probability distribution is the most complete description of the information content of the signal. At this point, a further assumption enters. Namely, the diverse micropscopic tissue contributions and their properties are superposed and averaged in MR signals, and on the macroscopic level of a discrete image with a finite set of voxels this gives rise to a seemingly continuum of parameter values. Therefore, one seems justified in assuming that the probability distribution is absolutely continuous, i.e., that there exists a continuous density function f describing the probability $f(x)\,\mathrm{d}x$ to observe a parameter value from a small interval $[x, x + \mathrm{d}x]$. As implied from the above, the value $f(x)\,\mathrm{d}x$ can be interpreted as the probability to observe such values of the imaging parameter if the origin of the signal is unknown (but constrained to lie in a certain ROI).

Since the probability density f is continuous, it contains an essentially infinite amount of information and cannot be directly assessed. The standard approach to deal with this problem has been to discretize the parameter range and to estimate the value of $f(x)$ in a small interval $[x, x + h]$ by the observed frequency of voxels exhibiting parameter values falling into this bin. Thereby, the continuous density f is estimated by a histogram.

Depending on the bin numbers and sizes, the amount of information is largely reduced. For example, total white matter might be found and measured in thousands up to a few million voxels (depending on scanner resolution), and commonly used histograms usually contain about 100 bins. This is still a lot of information, and the next task is therefore the extraction of a number of select statistical descriptors, also called summary statistics, that summarize this information in a more convenient and, hopefully, meaningful way. Statistical measures of this kind are the moments of the distribution and functions thereof, e.g., the mean, the standard deviation, kurtosis and skewness. With respect to parameters that exhibit *unimodal* densities, i.e., exhibit a marked single maximum, additional other measures have been considered. These are the height and location of the mode ("peak") of the density, and its broadeness, quantified by the width at half the height of the peak, the so-called *full width half maximum* (FWHM) statistic. The study of these measures is known under the name of "histogram analysis" in the field of qualitative MR image analysis.

A different way to characterize histograms or, more generally, densities by a number of features is offered by the methods of multivariate analysis. These cannot be applied to a single histogram but necessitate the use of multiple histograms, i.e.,

measurements from a population of subjects. Each bin is then considered a distinct variable and multivariate methods are applied to the collection of all such variables. For example, PCA then allows to extract linear combinations of bins that capture the most variation over the population. Thereby, each single histogram is decomposed into a number of additive components, and the percentage that each such prinicpal component (PC) contributes to it leads to a series of numerical scores. Since usually only a few PCs explain most of the variance, a few of these scroes characterize each single histogram, greatly reducing the amount of data even further.

Another advantage of the exploratory technique of PCA is that PCs can often be interpreted substantively, which might lead to insight into neurophysiological processes and possible generative approaches, i.e., microscopic models of tissue properties and their relationship with other factors, e.g., disease severity, cognitive abilities, age and time, etc. In particular, PCA scores can be used in regression analysis to link parameter distributions with covariates of interest.

However, if the objective is to discriminate between different groups of subjects, e.g., healthy controls and one or more populations affected by disease, PCA is often not optimal. Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are generalizations of PCA that extract the features of histograms most useful for classificatory purposes. As in PCA, LDA and QDA determine linear combinations of bins, now chosen to separate two or more groups of histograms optimally. The difference between LDA and QDA consists in the underlying parametric assumptions. In QDA, each population of histograms is assumed to arise as a realization of a multivariate Gaussian distribution, and in LDA the covariance matrices of these Gaussians are furthermore assumed to be all equal (homoscedasticity assumption). It is questionable whether these assumptions are truly fulfilled, but this does not compromise classification. Indeed, LDA/QDA are easily implemented and interpreted, and although they might not be optimal if the assumptions are violated, as yet they have shown marked successes in discriminating and assessing neurological diseases.

4.3 Systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease that can affect any part of the body. It has a prevalence of about 40 per 100 000 people in Northern Europe (Rahman and Isenberg, 2008) and, as other autoimmune diseases, affects women more frequently than men, with a ratio of almost 9 to 1. Its diagnosis is complicated as its symptoms vary widely and can occur in unpredicatable outbreaks (“flares”) with intermediate remissions. Therefore, SLE is often mistaken for other diseases, and the diagnostic guidelines of the American College of Rheumatology rely on more than ten possible criteria (Tan et al., 1982). Even when SLE has been diagnosed, current treatments with immunosuppressive drugs have undesirable side

effects and it is imperative to quantify disease severity to improve the quality of life of patients. A further complication arises from the fact that 11 to 60 percent of patients exhibit neuropsychiatric symptoms, which is also called "active" SLE (NPSLE), whereas the remaining patients do not exhibit such symptoms (nonNP-SLE).

It has been shown in the past that noninvasive measurements of MTR distributions in the brain allow to detect SLE. The SLE group exhibits significant deviations of histogram-based measures (e.g., peak height) with regard to the control group. Classification on an individual basis has been attempted by Dehmeshki et al. (2002), where histograms where considered as multivariate measurements. Linear discriminant analysis (LDA) of histograms allowed to discriminate active SLE from controls perfectly (in 10 healthy and 9 SLE subjects) under leave-one-out crossvalidation, and to distinguish nonNP-SLE from SLE perfectly (in 9 SLE and 10 nonNP-SLE subjects). This study was one of the first to shift attention from a study of groupwise differences to classification on an individual level, i.e., to the assessment of the predictive value of MTR parameter distributions in the brain. Unfortunately, the proposed multivariate discriminant analysis (MDA) of histograms is faced with the problem that histogram data does not represent independent measurements, and collinearity in the data can compromise the validity of the classification results. In the following, we will therefore use principal component analysis (PCA) to first transform the data to a smaller and more robust set of scores, and then apply LDA to these to find an optimal, robust basis for classification. This approach, called linear discriminant principal component analysis (LDPCA, or MDA for simplicity) is further contrasted with (1) classification based on histogram measures, (2) a fitted model distribution, and (3) the distance-based analysis of MTR distributions. It will be shown that the best method for general classification and discrimination of diseased patients from healthy controls is the distance-based comparison, whereas slightly better results can be obtained by LDPCA/MDA for specific comparisons, e.g., when discriminating active NPSLE from nonNP-SLE.

4.3.1 Materials

A dataset of in total 54 subjects was measured in a 3 Tesla MR scanner at the Leiden University Medical Center. The subject population consisted of 19 healthy controls (abbreviated by code "H"), 14 NPSLE patients (code "N") and 20 patients with nonNP-SLE (code "O"). The age and gender structure of the population is shown in Figure 4.1.

All images were obtained in a $256 \times 256 \times 20$ matrix and automatically registered and segmented. This led to six distinct, partially overlapping datasets of MTR parameters for each subject: White matter (WM) and gray matter (GM) were separately available, and combined result in the total brain parenchym (PAR). Application of morphological erosion to each of these three image datasets removed one voxel from

Subject populations	
H	Controls ("healthy")
N / NPSLE	Systemic lupus erythematosus with neuropsych. problems
O / nonNP-SLE	Systemic lupus erythematosus without neuropsych. problems
S = N + O	Disased patients
Datasets	
WM	White matter
EWM	Eroded white matter
GM	Gray matter
EGM	Eroded gray matter
PAR	Brain parenchyma = White matter + Gray matter
EPAR	Eroded brain parenchyma
Other abbreviations	
CV	Cross-validated (leave-one-out)
FPR	False positive rate
FWHM	Full width half maximum
MDA / LDPCA	Multivariate discriminant analysis (of histogram data)
MR	Magnetic resonance imaging
MTR	Magnetic transfer ratio (an MR imaging parameter)
pu	Percentage unit
ROC	Receiver operator characteristics
TPR	True positive rate
T_2	Transverse / spin-spin relaxation time (an MRI parameter)

Table 4.1: Abbreviations used in this chapter.

the boundaries of the anatomical regions, thereby greatly reducing possible contamination by the partial-volume effect at the boundaries. This operation resulted in an additional three datasets of eroded white matter (EWM), eroded gray matter (EWM) and eroded parenchym (EPAR) per patient. Since we were only interested in global changes in MTR distributions, the information on the origin of the MTR values was discarded, leading to six sets of parameter values for each subject that consisted of about 60 000 (WM/EWM), 140 000 (GM/EGM) and 200 000 (PAR/EPAR) individual MTR values.

4.3.2 Histogram analysis

For each patient, the MTR values from the interval $(0, 1]$ were binned into 101 regular bins of width 1 percent unit (PU) each[1]. From these histograms, three measures were

[1] The bins have been chosen such that their centers coincide with the 101 MTR ratios from 0 to 1, in steps of 0.01 pu. Thereby, the bias of the histogram estimate at these MTR ratios is minimal, but the first and last bin represent values from the intervals (-0.05,0.05] pu and (0.95,1.05] pu, respectively, half of which are unphysical and cannot occur. This introduces a slight bias into their interpretation as frequencies/densities,

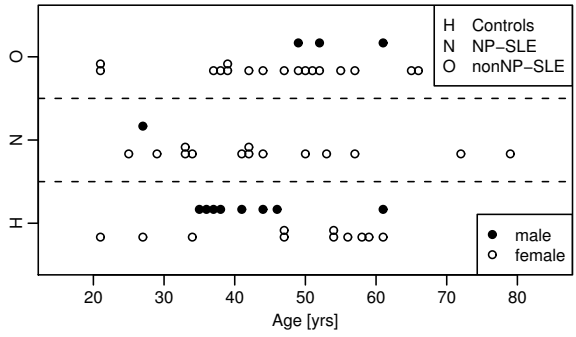


Figure 4.1: Age and gender structure of the sample dataset.

obtained: the location (bin center) of the peak, the height of the peak, and its FWHM (Figure 4.2, left panel). The mean values and groupwise standard deviations are given in Table 4.2 for the total brain parenchym. The peak height is expressed in the frequency with which MTR ratios from the same bin occur, i.e., normalized such that the frequency values of all 101 bins sum up to 1. It is seen that on the average the mode of the MTR ratios is located at about 0.39 and represents on average about 7.5% of all MTR values in the brain. In patients suffering from NPSLE the height of the mode is significantly reduced by about 1% and its location is shifted to slightly lower MTR ratios, with a group-wise average of about 0.38. This shift to lower MTR ratios leads to a broadening of the MTR distribution, as reflected in a significantly increased FWHM. The nonNP-SLE group shows intermediate changes in MTR distributions that suggest that the same structural changes have occured in their brains as in the NPSLE group, only less pronounced.

Cross-validated classification results are given in Tables 4.5-4.9 for all possible binary contrasts and the full classification task. These accuracies are based on a linear model that takes all three histogram measures into account, and range from about 60% in the full contrast to about 80% in the HN contrast. The receiver operating characteristic for the PAR dataset is given in Panel A of Figure 4.4, showing sensitivity (true positive rate) against 1-specificity (false positive rate). About half of the active NPSLE patients can be unambigously classified. Note that these accuracies drop immensely if not all three measures are used. For example, using only the peak height the cross-validated classification accuracy is about 13%, and using only the peak location it is only about 31%.

which is ignored here.

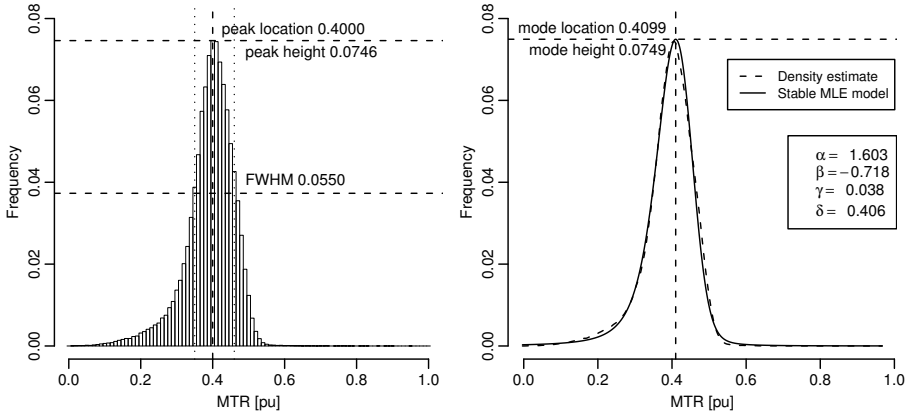


Figure 4.2: Histogram analysis. A: Traditional histogram measures. B: Fitting a stable distribution.

Group	Peak height	Peak location	FWHM
H	0.0746 (0.0068)	0.393 (0.012)	0.052 (0.005)
N	0.0626 (0.0084)	0.377 (0.021)	0.064 (0.011)
O	0.0713 (0.0073)	0.382 (0.014)	0.054 (0.007)
Contrast	Significance probabilities		
HN	0.0000913	0.00870	0.00021
HO	0.157	0.00882	0.565
HS	0.00379	0.00228	0.0206
NO	0.00343	0.350	0.00157

Table 4.2: Groupwise means and standard deviations for histogram-based measures and significance probabilities of differences (Wilcoxon two sample test) for PAR dataset.

4.3.3 Multivariate discriminant analysis

Prior to classification by LDA, we performed PCA on the histograms to extract those linear combinations of bins that represent the most variance in the dataset. The first three principal components, covering about 95.8% of the total variance, are shown in Figure 4.3, scaled according to their contribution to the variance. The first component is bimodal (with a shape typical for the first-derivative), representing the main shift in MTR ratios to lower values in SLE patients. The second component exhibits a similar shift at a different location, and the third component has a diffusive (second-derivative) character and partially represents the broadening of the MTR distribution.

The first panel in Figure 4.3 shows the experimental subjects in the space spanned

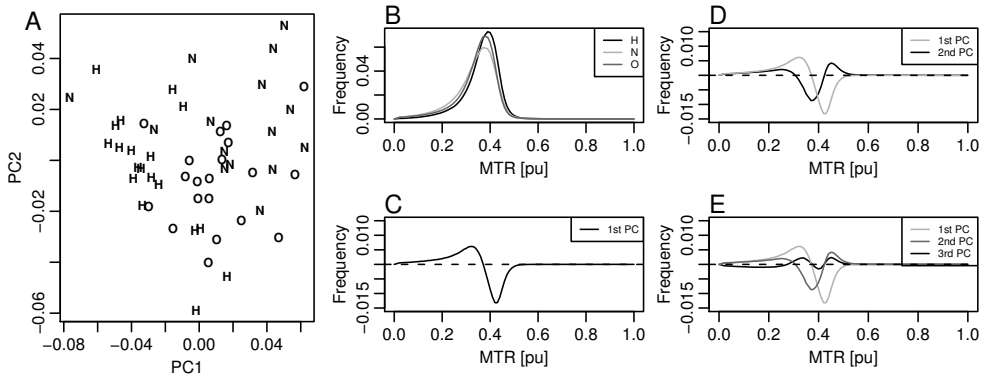


Figure 4.3: MDA/LDPCA of PAR dataset. A: Scores for first two principal components. B: Mean MTR histograms (100 bins). C: First principal component (64.4% of variance). D: First two second PCs (92.4% of variance). E: First three PCs (95.8% of variance).

by the first two principal components (PCs). Clearly, the first component represents the main signal differentiating healthy subjects from NPSLE patients, with nonNPSLE patients half-way in between. However, only about 85% of subjects can be classified correctly from the first PC, contrasted with about 88% for the first two PCs (cross-validated) and 91% for the first 11 PCs (covering 99.83% of the variance). The cross-validated accuracies and the number of principal components used for all datasets and contrasts are given in Tables 4.5-4.9. Note that the number of principal components was constrained to lie between 1 and 12 and was determined by searching for the smallerst number of PCs where the largest cross-validated accuracy occurs. The main reason for this was to allow for an essentially unbiased comparison with the distance-based accuracies (Section 4.3.5), although the accuracies might be slightly biased. Ideally, one would first determine the number of principal components on a training data set, and then evaluate its performance on the remaining, independent test dataset. The relatively small dataset precludes such ideal cross-validation, but the left Panel in Figure 4.6 shows that there is not much variation in the cross-validated accuracies with respect to the number of PCs, i.e., that the potential bias should be negligible.

The accuracies achieved with such an adaptive choice of PCs and LDA in the space of scores range from about 65% in the full contrast to more than 90% when distinguishing healthy controls from NPSLE patients. The ROC curve in Panel B of Figure 4.4 shows that essentially all positives (NPSLE patients) can be detected with only about 20% false positives.

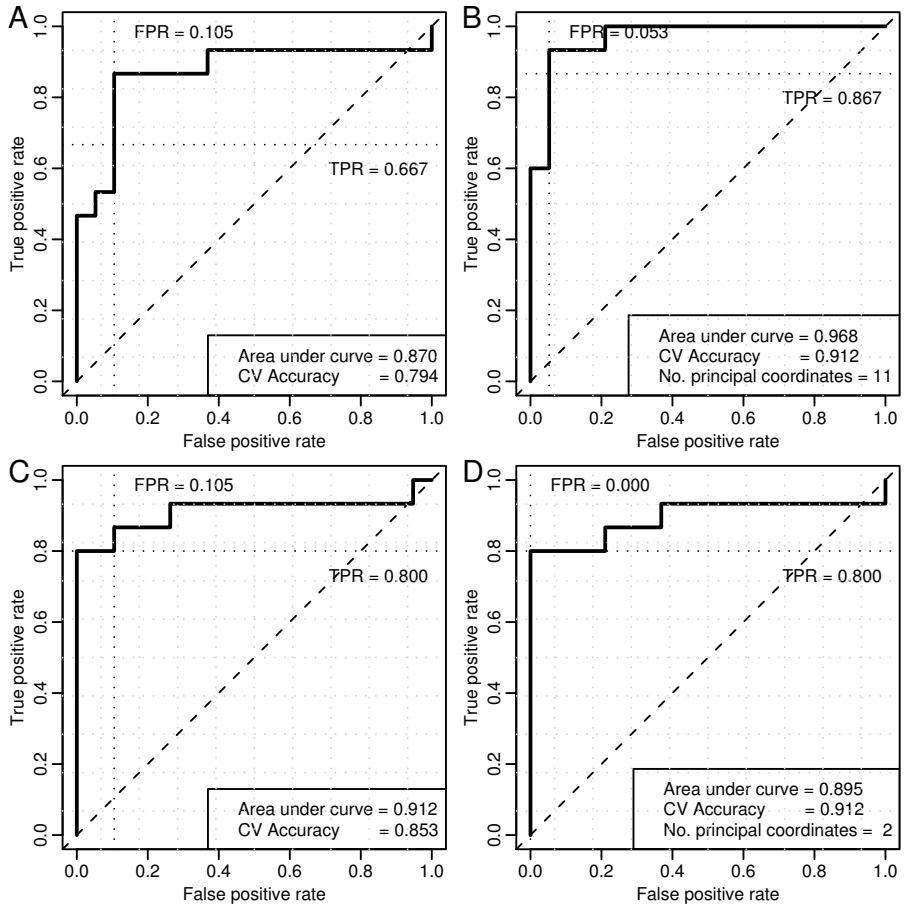


Figure 4.4: Receiver-Operator characteristics for discriminating active NPSLE from healthy controls in the PAR dataset. The stippled diagonal is the line of no discrimination, equal to the situation of a random guess. The total accuracy of classification (both positives and negatives) and the area under the ROC curve are given for further assessment. A: Based on traditional histogram measures. B: Based on histogram PCLDA/MDA. C: Based on fitting a stable distribution. D: Based on Wasserstein distances.

4.3.4 Fitting stable distributions

An interesting possibility in the analysis of MTR values is the fitting of experimentally measured distributions to a parametric model distribution. The class of stable distributions seems a promising candidate. This is a flexible four parameter family of distributions which are characterized by the property that they are attractors

Group	α	β	γ	δ
H	1.54 ± 0.06	-0.93 ± 0.05	0.039 ± 0.005	0.386 ± 0.008
N	1.56 ± 0.10	-0.95 ± 0.00	0.048 ± 0.008	0.365 ± 0.016
O	1.50 ± 0.08	-0.93 ± 0.07	0.041 ± 0.006	0.371 ± 0.010

Table 4.3: Fitting stable distributions to the PAR dataset by maximum likelihood estimation.

for properly normed sums of independent and identically-distributed random variables. In detail, a random variable X is stable if for two independent copies X_1 and X_2 of X and any constants $a, b > 0$ the following holds for some constants $c > 0$ and $d \in \mathbb{R}$ (in distribution):

$$aX_1 + bX_2 \stackrel{d}{=} cX + d. \tag{4.2}$$

. In other words, the shape of the distribution of X is preserved (up to scale and shift) under addition.

If the variance of these random variables were finite, one would get a Gaussian distribution as special case of a stable distribution. Without this assumption of finite variance, the limit may be a general stable distribution that can show various degrees of heavy-tailed ("power-law") behavior and skewness. Although there does not exist a general closed-form formula for stable distributions, they can be parameterized and numerically manipulated through series representations. We follow the parameterization of Nolan (2010), where a stable distribution is characterized by four parameters: an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\gamma > 0$ and a location parameter $\delta \in \mathbb{R}$. Maximum likelihood estimation of these parameters is provided by the fBasics package from the Rmetrics project[2]. The parameter estimates for the PAR dataset are shown in Table 4.3.

The location and scale parameters δ and γ are consistent with the histogram measures (peak location, FWHM) in Table 4.2, although both are slightly smaller by about 0.10 pu. The group-wise standard deviation is somewhat smaller than for the histogram measures. All fitted distributions are highly skewed, and both β and the stability index α are almost constant within the dataset (up to standard error). Classification by these parameters resulted in cross-validated accuracies ranging from about 65% for the full contrast to about 85% for the H-N contrast, improving on the histogram measures but not quite reaching the LDPCA/MDA accuracies (Tables 4.5-4.9).

[2] http://www.rmetrics.org

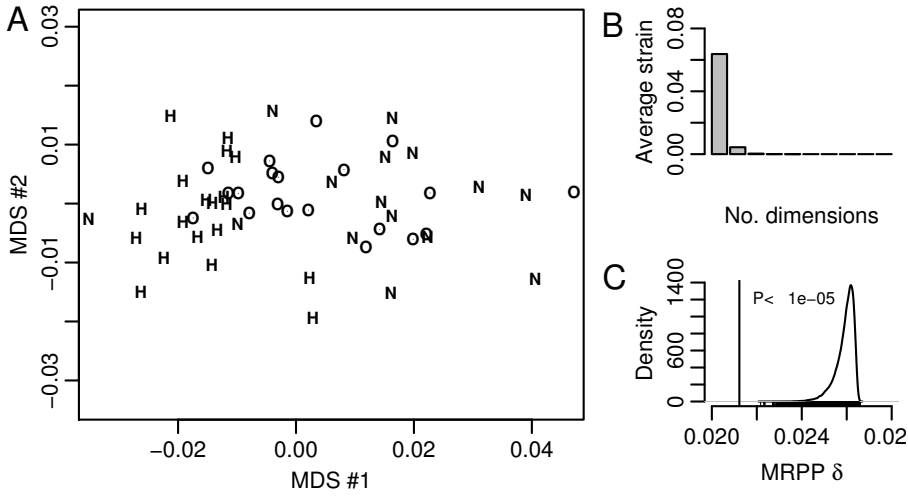


Figure 4.5: Wasserstein distances of PAR dataset. A: Two-dimensional MDS representation. Strain per point is depicted by circles but too small to be visible. B: Average strain per number of reconstruction dimensions. C: MRPP statistic indicates highly significant clustering.

Contrast	Best method			Second best method		
	CV Acc.	Method	Tissue	CV Acc.	Method	Tissue
H-N	0.970	mda (5)	EGM	0.939	dist (1)	EGM
H-O	0.872	histo	EWM	0.846	dist (6)	GM
H-S	0.870	dist (6)	PAR	0.852	mda (7)	PAR
N-O	0.829	mda (8)	PAR	0.800	dist (5)	PAR
Full	0.796	dist (6)	PAR	0.679	mda (2)	EGM

Table 4.4: Best crossvalidated classification for each contrast. mda: LDPCA/MDA; histo: LDA of histogram peak, location and FWHM; dist: 1D Wasserstein distances and LDA of MDS coordinates.

4.3.5 Distance-based analysis

For the distance-based analysis, we calculated 1D Wasserstein distances between the MTR distributions of all 54 subjects. Since these distances can be calculated efficiently in the one dimensional case, no bootstrapping was needed. Results are shown in Figure 4.5. The MRPP test confirms highly significant clustering of the three groups, with a chance-corrected withing-group agreement of about $A = 0.181$, indicating that almost 20% of the variance of the distances can be attributed to the group structure. The cross-validated classification accuracies range from about 75% for the full contrast to 90% for the HN contrast (Tables 4.5-4.9).

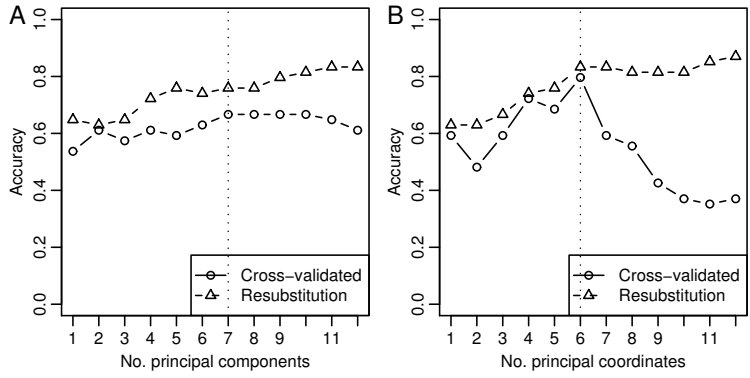


Figure 4.6: Influence of dimensionality on accuracies (triangles) and cross-validated accuracies (circles). A: LDPCA/MDA. Only small degradation of cross-validated accuracies occurs due to sligtly increasing numerical errors for a larger number of principal components. B: Wasserstein distances. Increasing the reconstruction dimension beyond six degrades cross-validated discrimination considerably. There are no negative eigenvalues, so this is a genuine overfitting effect.

Box 9. Systemic lupus erythematosus

- Subjects suffering from SLE can be successfully distinguished from healthy subjects by the statistical analysis of the distribution of the MTR imaging parameter.
- SLE patients with neuropsychatric problems exhibit marked differences in gray matter MTR distributions, whereas SLE patients without such problems differ mostly in white matter MTR properties.
- The best classification results for the discrimination between (1) healthy and diseased (NPSLE/nonNP-SLE) subjects and (2) for the full classification task are obtained by the distance-based comparison of their MTR distributions.
- The best results for the binary discrimination between (1) NPSLE and nonNP-SLE and (2) between healthy and NPSLE subjects is achieved by LDPCA of histograms.

4.3.6 Discussion

It has been established that MTR distributions of brain MR images are significantly changed in subjects suffering from SLE. These changes are more marked if subjects exhibit neuropsychiatric problems (NPSLE), but occur to a lesser extent also in subjects without these manifestations (nonNP-SLE). The true test whether these ob-

served differences can be attributed to SLE is the assessment of predictive accuracy in a blind classification task. Since at present there are not enough measurements available for such validation, we have used leave-one-out crossvalidation as an approximation of this ideal situation. In contrast to simply finding "significant" differences at the group-level, as is still common in most publications on SLE and other brain diseases, the much harder problem of individual classification of subjects has attracted little attention so far. The work of Dehmeshki et al. (2002) was the first to obtain predictive accuracies for SLE patients by a multivariate discriminant method (LDPCA/MDA). The accuracies obtained with this method are impressive, even when contrasting NPSLE with nonNP-SLE patients, but have been only assessed in a small group of 10+9+10 subjects and with a standard 1.5 Tesla MR system. Here we have presented results for a larger group of 19+15+20 obtained with a modern 3.0 Tesla MR scanner. These confirm the former results, improving about 10%-15% on the classification possible by the histogram-based measures mostly used.

Apart from independent confirmation of earlier approaches, our work introduces four novel aspects into this problem. First, we have distinguished between six different tissues on whose MTR values the classification has been based. The results show that healthy controls and NPSLE patients can be best classified by their gray matter MTR distributions, whereas the same controls and nonNP-SLE patients can be best classified by their white matter MTR distributions. In the full classification task, or when discriminating NPSLE from nonNP-SLE, the total parenchym MTR distributions result in the best results.

Secondly, we have shown that MTR distributions can be modelled by the family of stable distributions. Maximum likelihood estimation resulted in four parameters, and classification based on these improved on histogram measures based classification, although it did not reach quite the accuracies obtained in the multivariate LDPCA/MDA approach.

Thirdly, we have improved on quite a few aspects of the classification methodology. Instead of applying LDA to all histogram bin counts, we have avoided the collinearity problem by first reducing the histogram data to a few principal components. In contrast to the earlier approach by Dehmeshki et al. (2002), where this problem has not been noticed, our approach is sound, robust, and routinely available in all statistical software packages. The notes at the end of this thesis contain more comments regarding improvements of the histogram density estimate by kernel smoothing and consider alternative methods of data reduction.

Finally, we have applied our distance-based analysis to this data. Quantifying differences between univariate MTR distributions has been achieved by calculating one-dimensional Wasserstein distances. Since these transportation problems are one dimensional, they are very efficient to solve and no approximations are needed. Reconstruction of the distances by multidimensional scaling, and classification in the MDS coordinate space resulted in high classification accuracies. With about six di-

mensions, classification in the full contrast was possible with an improvement of more than 10%, compared to LDPCA/MDA, and also when discriminating healthy controls from all SLE patients, the distance-based approach seems superior. Discriminating between controls and NPSLE, or between NPSLE and nonNP-SLE patients, the LDPCA/MDA was still slightly better.

We conclude that the distance-based approach is very promising in the detection and classification of SLE. It would be interesting to see whether some of the six reconstruction dimensions can be attributed or correlated to some accessible covariate. However, this would need a much larger dataset than presently available.

4.3.7 Tables: Classification accuracies

Tissue	Method			
	Histogram	MDA†	Stable fit	Wasserstein
EPAR	0.794 / 0.870	0.912 (4) / 0.926	0.882 / 0.923	0.882 (4) / 0.909
EWM	0.824 / 0.846	0.853 (5) / 0.898	0.794 / 0.874	0.853 (1) / 0.895
EGM*	0.727 / 0.868	0.970 (5) / 0.959	0.879 / 0.906	0.939 (1) / 0.925
PAR	0.794 / 0.870	0.912 (11) / 0.968	0.853 / 0.912	0.912 (2) / 0.895
WM	0.824 / 0.842	0.882 (5) / 0.912	0.882 / 0.874	0.853 (1) / 0.902
GM*	0.818 / 0.872	0.939 (5) / 0.936	0.879 / 0.895	0.939 (2) / 0.921

Table 4.5: Cross-validated classification accuracies in the H-N contrast. First number: accuracy, second number: area under the receiver-operator characteristic. The number of principal or MDS components resulting in maximal cross-validated accuracy is given in brackets. *: MLE estimation did not converge for one subject, results therefore for a slightly smaller dataset. †: a maximum of 15 PCs tried.

Tissue	Method			
	Histogram	MDA†	Stable fit	Wasserstein
EPAR	0.600 / 0.683	0.800 (15) / 0.787	0.600 / 0.697	0.771 (4) / 0.843
EWM	0.600 / 0.687	0.771 (11) / 0.647	0.629 / 0.760	0.743 (4) / 0.767
EGM*	0.647 / 0.571	0.794 (6) / 0.793	0.647 / 0.754	0.794 (5) / 0.829
PAR	0.629 / 0.747	0.829 (8) / 0.817	0.600 / 0.740	0.800 (5) / 0.810
WM	0.514 / 0.620	0.714 (5) / 0.710	0.629 / 0.783	0.743 (6) / 0.770
GM*	0.588 / 0.714	0.765 (6) / 0.757	0.618 / 0.711	0.706 (7) / 0.714

Table 4.6: Cross-validated classification accuracies in the N-O contrast. First number: accuracy, second number: area under the receiver-operator characteristic. The number of principal or MDS components resulting in maximal cross-validated accuracy is given in brackets. *: MLE estimation did not converge for one subject, results therefore for a slightly smaller dataset. †: a maximum of 15 PCs tried.

Tissue	Method			
	Histogram	MDA†	Stable fit	Wasserstein
EPAR	0.590 / 0.705	0.769 (3) / 0.892	0.744 / 0.847	0.821 (1) / 0.913
EWM	0.872 / 0.861	0.795 (2) / 0.842	0.744 / 0.797	0.846 (6) / 0.797
EGM⋆	0.641 / 0.774	0.795 (2) / 0.826	0.769 / 0.876	0.821 (1) / 0.882
PAR	0.629 / 0.758	0.769 (3) / 0.871	0.821 / 0.882	0.821 (1) / 0.887
WM	0.821 / 0.916	0.795 (2) / 0.850	0.769 / 0.800	0.821 (1) / 0.918
GM⋆	0.744 / 0.871	0.769 (2) / 0.847	0.769 / 0.850	0.846 (6) / 0.934

Table 4.7: Cross-validated classification accuracies in the H-O contrast. First number: accuracy, second number: area under the receiver-operator characteristic. The number of principal or MDS components resulting in maximal cross-validated accuracy is given in brackets. ⋆: MLE estimation did not converge for one subject, results therefore for a slightly smaller dataset. †: a maximum of 15 PCs tried.

Tissue	Method			
	Histogram	MDA†	Stable fit	Wasserstein
EPAR	0.704 / 0.762	0.833 (7) / 0.890	0.796 / 0.865	0.833 (1) / 0.914
EWM	0.796 / 0.862	0.815 (4) / 0.833	0.815 / 0.829	0.833 (1) / 0.896
EGM⋆	0.736 / 0.814	0.830 (2) / 0.876	0.793 / 0.865	0.849 (1) / 0.906
PAR	0.704 / 0.793	0.852 (7) / 0.881	0.796 / 0.881	0.870 (6) / 0.916
WM	0.778 / 0.881	0.852 (10) / 0.874	0.796 / 0.833	0.833 (1) / 0.902
GM⋆	0.793 / 0.870	0.830 (5) / 0.923	0.755 / 0.862	0.849 (1) / 0.885

Table 4.8: Cross-validated classification accuracies in the H-S contrast. First number: accuracy, second number: area under the receiver-operator characteristic. The number of principal or MDS components resulting in maximal cross-validated accuracy is given in brackets. ⋆: MLE estimation did not converge for one subject, results therefore for a slightly smaller dataset. †: 15 PCs used.

4.4 Alzheimer's disease

Alzheimer's disease (AD) is one of the most common diseases affecting the elderly and will pose a large psychological and economical burden to Western society in the future. It is characterized by an excessive accumulation of amyloid-beta ($A\beta$) protein in neuronal synapses, cell bodies, and cerebral arteries in the form of pathological plaques. This induces inflammatory processes involving glial cells, neurofibrillary tangles involving tau protein from the cytoskeleton of affected neurons, and vascular lesions caused by arterial deposits (Figure 4.7). The result is neurodegenerative loss of neurons and brain atrophy, which leads to increasingly impaired cognitive ability, severe dementia, and ultimately death.

A few key mechanisms of the genesis of AD have been unraveled. There exists genetic predispositions with regard to the synthesis of $A\beta$ protein. Certain mutations in the amyloid protein precursor gene (APP) or the presilin (PS1/PS2) genes

Tissue	Method			
	Histogram	MDA†	Stable fit	Wasserstein
EPAR	0.556	0.667 (10)	0.611	0.722 (4)
EWM	0.574	0.648 (3)	0.630	0.741 (6)
EGM*	0.547	0.679 (2)	0.660	0.717 (4)
PAR	0.574	0.667 (7)	0.593	0.796 (6)
WM	0.574	0.667 (3)	0.630	0.722 (6)
GM*	0.604	0.660 (1)	0.623	0.736 (6)

Table 4.9: Cross-validated classification accuracies in the full contrast. The number of principal or MDS components resulting in maximal cross-validated accuracy is given in brackets. *: MLE estimation did not converge for one subject, results therefore for a slightly smaller dataset. †: a maximum of 15 PCs tried.

account for an estimated 10%-15% of early-onset cases of AD. Although this does not explain the development of AD in most humans, it has led to the development of small animal models that exhibit characteristic features of AD. The current state of these transgenic mouse models of AD has been reviewed in Muskulus, Scheenstra, Braakman, Dijkstra, Verduyn-Lunel, Alia, de Groot and Reiber (2009). The importance of these mouse models stems from the fact that they allow to study the development of $A\beta$ plaques in vivo, in large populations of animals, and over the course of time. Quantitative MR image analysis has detected significant changes in T_2 distributions of transgenic mice compared to normal controls (see loc. cit. for references); the T_2 parameter generally characterizes the local composition of tissue (an excellent overview of changes in T_2 due to pathologies has been given by Bottomley et al. (1987)). Interestingly, there exist pharmacological possibilities to slow and reduce the amyloid burden of the brain, so some of the main timely research questions about AD are the following: Can AD be reliably detected by studying T_2 distributions in a MR scanner? When and under which circumstances is this possible? Can disease burden and the severity of the disease be thereby quantified? Due to the obvious ethical problems, this kind of research is presently only possible with trangenic mouse models. Translated to this domain, the main question is then: From what age on can AD be detected in trangenic mice?

Unfortunately this is a difficult question. Not only is a large number of animals required for a reliable assessment of this possibility, it is also necessary to submit the animals to repeated MR scans of their brains. As this poses severe logistic problems, there are very few studies at present who have actually obtained such data, and usually these have only considered two distinct points in time (ages of the mice). Moreover, the analysis of this data has up to now focussed on specific regions in the brain, where the effects of AD are most pronounced. These include the hippocampus, the thalamus, the corpus callosum and a few other brain areas. In almost all studies the region of interest (ROI) was small and manually delinated in the brains,

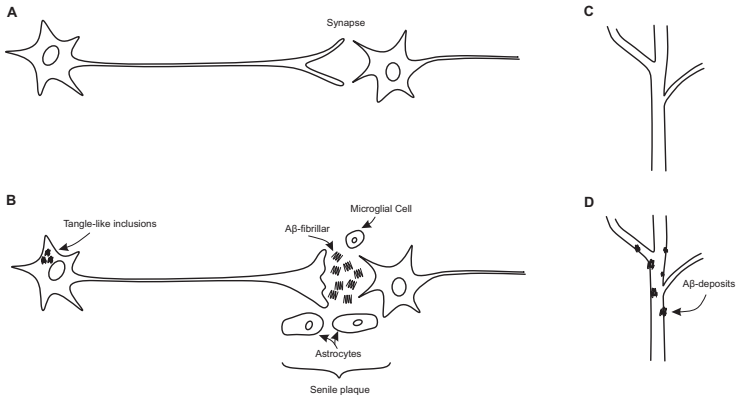


Figure 4.7: Structural changes characteristic of Alzheimer's disease.

which introduces a nontrivial source of bias and uncertainty. To overcome this, in (Falangola et al., 2005) nonlinear image registration was performed to align individual brain images to a common average brain image. The ROIs were still delineated manually, but only once on the average mouse brain.

Here we discuss preliminary results that were obtained by registering a complete three-dimensional MR scan of mice brains to a common brain atlas. This atlas does not only provide a common coordinate system, but has also been segmented, such that identification of subregions in the mice brains can now be achieved objectively and automatically. Since the total dataset, covering up to almost a hundred mice, measured at more or less regularly spaced points in time, will only be available in a few years, this section only discusses a limited, initial subset of nine mice, measured at one time point each.

4.4.1 Materials

Six wildtype (WT) and three transgenic (TG) mice were scanned in a Bruker MR system at an age between 16 and 17 months. A $proton - density$ T_1 imaging volume consisted of $256 \times 256 \times 256$ voxels that were used to align the imaging coordinate system to the Shiva brain atlas. A second T_2 imaging sequence was obtained for six transversal slices, more or less regularly spaced. This sequence resulted in 12 measurements of the T_2 relaxation process at 8.5 ms intervals. The resulting 12 images were fitted voxel-wise to the exponential model

$$I(t) = \alpha + \gamma \exp(-t/T_2), \tag{4.3}$$

minizing the residual least squares by a nonlinear Gauss-Newton iteration (algorithm nls in R). Four different initial conditions were tried if the iteration did not converge within a prescribed minimal stepsize of 10^{-5} and finally a few (less than 0.05 percent) of voxels were excluded. Also, voxels whose estimated standard deviation was larger than their T_2 value were excluded (again less than 0.05 percent).

4.4.2 Results

The resulting T_2 values were evaluated for either the total brain parenchym (PAR) or only the hippocampus (HC) and Figure 4.8 and Figure 4.9 show the corresponding histogram estimates for all mice.

Interestingly, the inter-group variability is extremely low for the transgenic mice, compared to the wildtype which exhibits more than a 25-fold increase in variance. As all the mice (apart from one) were scanned within a few days of each other, this phenomenon is unlikely to be caused by drift of the measurement apparatus, and has to be considered genuine. However, as is clear from the graphs, simple histogram based measures will not allow to discriminate between the two groups, since there is too much overlap between peak locations, peak heights and peak widths.

For the distance-based analysis, reconstructions of the mice in MDS space are shown in Figure 4.10. Although the number of mice is too small to allow for a reliable assessment, one general feature is visible. The MDS reconstruction for the HC dataset (Panel A) is essentially one-dimensional and corresponds roughly to the difference in peak location between the T_2 distributions of the mice. Classification with such data is difficult if not impossible. On the other hand, the PAR dataset (Panel B) hints at higher dimensionality, which corresponds to changes in the shape of the T_2 distributions (higher moments) instead of the first two moments only. There is not enough data to elaborate on these findings at the moment, but this will allow for an interesting future application of distance-based methods.

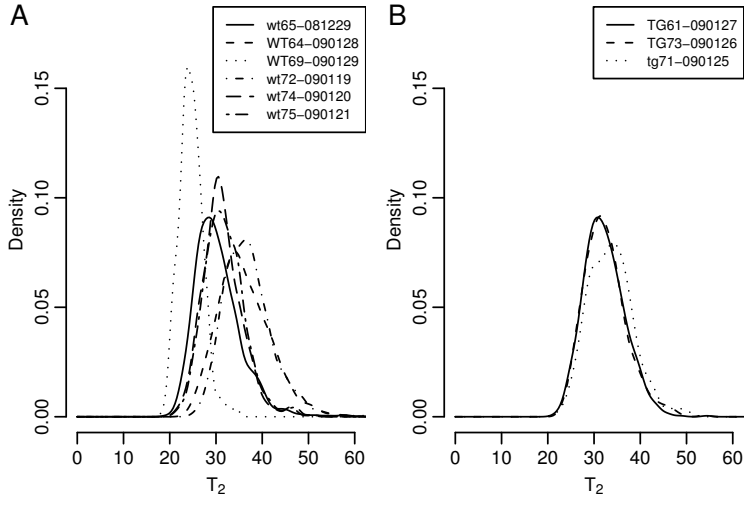


Figure 4.8: T_2 distributions of sample HC dataset. A: Wildtype controls. B: Transgenic mice. Curves based on a histogram estimate with 512 bins.

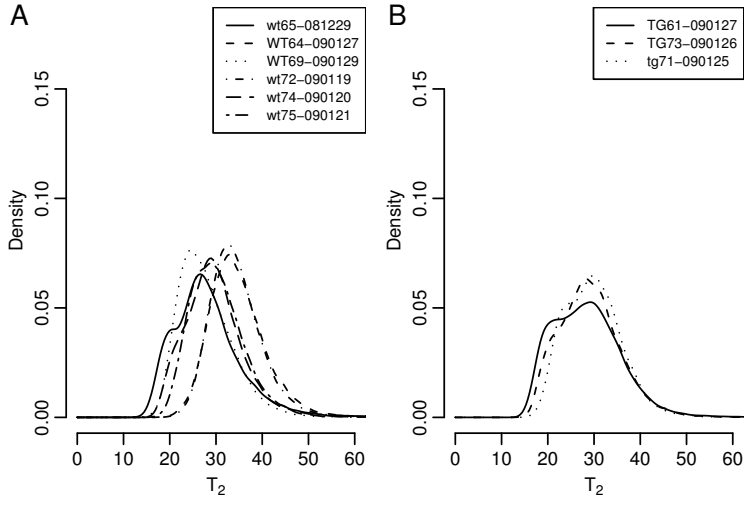


Figure 4.9: T_2 distributions of sample PAR dataset. A: Wildtype controls. B: Transgenic mice. Curves based on a histogram estimate with 512 bins.

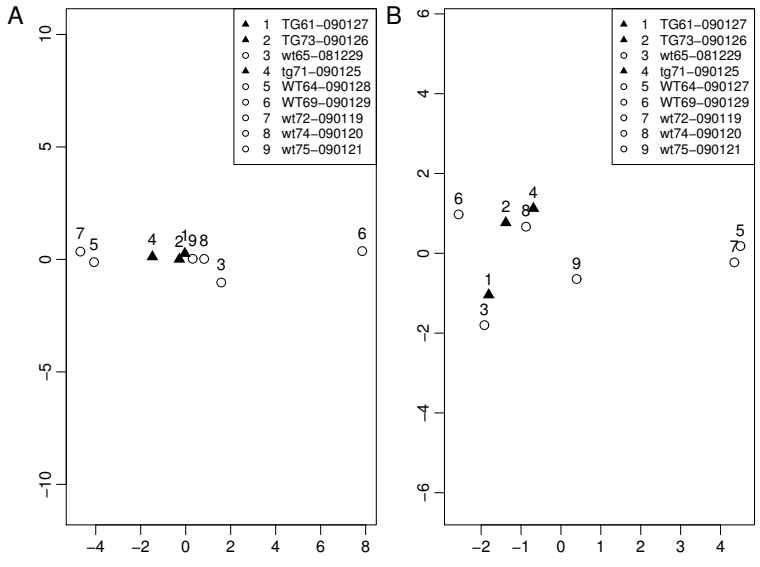


Figure 4.10: 2D MDS reconstruction of 1D Wasserstein distances. A: HC dataset (hippocampus). B: PAR dataset (total brain).

Box 10. Alzheimer's disease

- Alzheimer's disease can, to some extent, be conveniently studied in standardized, transgenic mouse models.
- The main questions are how to detect the disease (1) reliably and (2) as early as possible, and how to track its progress and quantify disease load (3). There exist pharmaceutical treatment options that could improve quality of life of patients if these problems were solved.
- Obtaining more than one time point should theoretically allow for much more sensitive detection. The methodological and practical tools for this exist, but the necessary data is not available yet.
- Is the variability of T_2 values decreased in transgenic (APP/PS1) mice, when large areas of the brain are pooled?

Chapter 5

Deformation morphometry

Abstract

The Rayleigh test is a popular one-sample test of randomness for directional data on the unit circle. Based on the Rayleigh test, Moore developed a nonparametric test for two-dimensional vector data that takes vector lengths into account as well, which is generalized to arbitrary dimensions. In the important case of three-dimensional data the asymptotic distribution is given in closed form as a finite combinatorial sum. This reduces the computational effort considerably. In particular, when analyzing deformation fields arising in nonlinear brain registration, the generalized Moore-Rayleigh test offers an efficient alternative to conventional permutation testing for the initial screening of voxels.

Simulation results for a few multivariate distributions are given and the test is applied to magnetic resonance images of mice with enlarged ventricles. Compared with the permutation version of Hotelling's T^2 test its increased power allows for improved localization of brain regions with significant deformations.

5.1 Overview

This chapter is an excursion into an important method that precedes the application of optimal transportation distances, namely, the localization of brain regions with significant deformations. It stands somewhat outside the general scope of this thesis but has been included for its innovative character and to illustrate the complexity of actual applications. After localization, parameter distributions in the highlighted brain regions can then be analyzed by optimal transportation measures as in the preceding chapter.

After introducing the problem in Section 5.2, the novel statistical test is described in Section 5.3. Its application in the two-sample case is discussed in Section 5.4, and illustrated by simulations in Section 5.5. Section 5.6 describes an application in transgenic mice. Finally, the method is discussed in Section 5.7.

5.2 Introduction

Consider the following illustrating example. In the voxel-based analysis of brain deformations, individual brain volumes are mapped to a reference brain image by a

nonlinear transformation (Kovacevic et al., 2004). This process of image registration results in a three-dimensional vector field of displacement vectors. The significance of local deformations between groups of subjects, usually a treatment and a control group, can be tested by either considering the Jacobian of the deformation field, or testing the displacement vectors directly (Chung et al., 2001). In the latter case, if one assumes that variations between subjects are given by a Gaussian random field, Hotelling's T^2 statistic can be used to test for significant differences between groups (Cao and Worsley, 1999). Its value is the squared sample Mahalanobis distance, estimated from the pooled covariance matrix, and the test assumes normality of the population of deformation vectors and equal covariances for the two groups. If these assumptions are not met, the T^2 test is known to fail gracefully, i.e., it will still be approximately conservative and the loss in power for the alternative will not be too dramatic for moderate violations of the assumptions. However, it is preferable to analyze deformation fields nonparametrically, as deformations are likely to be skewed and nonnormal.

Permutation tests, with their minimal assumptions, are the usual method of choice for this two-sample problem (Chen et al., 2005; Nichols and Holmes, 2007). However, they also rely on a test statistic that is evaluated for each labelling ("permutation"), and the null hypothesis is that this statistic is distributed symmetrically around zero. The usual choice for the statistic is again Hotelling's T^2, so permutation tests are not nonparametric, but rather result in *adjusted* significance probabilities (Davison and Hinkley, 1997, Chap. 4.5). For example, as shown in Lehmann and Romano (2005, Chap. 5.9), the adjusted one-dimensional version of the T^2 test, i.e., the permutation version of the classic t-test, is the uniformly most powerful test for the Gaussian alternatives with fixed variance, but fails to be uniformly most powerful against other alternatives.

A more serious practical problem is that, even for small sample sizes, the number of permutations to consider for an exact test is prohibitively large. Especially so, if the number of voxels, i.e., the number of tests, is on the order of hundreds of thousands, as common in neuroimaging applications. Therefore, in current analyses one often limits the data to only 10 000 or less random labellings per voxel, at the expense of increasing the simulation error. Moreover, correcting for multiple comparisons imposes severe lower bounds on the numbers of labellings needed per voxel for testing at realistic significance levels, i.e., on the sample size and simulation time. Particularly for small sample sizes that occur in prospective studies, permutation tests cannot resolve low enough significance probabilities to allow for strong control of the family-wise error. Even the modern, liberal approach of limiting the False Discovery Rate (Benjamini and Hochberg, 1995; Schwartzman et al., 2009) does often not lead to useful results in these datasets. This implies that although permutation tests are elegant and theoretically well understood, they can not be used on a routine basis (e.g. in a clinical setting) to assess and quantify brain changes.

For these reasons, in the analysis of magnetic resonance (MR) images classical hypothesis testing is still unmatched in its efficiency and speed. In this article we describe a new nonparametric statistical test that allows to efficiently perform a large number of such tests on vector data. The two-sample version of the test is not provably conservative, but its advantage is that it can be used for the initial screening of voxels. It is sensitive enough to work even under the conservative Bonferroni correction. Voxels where the null hypothesis is rejected can then be analyzed further by this test under the permutation distribution of the data; alternatively a different test statistic can be employed.

This problem of testing one or more groups of vectors for distributional differences does not only arise in neuroimaging, but also in a number of other disciplines and diverse contexts, e.g. in geostatistics, human movement sciences, astronomy and biology. In the two-dimensional case, a natural nonparametric test for such problems has been given by Moore (1980), which we describe next. After generalizing this test to arbitrary dimensions, in Section 5.3.2 we focus on the three-dimensional case, being the most important one for applications.

5.3 The Moore-Rayleigh test

Let $X = (X_1, \ldots, X_N)$ be a finite sample of real k-vector-valued random variables

$$X_i = (X_{i,1}, \ldots, X_{i,k}). \quad (n = 1, \ldots, N). \tag{5.1}$$

If we assume that the X_i are independently drawn from a common absolutely continuous probability distribution with density $f : \mathbb{R}^k \to [0, \infty)$, then the null hypothesis is:

$$H_0 : \text{The probability density } f \text{ is } \textit{spherically symmetric}.$$

Consequently, this implies that the density f is spherically decomposable. It factors into the product of a radial density $p_r : [0, \infty) \to [0, \infty)$ and the uniform distribution on each hypersphere $rS^{k-1} = \{x \in \mathbb{R}^k \mid ||x|| = r\}$, such that $f(x) = p_r(||x||)/\text{vol}(||x||S^{k-1})$. We can then write $X_i = R_i U_i$, where $R_i \sim p_r$ and U_i is distributed uniformly on the k-dimensional unit sphere S^{k-1}. The latter distribution can be realized as the projection of a k-dimensional diagonal Gaussian distribution with equal variance in each coordinate. The sum $\sum_{i=1}^{N} X_i$, where the X_i are independently distributed according to a common, spherically symmetric distribution, is easy to interpret. It corresponds to a Rayleigh random flight (Dutka, 1985) with N steps, whose lengths are distributed according to p_r.

Scaling the vector-valued random variables X by the ranks of their lengths, the

distribution of the resultant vector

$$S_N = \sum_{i=1}^{N} \frac{iX_{(i)}}{||X_{(i)}||}, \tag{5.2}$$

where $X_{(i)}$ denotes the i-th largest vector in the sample (with ties being arbitrarily resolved), is independent of p_r; consequently, a test based on S_N is nonparametric. The test statistic of interest here is the asymptotically scaled length of the resultant,

$$R_N^* = \frac{||S_N||}{N^{3/2}}. \tag{5.3}$$

A large value of R_N^* for a given sample X from an unknown distribution indicates a deviation from spherical symmetry. This test was introduced by Moore (1980), who treated the two-dimensional case numerically, and has been used in neuroscience (Kajikawa and Hackett, 2005; Tukker et al., 2007; Richardson et al., 2008), human movement science (van Beers et al., 2004) and avian biology (Able and Able, 1997; Mcnaught and Owens, 20002; Burton, 2006; Chernetsov et al., 2006). In contrast to the Rayleigh test of uniformity (Mardia and Jupp, 2000, Chap. 10.4.1), where the X_i are constrained to lie on (alternatively, are projected onto) the unit sphere, in the Moore-Rayleigh test also the vector length influences the test statistic. This follows the observation of Gastwirth (1965), that differences in scale between two distributions will be mostly evident in their (radial) tails, i.e., when moving away from the mean. The interpretation of R_N^* is not so easy as in the Rayleigh test, however, where the test statistic is a measure of *spherical variance*.

Consider the projections

$$S_{N,j} = \sum_{i=1}^{N} \frac{iX_{(i),j}}{||X_{(i)}||}, \qquad (j = 1, \ldots, k). \tag{5.4}$$

A direct calculation shows that under the null hypothesis the variance of $X_{(i),j}/||X_{(i)}||$ is $1/k$, and that

$$\sigma^2 = \mathrm{var}(S_{N,j}) = N(N+1)(2N+1)/(6k). \tag{5.5}$$

As $E(S_{N,j})^3 = 0$ and $\sigma^2 < \infty$, the Lyapunov version of the Central Limit Theorem implies that the random variables $S_{N,j}$ approach Gaussian $\mathcal{N}(0, \sigma^2)$ distributions for large sample sizes N. Although the random variables $||S_{N,j}||$ are obviously not independent, by the same argument as in Stephens (1962) the corresponding distribution of $||S_N||^2/\sigma^2$ asymptotically approaches a χ^2_k distribution.

Let $\alpha_N = N^{3/2}$. The exact null distribution of $R_N = \alpha_N R_N^*$ in k dimensions,

$k \geq 2$, is given by

$$\mathrm{pr}\left(R_N \leq \alpha_N r; k\right) = r\left[\Gamma\left(\frac{k}{2}\right)\right]^{N-1} \int_0^\infty \left(\frac{rt}{2}\right)^{\frac{k-2}{2}} J_{\frac{k}{2}}(rt) \prod_{n=1}^{N} \frac{J_{\frac{k-2}{2}}(nt)}{(nt/2)^{\frac{k-2}{2}}}\, \mathrm{d}t, \quad (5.6)$$

where J_l denotes the Bessel function of order l; see (Lord, 1954).

5.3.1 The one-dimensional case

In one dimension, the Moore–Rayleigh statistic for the null hypothesis corresponds to a symmetric random walk with linearly growing steps,

$$S_N = \sum_{i=1}^{N} \gamma_i i, \qquad \text{where } \gamma_i = \pm 1 \text{ with equal probability.} \quad (5.7)$$

Proposition 4. The probability mass function $\mathrm{pr}\left(S_N = r\right) \stackrel{\mathrm{def}}{=} p(r, N)/2^N$ is given by the recurrence

$$p(r, N) = p(r - n, N - 1) + p(r + n, N - 1) \quad (5.8)$$

with initial condition $p(0, 0) = 1$ and $p(r, 0) = 0$ for $r \neq 0$.

Rewriting Eq. 5.7 as

$$\sum_{\{\gamma_i = +1\}} i = \frac{1}{2}\left(S_N + \frac{1}{2}N(N+1)\right), \quad (5.9)$$

where the sum runs over all step sizes $i \in \{1, \ldots, N\}$ that have positive sign γ_i, shows that the numbers $p(r, N)$ have a well-known combinatorial interpretation.

Proposition 5. The numbers $p(r, N)$ count the number of partitions of $\frac{1}{2}(r + \frac{1}{2}N(N+1))$ with distinct parts less or equal to N.

As before, denote the length of the resultant by $R_N = ||S_N||$. Its probability function $\mathrm{pr}(R_N = r)$ is given by

$$\mathrm{pr}(R_N = r) = \begin{cases} p(r, N)/2^{N-1} & \text{if } r > 0, \\ p(0, N)/2^N & \text{if } r = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (5.10)$$

In the sequel, we also need the random signs defined by

$$\epsilon_N = \prod_{i=1}^{N} \gamma_i, \tag{5.11}$$

conditional on the resultant S_N: Let $\epsilon_{r,N}$ denote the average sign of the partitions of $\frac{1}{2}(r + \frac{1}{2}N(N+1))$ with distinct terms less or equal to N, i.e.,

$$\epsilon_{r,N} \stackrel{\text{def}}{=} E(\epsilon_N \mid S_N = r). \tag{5.12}$$

Anticipating the two-sample Moore-Rayleigh test discussed in Section 5.4, we note the following:

Remark 3 (Relation to the Wilcoxon signed-rank test)**.** In the Wilcoxon signed-rank test for two paired samples X and Y of equal size $|X| = |Y| = N$, the null hypothesis is that the paired differences $Z_i = Y_i - X_i$ are distributed (independently and identically) symmetrically around zero (Wilcoxon, 1945). The test statistic is the sum $W_+ = \sum_{i=1}^{N} iI(Z_i > 0)$, where $I(\cdot)$ is an indicator function. Under the null hypothesis we have that $\mathrm{pr}(Z_i > 0) = \mathrm{pr}(Z_i < 0) = \frac{1}{2}$. Assuming that $\mathrm{pr}(X_i = Y_i) = 0$, which is fulfilled with probability 1 for continuous distributions, we can then identify $I(Z_i > 0) - I(Z_i < 0)$ with a random sign γ_i, such that

$$\begin{aligned}\sum_{i=1}^{N} \gamma_i i &= \sum_{i=1}^{N} iI(Z_i > 0) - \sum_{i=1}^{N} (1 - I(Z_i > 0))i \\ &= 2W_+ - \frac{1}{2}N(N+1).\end{aligned}$$

Therefore, testing for symmetry of the Z_i under the one-dimensional Moore-Rayleigh test is equivalent to the signed-rank Wilcoxon two-sample test of X and Y, with

$$\mathrm{pr}(W_+ = r) = \mathrm{pr}(S_N = 2r - \frac{1}{2}N(N+1), N).$$

This approach easily generalizes to more than one dimension.

Remark 4 (Testing for radial dependence)**.** Assume the density f decomposes spherically, such that $X_i = R_i U_i$, with $R_i \sim p_r$ and $U_i \sim u$, where $p_r(r) = \mathrm{pr}(||X_i|| = r)$ and $u(x) = \mathrm{pr}(X_i/||X_i|| = x)$. In one dimension, u can only attain the values $\{-1, +1\}$ and $u(\mp 1) = \mathrm{pr}(X_i \lessgtr 0)$. If the mean of f is zero, i.e., $E(X_i) = 0$, then $\mathrm{pr}(X_i > 0) = \mathrm{pr}(X_i < 0) = 1/2$, and this implies that f is (spherically) symmetric. The Moore-Rayleigh test, under the assumption that $X_i = R_i U_i$, therefore tests the null hypothesis that $E(X_i) = 0$.

On the other hand, assume that $E(X_i) = 0$. If the Moore-Rayleigh test finds a significant departure from uniformity, then this leads to the rejection of the hypoth-

esis that the density f decomposes in such way, i.e., to accept the alternative that the common distribution of the random variables X_i is conditional on the length $||X_i||$. In practice, centering $X = (X_1, \ldots, X_N)$ by the sample mean, the Moore-Rayleigh test could be used to detect such radial dependence. However, its power would be quite limited and it seems likely that directly testing for differences in the two tails $\{X_i > x\}$ and $\{X_i < -x\}$ will be more powerful.

5.3.2 The three-dimensional case

Taking derivatives, the distribution function of $R_N = \alpha_N R_N^*$, given in Eq. (5.6), reduces to the density

$$\mathrm{pr}\,(R_N = r) = \frac{2r}{\pi} \int_0^\infty t \frac{\sin rt/\alpha_N}{r} \prod_{n=1}^{N} \frac{\sin nt}{nt}\, \mathrm{d}t \tag{5.13}$$

in the three-dimensional case ($k = 3$). This formula can alternatively be derived by using characteristic functions; see Eq. 16 in Dutka (1985). The oscillating integral in Eq. (5.13) can be evaluated by numerical quadrature, but it is difficult to calculate its tail accurately. Another approach to evaluate this integral is based on a finite series representation, following an idea originally due to G. Pólya. Let $N_{\max} = N(N+1)/2$. If we expand $\sin(nt) = (e^{nt} - e^{-nt})/2i$ and integrate the oscillating integral in Eq. (5.13) by parts $N-2$ times as in Borwein and Borwein (2001), a simple but tedious calculation (which we omit) results in the following representation:

Theorem 2. The probability density of R_N^* under the null hypothesis can be evaluated as

$$\mathrm{pr}\,(R_N^* = r) = \frac{2rN^3}{N!(N-2)!} \sum_{\substack{k\in\mathbb{N}:\\ \alpha_N r < k \le N_{\max}}} \epsilon_{k,N} (\alpha_N r - k)^{N-2}, \tag{5.14}$$

where $\epsilon_{k,N}$ is given by Eq. 5.12.

This is a generalization of Treolar's representation for the random flight with equal step sizes (Dvorák, 1972). We see that, interestingly, the density of the three-dimensional case can be expressed in terms of statistical properties of the one-dimensional case. Integrating Eq. 5.14 term-by-term from r to infinity, we have the following corollary.

Corollary 1. The cumulative distribution function of R_N^* under the null hypothesis can be evaluated as

$$\mathrm{pr}\,(R_N^* \le r) = 1 - \frac{2}{N!N!} \sum_{\substack{k\in\mathbb{N}:\\ \alpha_N r < k \le N_{\max}}} \epsilon_{k,N} (\alpha_N r - k)^{N-1} (\alpha_N r(1-N) - k). \tag{5.15}$$

Table 5.1: Critical values of Moore-Rayleigh statistic in 3D

Sample size	Probability			-Log(Probability)						
N	0·100	0·010	0·001	4	5	6	9	12	15	18
2	1·013	1·056	1·061							
3	0·973	1·100	1·138	1·150	1·153	1·155				
4	0·948	1·116	1·189	1·222	1·237	1·244	1·250			
5	0·930	1·124	1·221	1·275	1·304	1·321	1·338	1·341	1·342	
6	0·916	1·129	1·245	1·314	1·357	1·384	1·418	1·427	1·429	
7	0·905	1·132	1·262	1·344	1·398	1·435	1·488	1·505	1·510	1·511
8	0·897	1·133	1·275	1·368	1·432	1·477	1·549	1·576	1·586	1·588
9	0·890	1·134	1·284	1·387	1·460	1·513	1·603	1·640	1·656	1·659
10	0·885	1·134	1·292	1·402	1·483	1·543	1·649	1·698	1·720	1·726
12	0·877	1·133	1·303	1·426	1·519	1·590	1·727	1·797	1·834	1·844
14	0·871	1·133	1·310	1·443	1·545	1·626	1·788	1·879	1·931	1·946
16	0·866	1·132	1·316	1·455	1·565	1·654	1·838	1·947	2·013	2·033
18	0·863	1·132	1·320	1·464	1·580	1·675	1·878	2·003	2·083	2·108
20	0·860	1·131	1·323	1·472	1·593	1·693	1·911	2·051	2·144	2·174
30	0·851	1·129	1·331	1·493	1·629	1·746	2·016	2·209	2·350	2·399
40	0·847	1·128	1·335	1·503	1·647	1·771	2·071	2·294	2·467	2·529
50	0·844	1·127	1·337	1·509	1·657	1·787	2·103	2·347	2·540	2·612
60	0·843	1·126	1·338	1·513	1·664	1·797	2·125	2·382	2·590	2·668
∞	0·834	1·123	1·345	1·532	1·697	1·846	2·233	2·559	2·847	3·108

In particular, $\mathrm{pr}(R_N^* > (N+1)/(2\sqrt{N})) = 0$.

Note that because of the representation (5.15) for smaller r successively more and more terms enter the sum in the calculation of $\mathrm{pr}\,(R_N^* > r)$. The numerical accuracy is therefore higher for larger r, i.e., in the tail of the distribution.

The representations (5.14) and (5.15) therefore allow the efficient computation of exact significance probabilities for the test statistic R_N^* for small to moderately large sample sizes N (e.g., for $N \lesssim 60$ under double precision IEEE 754 arithmetic). This restriction on the sample size is only due to numerical accuracy; for larger N approximations of the Gamma function can be used.

Remark 5 (What is tested by the Moore-Rayleigh test?)**.** As in Remark 4, assume that $X_i = R_i U_i$, with $R_i \sim p_r$ and $U_i \sim u$, where $p_r(r) = \mathrm{pr}(||X_i|| = r)$ and $u(x) = \mathrm{pr}(X_i/||X_i|| = x)$ are arbitrary. If $E(X_i) = 0$, this implies $E(U_i) = 0$, and suggests that $\sum_i U_i \approx 0$ for a sample. More precisely, an upper bound for the variance of the test statistic R_N^* is realized by the one-dimensional Moore-Rayleigh null hypothesis, whose distribution is similar to the null hypothesis of the three-dimensional case (confer Figure 5.5). Therefore, as in the one-dimensional case, the Moore-Rayleigh test under the assumption of radial decomposability tests mostly for differences in location. Note that symmetry of the U_i, i.e., $\mathrm{pr}(U_i = u) = \mathrm{pr}(U_i = -u)$, implies

that $E[\sum_i U_i] = 0$. Thus, under the assumption of decomposability, testing for spherical symmetry and testing for symmetry are approximately equivalent, i.e., the Moore-Rayleigh test will not be sensitive to deviations from spherical uniformity if the underlying distribution is merely symmetric or mean-centered. This is actually an advantage when the Moore-Rayleigh test is considered as a two-sample test (see below).

5.3.3 Power estimates

To evaluate the performance of the three-dimensional Moore-Rayleigh test (MR3), power functions for a number of distributions were obtained by Monte-Carlo simulation. These show the fraction of rejections of the null hypothesis for a specific distribution, significance level α, and sample size N. The left panel of Figure 5.1 shows the power function for a family of diagonal Gaussian distributions with unit variances, shifted away from zero (along the z-axis) a constant distance $\mu \geq 0$. Each point power estimate was obtained by 1000 realizations of the distributions and represents the fraction of significance probabilities ("p-values") less than the nominal significance level α. The test was performed on $N = 10$ randomly drawn samples, and is compared to Hotelling's (non-randomized) T^2 one-sample test of location (Hotelling, 1931), as implemented in the R package ICSNP (Nordhausen et al., 2007), and to the spherical uniformity permutation test of Diks and Tong (1999), under 10^4 resamplings. The test statistic of the latter is an U-estimator of the difference between two probability distributions of vectors, calculated by a Gaussian kernel with a bandwidth parameter. The choice of the proper bandwidth is the subject of ongoing research; we show results for the two bandwiths $b_1 = 0.25$ and $b_2 = 2.5$, and denote the corresponding tests by "Diks1" and "Diks2", respectively.

In comparison with the T^2 test, MR3 shows larger power, an effect that is more pronounced for lower significance levels. It is thus a more sensitive measure of changes in location. Note that this does not contradict the well-known optimality of Hotelling's T^2 test for the family of multivariate Gaussian distributions, since in the calculation of T^2 the covariance matrix needs to be estimated from the data. In the special case of equal covariances considered here, the Moore-Rayleigh test can therefore exhibit larger power. Also note that the test of Diks & Tong can be more powerful than the MR3 test, but as its results depend strongly on the bandwidth parameter, it is difficult to apply it routinely.

In Figure 5.2, power functions are shown for a family of diagonal Gaussian distributions where the standard deviation of one axis was varied from $\sigma = 0.1$ to $\sigma = 5.0$ in steps of 0.1, the other standard deviations were kept at unity. As expected from Remark 5, the MR3 test performs poorly for this specific violation of spherical symmetry. The remaining symmetry in the distribution means that although sample points are now increasingly less concentrated on one axis, on average their contribu-

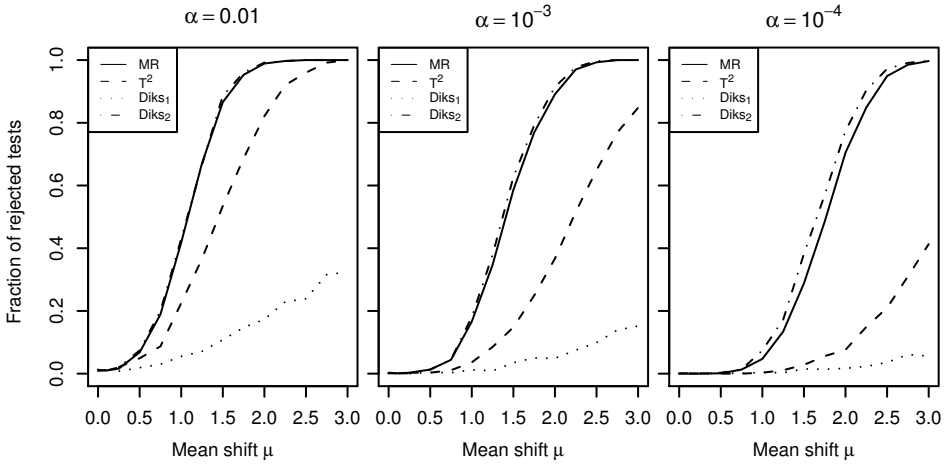


Figure 5.1: Estimated power functions for the family of Gaussian distributions with covariance matrix the identity and mean shifted a distance μ away from zero. Sample size $N = 10$.

tions to the resultant length still mostly cancel each other. Analogously, the T^2 test has only nominal power for the anisotropic multivariate Gaussian, being a test of location only. Note that MR3 shows slightly more power than the nominal significance levels α for $\sigma \neq 1$, as do the Diks1 and Diks2 tests.

To assess the effect of asymmetry of the sample distribution, we employ the Fisher distribution, also known as the Fisher–Bingham three-parameter distribution. This is the $k = 3$ case of the k-dimensional von–Mises Fisher distributions commonly used in directional statistics (Mardia and Jupp, 2000, Chap. 9.3.2). It is a singular distribution on the hypersphere S^{k-1} whose density $f(x), x \in \mathbb{R}^k$, is proportional to $e^{\lambda \xi^t x}$, where ξ^t denotes the transpose of ξ. The mean direction ξ is constrained to be a unit vector, and $\lambda \geq 0$ is a concentration parameter. Without restricting generality, we let $\xi = e_k$ be the unit vector in the k-th dimension, so $f \sim e^{\lambda x_k}$ only depends on the last coordinate, and we are left with a one-parameter family of distributions.

Following Ulrich (1984) and Wood (1994), a random variate distributed according to the von–Mises Fisher distribution is obtained by generating a random variate W for the last coordinate, by the density proportional to

$$e^{\lambda w}(1 - w^2)^{(k-3)/2}, \quad w \in (-1, 1), \quad k \geq 2,$$

and a $k - 1$ dimensional variate V uniformly distributed on the hypersphere S^{k-2}. The vector

$$X = (\sqrt{1 - W^2} \cdot V^t, W) \in \mathbb{R}^k$$

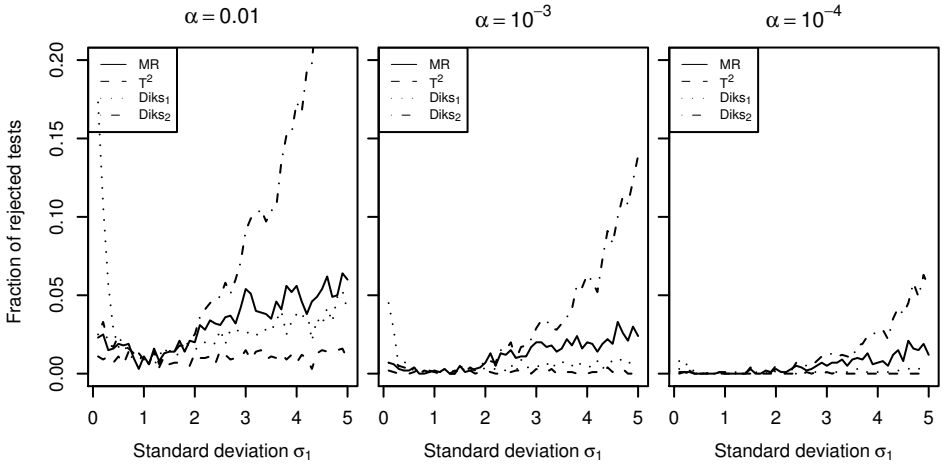


Figure 5.2: Estimated power functions for the family of Gaussian distributions, varying the standard deviation σ of a single axis. Sample size $N = 10$. Note the small range of the power.

then has the desired density. In $k = 3$ dimensions the former can be achieved by integrating the distribution function of W directly. Choosing a uniform variate U on the interval $[-1, 1]$, a random variate W is clearly given by

$$W = \frac{1}{\lambda} \log(2U \sinh \lambda + e^{-\lambda}).$$

We denote the Fisher distribution with concentration parameter λ (and with the choice $\xi = e_3$) by $F3_\lambda$. To avoid degeneracies due to its singular character, the $F3_\lambda$ distribution is multiplied by $1 - Z$, where $Z \sim \mathcal{N}(0, 0.1)$. Figure 5.3 shows three examples of $N = 1000$ random variates obtained from these "scattered" Fisher distributions for distinct values of the concentration parameter λ, with increasingly larger deviation from the uniform distribution.

The power of MR3 for the family of scattered Fisher distributions, varying the concentration parameter, is comparable to the power of the other tests (not shown). Let us now consider a mixture, where the samples are chosen either (i) from the uniform distribution on the unit sphere, or (ii) from the scattered Fisher distribution $2F3_5$. The probability $0 \leq p \leq 1$ for each sample vector to be chosen from the second distribution is the parameter of this family of mixture distributions, with larger p indicating stronger deviations in uniformity for the larger vectors. Figure 5.4 depicts the estimated power for this family under variation of the mixture probability p. Compared to the T^2 test, the MR3 test is seen to be more sensitive to these specific departures from uniformity. It should be noted that reversing the situation, e.g., by

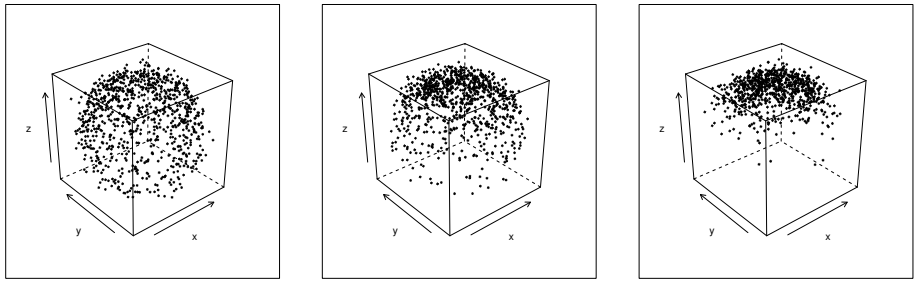


Figure 5.3: Scattered Fisher distribution, visualized by 1000 randomly drawn points in the unit-cube. Left: Concentration parameter $\lambda = 1$. Middle: Concentration parameter $\lambda = 2.5$. Right: Concentration parameter $\lambda = 5$.

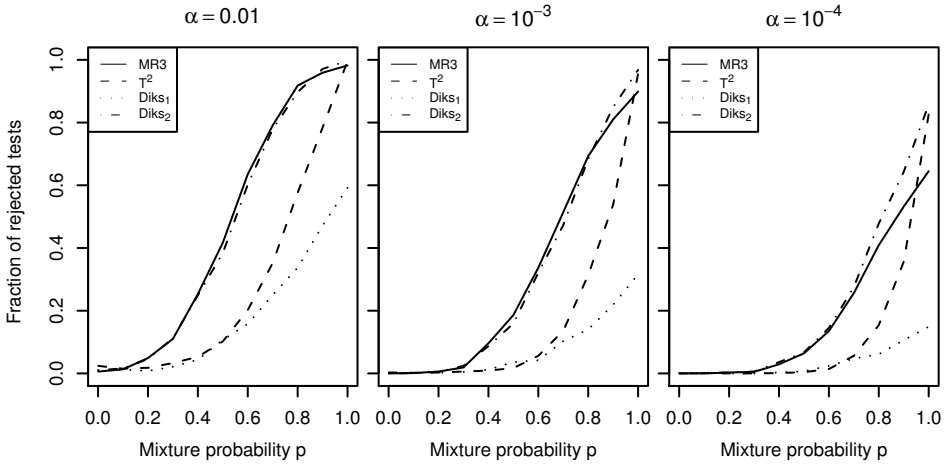


Figure 5.4: Estimated power functions for the mixture of the scattered Fisher distribution $2F3_5$ with the uniform distribution on the sphere S^2, varying the mixture probability p that a sample vector arises from the first distribution. Sample size $N = 10$.

considering $F3_5/2$ instead of $2F3_5$, such that the smaller vectors exhibit deviations from uniformity, the power of MR3 becomes less than that of the T^2 test (not shown).

5.4 The two-sample test

The most interesting application of the Moore-Rayleigh test is the two-sample problem. There, we are given two vector-valued random variables

$$X = (X_1, \ldots, X_N) \quad \text{and} \quad Y = (Y_1, \ldots, Y_N), \tag{5.16}$$

and we assume that they are identically and independently distributed with densities f and g, respectively. The differences $Y_j - X_i$ are then distributed according to the convolution $g * (f^-)$, whose density is

$$\operatorname{pr}(Y - X = x) = \int \operatorname{pr}(Y = u) \operatorname{pr}(X = u + x) \ \mathrm{d}^k u. \tag{5.17}$$

Under the null hypothesis that the X_i and Y_j come from a common probability density f, this reduces to the *symmetrization* of f, with density

$$\operatorname{pr}(Y - X = x) = \int \operatorname{pr}(X = u) \operatorname{pr}(X = u + x) \ \mathrm{d}^k u. \tag{5.18}$$

If the probability density f is *spherically symmetric* around its mean μ, i.e., uniform on each hypersphere $\{x \mid ||x - \mu|| = r\}$, then Eq. (5.15) gives the significance probability of a deviation from the null hypothesis. In particular, this applies when f is assumed to arise from a multivariate normal distribution, justifying the use of the Moore-Rayleigh statistic in many practical situations.

5.4.1 Testing for symmetry

In general, however, the distribution of $h = f * (f^-)$ is merely *symmetric*, i.e., $h(x) = h(-x)$ for all $x \in \mathbb{R}^k$. This follows from

$$\int \operatorname{pr}(X = u) \operatorname{pr}(X = u + x) \ \mathrm{d}^k u = \int \operatorname{pr}(X = u) \operatorname{pr}(X = u - x) \ \mathrm{d}^k u. \tag{5.19}$$

The following demonstrates the difference.

Example 1. Consider the symmetric singular distribution $B_x \stackrel{\text{def}}{=} \frac{1}{2}\delta_x + \frac{1}{2}\delta_{-x}$, where δ_x is the Dirac measure concentrated at the point $x \in \mathbb{R}^k$. The distribution B_x leads to an embedding of the one-dimensional Moore-Rayleigh null distribution in three dimensional space. Its realizations take values x and $-x$ with equal probability, and it is not spherically symmetric. As it is, B_x is neither absolutely continuous, nor can it arise as the the symmetrization of a distribution. Nevertheless, it is a model for a distribution that can arise in practice: First, the Dirac measures can be approximated, e.g., by a series of Gaussian distributions with decreasing variance. Secondly, con-

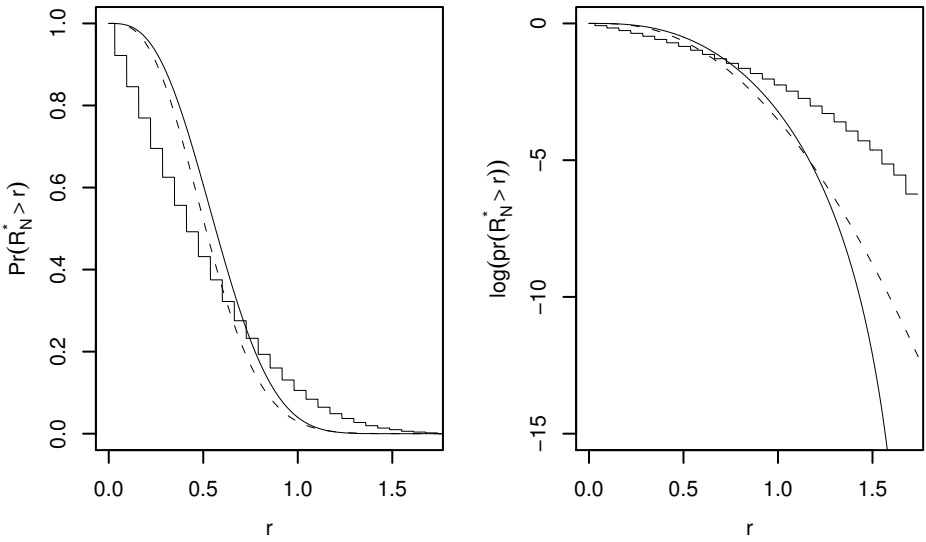


Figure 5.5: Comparison of significance probabilities for resultant lengths of spherically symmetric (smooth curve) and one-dimensional symmetric random walk (piecewise-linear curve) in three dimensions for $N = 10$ steps. Dotted curve shows the asymptotic case.

sider the singular distribution B^x that is concentrated on a line $\{\lambda x \mid \lambda \in \mathbb{R}\} \subseteq \mathbb{R}^k$ through the origin. Applying the Moore-Rayleigh test to B^x is equivalent to calculating the test statistic from B_1, since B^x is invariant under symmetrization and is projected, before ranking, to the sphere $S^0 = \{-1, +1\}$.

The distribution B_1 is a representative of the class of "fastest growing" random flights in three dimensions, since any other distribution of increments has less or equal probability to reach the highest values of the test statistic. On the other hand, the uniform distribution on the sphere, which represents the null hypothesis of the Moore-Rayleigh test statistic R_N^*, will attain lower values of R_N^* with higher probability, as the uniform random walk can do "orthogonal" steps that increase the distance from the origin faster than in B_1 (on the average). To be specific, if the finite sample X is distributed according to B_1, the n-th step of the scaled random walk either increases or decreases the distance from the origin by n (when crossing the origin, there is an obvious correction to this). However, if the n-th step were taken in a direction that is orthogonal to the resultant obtained so far, the distance will increase from R to $\sqrt{R^2 + n^2} \approx R + n/(2R)$, with probability 1 (conditional on the orthogonality).

Figure 5.5 compares significance probabilities for B_1 with those of the uniform

random flight that represents the null hypothesis of the Moore-Rayleigh test, for $N = 10$ sample points. There exists a value of the test statistic where the two curves cross (at about $p = 0.20$), and after which the distribution function (significance probability) of the one-dimensional random walk B_1 lies below (above) the one for the uniform random flight.

The two-sample Moore-Rayleigh test, interpreted as a goodness-of-fit test, is therefore liberal, which has escaped Moore (1980) and casts doubt on the applicability of the test in this setting. The optimal upper bound for a conservative significance probability would be

$$G_N^*(r) = \sup_{\Psi_N} \mathrm{pr}(|S_N| \geq r), \tag{5.20}$$

where the supremum is taken over the set Ψ_N of all possible symmetric probability distributions for N increments. More precisely, these increments are not independent but arise from a mixture of independent distributions by the order distribution (due to the ranking of vector lengths) of their radial projections. Even if one restricts this to the class where only independent, not necessarily identical symmetric probability distributions for each step are considered, this is a difficult problem. First steps in this direction have been made by Kingman (1963), where the three-dimensional problem is reduced to a similar problem in one dimension by the familiar tangent-normal decomposition of the sphere. Apart from that, there has not been much progress in determining the envelope in Eq. 5.20. Even in the one-dimensional case it is not clear what the "fastest" random flight with linearly bounded increments is.

If a liberal test is admissible for the specific problem at hand, e.g., in exploratory data analysis, MR3 offers an efficient two-sample test. Moreover, the Remarks and Figure 5.2 suggest that the significance probabilities are only liberal for relatively large values of the test statistic. Studies with synthetic data seem to confirm that the MR3 test fails gracefully, if at all, for distributions expected in biomedical imaging practice (Scheenstra et al., 2009).

Since the assumed null hypothesis is stronger than mere symmetry, MR3 can also be used for negative testing, i.e., if the null hypothesis of the uniform random flight cannot be rejected for a sample of difference vectors, then the modified null hypothesis that $g * (f^-)$ is symmetric, not necessarily spherically symmetric, cannot be rejected. For the null hypothesis of mere symmetry, there does not exist an accessible sufficient statistic and existing tests are either only asymptotically nonparametric or require further randomization of the underlying distribution (Aki, 1987; Jupp, 1987; Diks and Tong, 1999; Henze et al., 2003; Fernández et al., 2008; Ngatchou-Wandji, 2009), so the MR3 test offers a simpler and much more efficient alternative, albeit with the disadvantage that it is potentially liberal.

5.4.2 Further issues

For a truly conservative test it is possible to adjust the p-values of the MR3 test by bootstrapping the distribution of p as in Davison and Hinkley (1997). In practice this makes use of the exchangeability of the vectors from X and Y, assuming that they both arise from the same distribution. For each pair $Y_i - X_i$ we can therefore introduce a random sign $\epsilon_i \in \{-1, +1\}$. The fraction of the test statistics $R_N^*(\epsilon)$ under all 2^N possibilities of the signs $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ that result in a larger value than the one for the trivial signs (all positive) results in an exact p-value; confer Diks and Tong (1999) for elaboration and also Lehmann and Romano (2005) for general background on symmetries and invariance in hypothesis testing. The drawback of this adjustment is that it is not routinely feasible, as it suffers from the same computational complexity problems that affect conventional permutation tests. However, the calculation of the Moore-Rayleigh test statistic is much faster than the computation of Hotelling's T^2, and in some applications this difference might be crucial.

A different issue with the Moore-Rayleigh test arises in the (usual) case of unpaired samples. The two sample test we have presented up to now assumes paired vectors, and this approach reduces the symmetry group of the null hypothesis from the group of permutations to the much smaller group of reflection symmetry of the given pairs. The main reason here is simplicity in applications and reproducibility of the test statistic. If there is no natural pairing, it seems advisable to randomly pair samples, as e.g. Moore (1980) advocates. However, a drawback is that the test statistic then becomes a random variable, and replications of the test will result in distinct significance probabilities. This is undesirable, for example, in a clinical context. Bootstrapping the test, i.e., considering the mean of the test statistic R_N^* obtained during a large enough number of resamples from the empirical distributions, is a natural way to obtain more or less replicable significance probabilities, but on the expense of computational time. It is also not precisely known at present what the convergence properties of such an estimator are.

A different approach would be to pair samples based on a measure of optimality. This seems natural enough, but has the undesirable feature that the test might become biased, e.g., too sensitive in case the sample points are matched by the method of least-squares or the Wasserstein distance. Therefore, as a practical solution in a context where reproducibility is desired, we propose to pair samples based on their ranks, such that $X_{(i)}$ is matched with $Y_{(i)}$, $i = 1, 2, \ldots, N$ (with ties resolved arbitrarily). Under the null hypothesis, the decomposability of the common distribution of X and Y guarantees the asymptotic unbiasedness of this approach, although for finite samples a slight bias is expected.

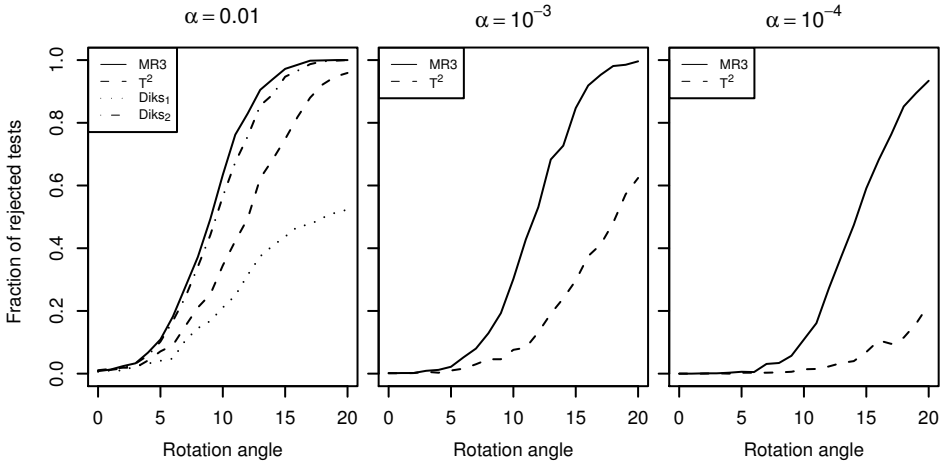


Figure 5.6: Estimated power for translated (10 standard deviations), then rotated diagonal Gaussians with unit variances as a function of relative rotation angle. Sample size $N = 10$.

5.5 Simulation results

In this section we show the results of a number of numerical simulations for the two-sample problem and compare them with Hotelling's T^2 test and the Diks1 and Diks2 tests. Throughout, we use random matching of samples and $N = 10$.

Figure 5.6 shows the results for two standard Gaussian distributions that were first translated in the same direction by ten standard deviations, and then one of them was rotated against the other (with the origin as the center of rotation), for 1000 realizations. The Moore-Rayleigh test performs well: Its power for the trivial rotation is nominal, and for larger rotation angles higher than the power of the T^2 test. Similar results are obtained when rotating Fisher distributions (not shown). Note that the Diks1/Diks2 tests are performed on the group of symmetric sign changes (of order 2^{10}), in contrast to the previous section where the full symmetry group of all rotations (of infinite order) was used, and do not resolve significance probabilities smaller than $1/1000$, i.e., their power is zero for the lower significance levels, and therefore not indicated.

Figure 5.7 compares the Gaussian distribution with the distribution $R \cdot F3_\lambda$, $R \sim \mathcal{N}(0, 1)$, when both distributions are first translated and then rotated against each other, with similar results.

Finally, Figure 5.8 shows adjusted p-values, for 10^4 permutations and 100 realizations each. The Moore-Rayleigh test again shows slightly better power then the T^2 test. More importantly, there is not much difference with the unadjusted power

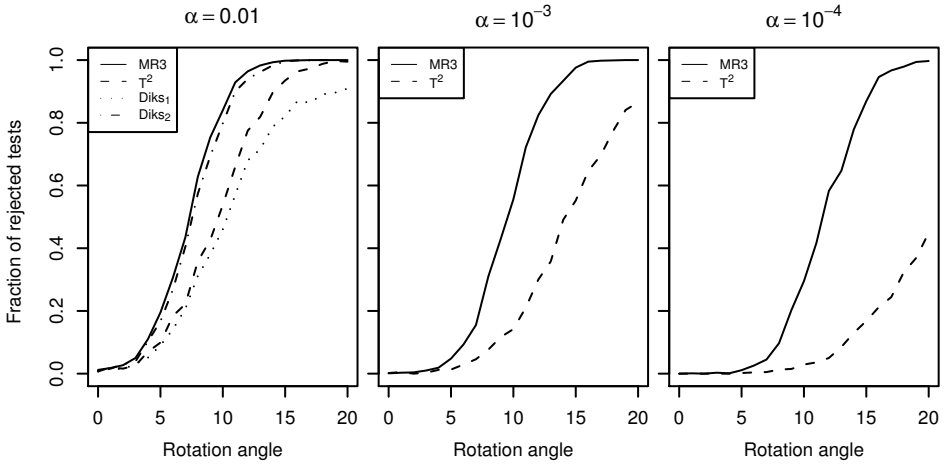


Figure 5.7: Estimated power functions for translated (10 standard deviations), then rotated diagonal Gaussians with unit variance against the scattered Fisher distribution (scaled by a unit Gaussian) as a function of relative rotation angle. Sample size $N = 10$. Note the small power at angle zero.

functions (Figures 5.6 and 5.7). These results are based on 100 realizations only, to speed up the considerable amount of computations, which accounts for the visible fluctuations.

5.6 Application: deformation-based morphometry

As remarked in the introduction, an important field of application of the Moore-Rayleigh test is the morphometric analysis of MR images.

5.6.1 Synthetic data

The Moore-Rayleigh test was validated on a synthethic $50\times50\times80$ three-dimensional image domain. Five spherical deformations were added in two distinct regions, introducing characteristic localized changes. The volume in each sphere was mapped radially, linearly expanding by a factor $\lambda_1 = 1.8$ from the centerpoint to half radius distance, and then contracting linearly by $\lambda_2 = 2 - \lambda_1$, resulting in a one-to-one transformation of each spherical volume. Although the transformations were not smooth, interpolation at subpixel level guaranteed that they were indeed local diffeomorphisms. Figure 5.9 shows the transformation along a radial direction.

A Gaussian noise process (zero mean, SD = 1.0) was added to the deformed im-

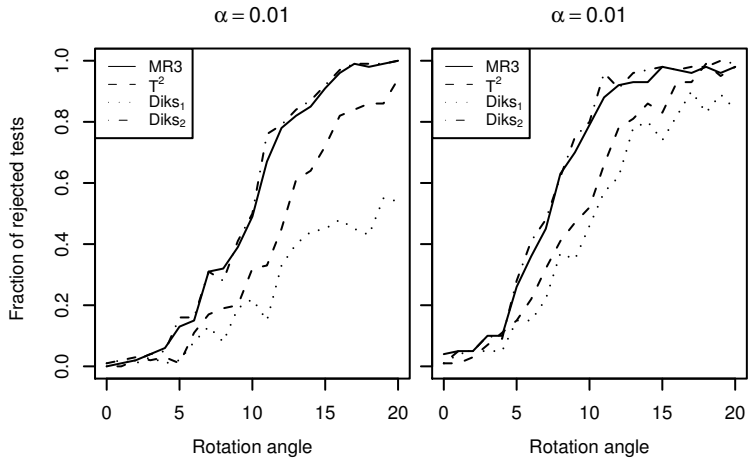


Figure 5.8: Adjusted estimated power functions. Left: translated (10 standard deviations), then rotated diagonal Gaussians with unit variances as a function of relative rotation angle. Right: translated (10 standard deviations) then rotated diagonal Gaussians with unit variance against the scattered Fisher distribuion (scaled by a unit Gaussian) as a function of relative rotation angle. Sample size $N = 10$. Results based on 100 realizations of 10^4 permutations each.

age, for a total of 15 distinct realizations, thereby simulating natural variation in brain structure. Panel A in Figure 5.10 shows the average lengths of deformation vectors in a cental slice of this image set. Two spherical deformations in the lower part (Region I) with radii 6 voxels (S_4, left) and 9 voxels (S_5, right) were created at a distance of 25 voxels. In the upper part (Region II) one sphere of radius 9 voxels (S_2) and two spheres of radius 6 voxels (S_1 and S_3) were created at successive distances of 12.5 voxels between their center points, creating a more complex deformation due to partial overlap in the superposition of deformation fields.

A second group of 15 images was created, with a reduced radius of 6 voxels for the spheres S_2 and S_5. Panel B in Figure 5.10 depicts the absolute differences in deformation vector lengths between the average deformation fields of both groups in the central slice.

For the evaluation of the statistical tests, ground truth, i.e., voxels for which the null hypothesis of no group difference should be rejected, was taken to be the total volume of the two spheres S_2 and S_5. This approximation allowed the estimation of *precision* and *recall* from the numbers of true positives (TP), false positives (FP, type I

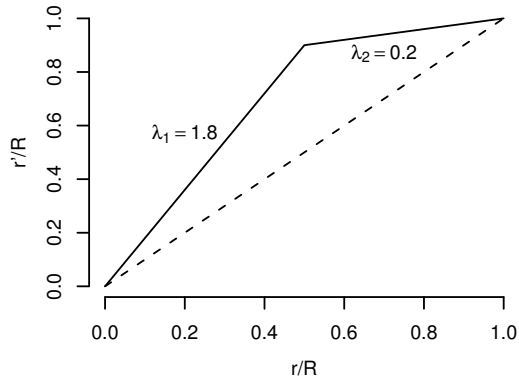


Figure 5.9: Generation of spherical deformations. The plot shows the behaviour of the deformation field in a one-dimensional projection along a radial direction. The volume at normalized distance r/R from the centerpoint of the sphere (radius R) is mapped radially to r'/R. For $r < R/2$ the length of the deformation vectors expands linearly, $r' = \lambda_1 r$, attaining its maximum at half radius $r = R/2$. For $r > R/2$ the length shrinks linearly by $\lambda_2 = 2 - \lambda_1$, ensuring continuity at the boundary. The stippled line shows the case of no deformation ($\lambda_1 = 1$).

Table 5.2: Precision and recall for synthethic dataset

Test	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$		$\alpha = 2.5 \cdot 10^{-7}$	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
MR3	0.07	0.81	0.21	0.63	0.59	0.39	1	0.04
HT2	0.07	0.77	0.21	0.54	0.56	0.28	1	0.01
permuted HT2	0.07	0.77	0.21	0.54	0.57	0.28	0	0
Diks1	0.03	0.38	0.1	0.23	0.35	0.11	0.69	0.04
Diks2	0.07	0.80	0.22	0.59	0.56	0.31	0.77	0.08

error) and false negatives (FN, type II error), where

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The results are shown in Table 5.2 for different significance levels α. The rightmost level $\alpha = 2.5 \cdot 10^{-7}$ corresponds to 0.05 under Bonferroni correction with 200 000 voxels. The performance of all four tests is comparable, with the Moore-Rayleigh test exhibiting better recall and precision rates then the other tests. Note that the results of the permutation version of Hotellings T^2 test are limited by the number of relabelling ($N = 10\,000$), such that Bonferroni correction for multiple comparisons did not result in any significant voxels.

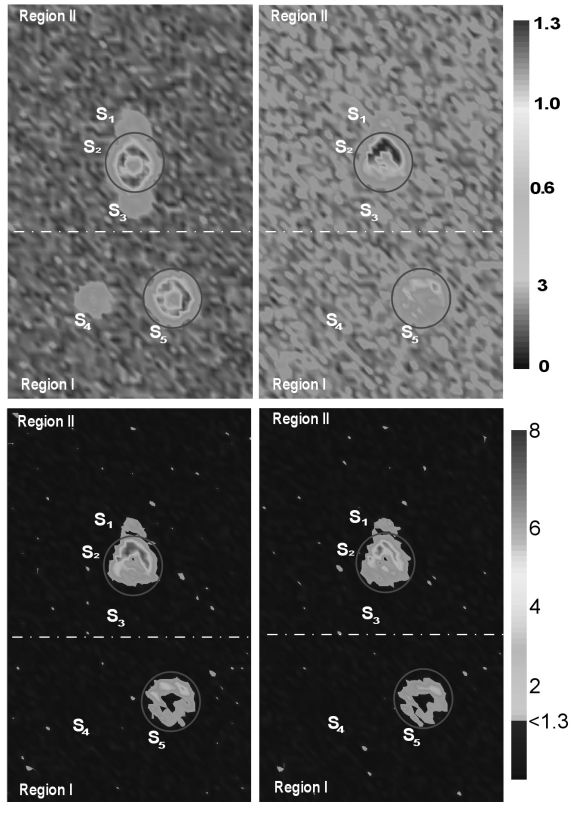


Figure 5.10: Validation with a synthetic dataset. Upper part: Central slice from the deformation field of $50 \times 50 \times 80$ voxels (isotropic spacing of 1.00 mm), showing the five spherical deformations that were added to it (see text for details). The color indicates the length of the deformation vectors (in units of voxel dimensions). A: Mean deformation field for the first group. B: Difference between deformation fields for the two groups (smaller deformations in spheres S_2 and S_5 in the second group). Lower part: Negative logarithms of significance probabilities for the statistical tests. C: Moore-Rayleigh test. D: Hotellings T^2 test.

5.6.2 Experimental data

To demonstrate the Moore-Rayleigh test in a clinical setting, MR images of five mice with enlarged ventricles (Panel A in Figure 5.11) were compared with images obtained from a group of five control animals (Panel B). The affected mice were selected from a large set of T2-weighted MR scans used for general mouse phenotyping by a trained observer, which included transgenically altered mice. The inclusion criterion was the existence of visibly enlarged ventricular spaces. This dataset exhibits

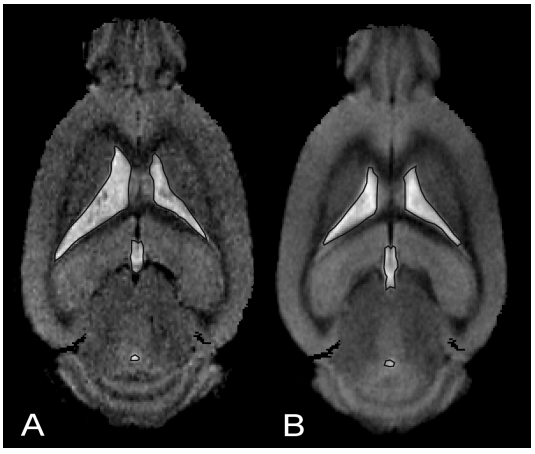


Figure 5.11: Deformation-field analysis of mouse brains. A: Slice of a MR image of a mouse with enlarged ventricles (here: especially the left ventricle). B: The same slice from the average MR image of the control mice with normal ventricles. The ventricles (main regions of interest) are manually delineated.

characteristic properties of clinical data and was selected on the following grounds:

- The pathology of the diseased mice is clearly visible and allows to validate the results.
- Relatively large levels of noise occur.
- Small sample size, since in prospective studies a typical dataset of mice consists of 5–10 animals per group.

All MR scans were normalized for global orientation, size and shape (by an affine transformation) and resampled to the same coordinate space with equal dimensions ($160 \times 132 \times 255$ voxels) and isotropic voxel size (0.06 mm), thereby allowing voxel-wise comparison between different scans. Nonlinear registration was then performed to obtain deformation fields, utilizing the symmetric demons algorithm (Thirion, 1998), as implemented in the Insight Toolkit (Yoo, 2004). After normalization, the images of the control group were registered to the corresponding group average under a leave-one-out design. Thereby, to reduce correlations due to the small sample size, each image was registered to the average obtained from the remaining images of the group. The images of the mice with enlarged ventricles were registered to the average image of all controls (Figure 5.11, Panel B). Spherical symmetry of voxel-wise deformation fields (due to intra-group variation) should then hold for the controls, and under the null hypothesis of no group-wise difference also for the mice with enlarged ventricles.

Negative logarithms of significance probabilities are shown as statistical parametric mappings in Figure 5.12, overlaid on the average image of the normalized

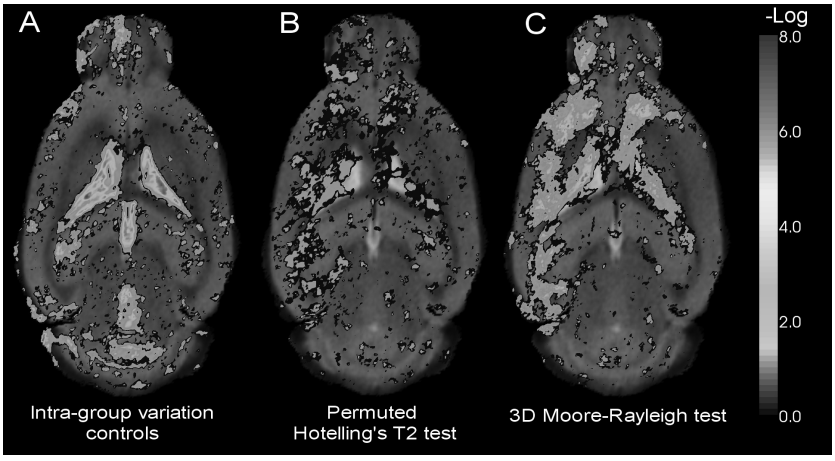


Figure 5.12: Average MR image of control mice overlaid with significance probabilities obtained by statistical tests. A: The one-sample Moore-Rayleigh test indicates the loss of spherical symmetry at various places in the control group. B: Hotellings T^2 two-sample test ($N = 10\,000$ relabellings). C: The two-sample 3D Moore-Rayleigh test. In all images negative logarithms of significance probabilities are shown for better visualization, and only significant ($p < 0.05$) voxels are colored.

control brains. Only significant ($p < 0.05$) voxels are indicated. Note that only one central slice (2D image) is shown, although the registration is performed in 3D. Furthermore, only voxels inside the brain were analyzed, resulting in about 1.9 million voxels in total.

Compared with Hotelling's T^2 permutation test (Figure 5.12, middle frame), the two-sample Moore-Rayleigh test exhibits lower p-values (Figure 5.12, right frame) of which a few hundred remain significant even under Bonferroni correction for multiple testing (p-value lower than 10^{-6}). Note that the T^2 test does not show any significant voxels after Bonferroni correction, so it cannot be reliably decided which (if any) voxels exhibit structural changes.

Regions where significant voxels were found correspond well between the two tests and conform with established knowledge of the pathology of enlarged ventricles. Differences in and around the ventricles (delineated in Figure 5.11) are of course expected. As the enlarged ventricles cause the surrounding brain tissue to shift location, they thereby induce secondary deformations (shrinkage to account for the expansion of the ventricles), which also seem to have been highlighted well by the tests. In particular, both the MR3 and the T^2 test display more significant voxels in the left side of the brain, corresponding to the fact that the left ventricles were slightly larger than the right ventricles in the group with enlarged ventricles.

However, the distribution of deformations was not spherically symmetric in all voxels of the control group (Figure 5.11, left frame), as assessed by the one-sample MR3 test. This indicates systematic variations that possibly arise from nonnormal differences in cerebrospinal fluid content in control mice. In these voxels, the two-sample MR3 test could be potentially liberal, and further tests should be considered.

In fact, both tests also detect significant spurious differences at other places of the brain, some of which probably need to be considered artifacts of image (mis-) registration, due to varying brain shapes between individual mice and the small sample size. Since the null hypothesis was not valid in parts of the brains of the control group, the test results in these regions also have to be considered with care. This shows the importance, but also the difficulties, of proper validation in a clinical context. As always in hypothesis testing, results should be carefully interpreted, not only by statistical significance, but also guided by neuroanatomical insight. Here, if spurious voxels are excluded on a priori grounds, the Moore-Rayleigh test detects significant deformations in brain structure with strong control of the family-wise error rate. Voxels whose null hypothesis has been rejected could then be subjected to further statistical tests, analyzed with regard to what kind of structural change is the most probable cause of these deviations, or lead to further experimental procedures, e.g., targeted biopsies. The nonnormal variation in brain regions of the control mice is also potentially interesting, since this contrasts with widely held assumptions.

5.7 Discussion

It is possible to test spherical symmetry in three dimensions with high numerical accuracy by using the combinatorial sum representation given in Eq. (5.15). In combination with Kahan summation (Goldberg, 1991), this representation makes it feasible to routinely calculate p-values for finite sample sizes that allow to assess statistical significance. Even for hundreds of thousands of multiple comparisons with a Bonferroni correction, as is common practice in neuroscientific imaging applications, the proposed approach is effective. Permutation methods, although theoretically preferred, are difficult to use in this setting due to practical limitations. The standard approaches to cope with these limitations, based on either saddle-point approximations to permutation tests (Robinson, 1982) or on permutation tests for linear test statistics, where the conditional characteristic function can be rewritten as a convergent approximating series (Gill, 2007), are not directly applicable because these statistics usually do not arise in these practical problems or are too involved in the multivariate case. An alternative might be the use of optimal (Bayesian) stopping rules in the resampling process (Besag and Clifford, 1991; Fay et al., 2007). However, small sample sizes can still seriously restrict the possible range of the significance probabilities.

In the special case of the two-sample problem, the distribution of the null hy-

pothesis is conditional on the unknown distribution of the data, and the generalized Moore-Rayleigh test is only approximately valid, a feature that all other (non-randomized) tests of symmetry exhibit. In Section 5.6 we evaluated the properties of this generalized Moore-Rayleigh test empirically with simulated imaging data of known ground-truth and by comparison with other nonparametric tests; for a different comparative study see Scheenstra et al. (2009). Even though the test is theoretically liberal, it seems to work well in practice, as it is not particulary sensitive to the difference between symmetry and spherical symmetry. An exact test is furthermore available by the permutation variant of the Moore-Rayleigh test, with slightly improved power when compared with conventional permutation testing. This can be used in a second stage after initial screening with the fast, unadjusted Moore-Rayleigh test. Although such screening could also be realized by the T^2 test, the MR3 test seems better suited to this problem due to its enhanced power, which allows for strong control of the family-wise error. In contrast, the T^2 test does often not allow the localization of individual voxels, as demonstrated in the example on deformation morphometry in brain scans. It should be noted that we have only considered the conservative Bonferroni correction here, for simplicity, but it is expected that the MR3 test remains a more sensitive instrument also under modern step-down multiple comparison procedures (as described in, e.g., Nichols and Holmes (2007)).

Chapter 6

Electrophysiology of the brain

Abstract

The analysis of functional and effective brain connectivity forms an important tool for unraveling structure–function relationships from neurophysiological data. It has clinical applications, supports the formulation of hypotheses regarding the role and localization of functional processes, and is often an initial step in modeling. However, only a few of the commonly applied connectivity measures respect metric properties: reflexivity, symmetry, and the triangle inequality. This may hamper interpretation of findings and subsequent analysis. Connectivity indices obtained by metric measures can be seen as functional distances, and may be represented in Euclidean space by the methods of multidimensional scaling. We sketch some classes of measures that do allow for such a reconstruction, in particular the class of Wasserstein distances, and discuss their merits for interpreting cortical activity assessed by magnetoencephalography. In an application to magnetoencephalographic recordings during the execution of a bimanual task, the Wasserstein distances between relative circular variances indicated cortico-muscular synchrony as well as cross-talk between bilateral primary motor areas in the β-band.

6.1 Introduction

Functional connectivity can be defined as the occurrence of a significant statistical interdependency between activities of distant neurons or neural populations. In combination with the constraining anatomy, this definition forms a proper starting point for unraveling the relationship between structural and functional features of the brain. Down to the present day, the quest for a comprehensive understanding of structure-function interaction has attracted a lot of attention (Stephan et al., 2008; Lee et al., 2003). Structure can be rather complicated but is typically considered material and fixed, whereas function reflects statistical similarity between dynamical processes in the brain. Related concepts are anatomical and effective connectivity, respectively, where the latter refers to causal relationships between signals (Friston et al., 1993; Ramnani et al., 2004). A functional connectivity analysis often precedes the formulation of a causal (or directed) model, yielding numerous applications. Apart from its fundamental role in determining functionally important neuronal processes, it has important clinical applications (Stam, 2005). The neurophysiological underpinnings, however, are still under debate, partly because of the huge

variety of connectivity measures employed, rendering methods inscrutable (Pereda et al., 2005) and questioning the possible contribution of functional connectivity to an integrative understanding of brain functioning (Horwitz, 2003; Fingelkurts et al., 2005). To dispel doubts, we outline general properties of functional connectivity measures and their implications for analysis. We aim for facilitating the selection of proper measures and, by this, distill convincing arguments for their relevance for an understanding of brain dynamics.

In a nutshell, the large majority of commonly implemented connectivity measures do not respect fundamental metric properties. Here, we focus on three important properties: (*i*) reflexivity, (*ii*) symmetry, and (*iii*) the triangle inequality. If a connectivity measure disregards one or more of these three properties, its interpretation may be ambivalent when multiple signals are assessed. Put differently, such measures can be very successful for a pair-wise comparison of signals, i.e., in the bivariate case, but they may lead to spurious results in multivariate settings (Kus et al., 2004). On the contrary, if a connectivity measure does respect all properties (*i*)-(*iii*), then it describes a proper *functional distance*. We argue that this is a necessary condition for a truly integrative analysis. Of course, genuine multivariate statistical methods suffice for this purpose, but implementation can be cumbersome and results might be difficult to interpret. More important, commonly used multivariate methods require explicit assumptions about the data, e.g., signals ought to be normally distributed to apply principal or independent component analysis, cluster analysis typically requires an educated guess regarding numbers of cluster, and so on.

An intermediate form of analysis is the transformation of proper functional distances into a low-dimensional representation as points in a Euclidean space. This technique, commonly referred to as *multidimensional scaling* (MDS), was successfully applied in anatomical studies (Young et al., 1995; Goodhill et al., 1995) as well as in the context of functional connectivity (Friston et al., 1996). In general, MDS allows for visualizing signals in a 'functional space', which may facilitate hypothesis-finding regarding relevant interactions. The MDS representation can also be used for classification and discrimination, which is particularly interesting from a more clinical perspective. However, the necessary statistical methodology for proper classification has been developed only recently (Anderson and Robinson, 2003; Trosset and Priebe, 2008). MDS requires functional distances to work with, that is, all the metric properties (*i*)-(*iii*) need to be respected. As we will show below, the number of connectivity measures forming proper functional distances is far and few between. Therefore, we additionally discuss the so-called Wasserstein distances (Villani, 2003), which are general distances between probability distributions: total variation, i.e., the area between two probability densities, is a common example. We illustrate the application of Wasserstein distances using source-localized magnetoencephalographic (MEG) recordings obtained during bimanual isometric force production.

6.2 Distance properties

There exists a plethora of connectivity measures to assess statistical similarities of dynamical processes in the brain (Quiroga et al., 2002; Pereda et al., 2005). Below we list several explicit examples of commonly used methods. To anticipate the subsequent discussion, Table 6.1 provides an overview of commonly applied connectivity measures and their metric properties. All methods are (usually) bivariate, i.e., offer a pair-wise link between signals, and they result in a single scalar number, which is interpreted as either similarity or dissimilarity. The pair-wise 'distances' of a set of N signals or channels are combined in a single $N \times N$ connectivity matrix

$$\mathbf{\Delta} = \{\Delta_{ij}\}_{1 \leq i \leq N, 1 \leq j \leq N} \,. \tag{6.1}$$

That is, the elements Δ_{ij} are functional connectivities that stem from a fixed, scalar-valued connectivity measure. In the following we assume that the Δ_{ij}'s are dissimilarities, such that small values of Δ_{ij} are interpreted as a functional similarity between the i-th and j-th signal[1].

6.2.1 Metric properties

As mentioned in the *Introduction*, a connectivity measure has to be reflexive, symmetric, and it has to fulfill the triangle inequality in order to represent functional distances.

(*i*) If the diagonal elements of the connectivity matrix $\mathbf{\Delta}$ vanish, then its underlying measure is *reflexive*. That is, reflexivity is the property that

$$\Delta_{ii} = 0 \tag{6.2}$$

holds for all signals $i = 1, \ldots, N$. This is often a trivial property that holds for most connectivity measures by construction, or can be obtained by some simple transformation; see Table 6.1 for an overview. If furthermore $\Delta_{ij} > 0$ for all signals $i \neq j$, then the measure is *strongly reflexive*. Although strictly speaking strong reflexivity is necessary for a metric, the property of reflexivity is enough for most applications. Technically, reflexive connectivity measures lead to a pseudo-metric, i.e., certain kinds of interactions might not be distinguished by them, but in the following we do not emphasize this distinction.

(*ii*) A connectivity matrix Δ is *symmetric* if

$$\Delta_{ij} = \Delta_{ji} \tag{6.3}$$

[1] Measures that represent similarities, e.g. correlation coefficients, can be turned into dissimilarities by various transformations.

holds for all pairs (i, j). In fact, symmetry is often unwanted because it does not allow for assessing the direction of connectivity, e.g., the flow of information, or the directionality of dynamic coupling. Instead, symmetric measures determine the commonality of two signals, a necessary feature to provide a unique distance. It is important to note that the analysis of symmetric connectivity measures has an analogue for asymmetric measures. To explain this, first note that a general asymmetric connectivity matrix $\boldsymbol{\Delta}$ can be uniquely decomposed as

$$\boldsymbol{\Delta} = \boldsymbol{S} + \boldsymbol{A}, \tag{6.4}$$

where $\boldsymbol{S}$ is symmetric and $\boldsymbol{A}$ is anti-symmetric, i.e., $S_{ij} = S_{ji}$ and $A_{ij} = -A_{ji}$, respectively. Moreover, the sum-of-squares of $\boldsymbol{\Delta}$ decomposes as

$$\sum_{ij} \Delta_{ij}^2 = \sum_{ij} S_{ij}^2 + \sum_{ij} A_{ij}^2. \tag{6.5}$$

since the cross-product term $\sum_{ij} S_{ij} A_{ij}$ vanishes, and the trace of $\boldsymbol{SA}$ is zero, due to the fact that $\boldsymbol{S}$ and $\boldsymbol{A}$ are orthogonal. Hence, the analysis of an asymmetric connectivity matrix can be split into the analysis of its symmetric part and a slightly modified analysis for the anti-symmetric part; see Section 6.2.2 for more details.

(*iii*) The *triangle inequality* is the property that

$$\Delta_{ij} \leq \Delta_{ik} + \Delta_{kj} \tag{6.6}$$

holds for all triples (i, j, k). It formalizes the well-known geometric notion that, given a (shortest) distance between two points, there can be no further 'short-cuts'. In the current context the triangle inequality tells us whether a given connectivity measure reflects genuine information about the commonality of two signals or not. Put differently, violations of the triangle inequality are methodologically very important as they indicate that the signals might be a mixture of two or more distinct processes or subsystems[2]; see Figure 6.1.

The applicability of the triangle inequality may depend on the scale in which connectivity is measured. For example, if the maximal violation of the triangle inequality in a given connectivity matrix Δ is $\varepsilon > 0$, then adding ε to all dissimilarities, i.e., $\Delta_{ij} \mapsto \Delta_{ij} + \varepsilon$, trivially restores the triangle inequality. This process can be interpreted as 'flattening' of the data, since smaller dissimilarities are affected relatively more than larger ones. It would be good practice if connectivity studies would consider the triangle inequality and, by the same token, publish numerical values of its most extreme violation. Likewise, if the triangle inequality holds, then one should ask which maximal ε can be subtracted from the off-diagonal elements of the con-

[2] Violations of the triangle inequality may also imply that the resolution of recording is limited, e.g., the number of signals used is too small, as they may be caused by the existence of opaque subsystems.

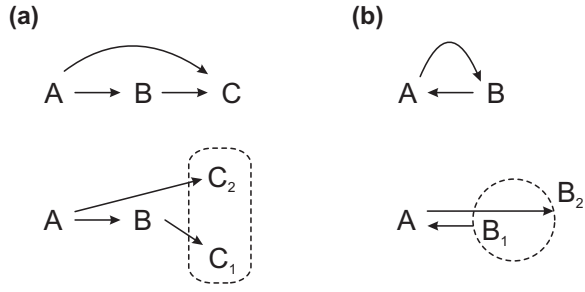


Figure 6.1: Violations of metric properties. a) A violation of the triangle inequality (top) is an indication that one measures functional properties between two distinct subsystems (bottom). b) An asymmetry (top) may also indicate the existence of two subsystems with distinct functional roles (bottom). It can be conveniently visualized by the radius-distance (see text).

nectivity matrix before the inequality fails, as this may provide insight into the robustness of estimates.

Connectivity measure	S	R	T	References
Covariance & Correlation	+	0	0	
Mutual information	+	+	0	(Gray, 1990; Cover and Thomas, 1991) (Kraskov et al., 2004)
Granger causality	-	-	-	(Granger, 1969; Geweke, 1982)
Coherence	+	0	-	(Brillinger, 1981)
Imaginary part of coherency	-	-	-	(Nolte et al., 2004, 2008)
Relative circular variance	+	+	-	(Lachaux et al., 1999; Mormann et al., 2000)
Synchronization likelihood	+	+	-	(Stam and van Dijk, 2002)
Wasserstein distance	+	+	+	(Moeckel and Murray, 1997) (Muskulus and Verduyn-Lunel, 2008b)

Table 6.1: Overview of commonly used measures in connectivity analysis. S: Symmetry, R: Reflexivity, T: Triangle inequality; '+' indicates the measure respects the given property, '-' indicates the opposite, '0' indicates that the measure does not respect the property, but that extensions and/or derived versions of it exist that do.

6.2.2 Embeddability and MDS

The strict metric properties (*i*)-(*iii*) are not the only important properties of connectivity measures. For instance, there exists a hierarchy of generalized triangle inequalities, which are usually expressed in terms of the so-called Cayley-Menger determinants (Blumenthal, 1953). The hierarchy contains inequalities for quadruples, quintuples, etc., of points, which need to be fulfilled if they are to lie in a Euclidean

space of given dimension.

Let $M \leq N$ be the dimension of a Euclidean space in which the recorded signals can be represented as points such that their Euclidean distance is equal to the dissimilarities Δ_{ij}. We ask how this M-dimensional space can be determined. We consider a $N \times N$ connectivity matrix $\boldsymbol{\Delta}$ that fulfills (*i*)-(*iii*). In addition, let $\boldsymbol{D}^2$ be the matrix of squared dissimilarities such that $D_{ij}^2 = (\Delta_{ij})^2$ holds. In general, $\boldsymbol{D}^2$ can be expanded as

$$D_{ij}^2 = \sum_k \left(\xi_{ik}^2 + \xi_{jk}^2 - 2\xi_{ik}\xi_{jk}\right), \tag{6.7}$$

where $\xi_{i1}, \xi_{i2}, \ldots, \xi_{iN}$ denote coordinates of the i-th 'signal' embedded into a yet unknown, N-dimensional space that will reduce to M-dimensions. Eq. (6.7) can be re-written in terms of

$$\boldsymbol{D}^2 = \boldsymbol{\nu}\mathbf{1}_N + \mathbf{1}_N\boldsymbol{\nu}^T - 2\boldsymbol{\xi}\boldsymbol{\xi}^T, \tag{6.8}$$

in which $\boldsymbol{\xi}$ is the matrix with elements $\xi_{1\ldots N,1\ldots N}$, the vector $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_N)^T$ consists of the squared norms of $\boldsymbol{\xi}_i$, i.e., $\nu_i = ||\boldsymbol{\xi}_i||^2 = \sum_k \xi_{ik}^2$, and $\mathbf{1}_N$ is an $N \times 1$ vector of ones; superscript T denotes the matrix transpose. Inverting this identity yields the matrix $\boldsymbol{\xi}\boldsymbol{\xi}^T$ of scalar products (Gram matrix) in terms of[3]

$$\boldsymbol{\xi}\boldsymbol{\xi}^T = -\frac{1}{2}\left(\boldsymbol{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)\boldsymbol{D}^2\left(\boldsymbol{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right). \tag{6.9}$$

This matrix is positive semi-definite, i.e., all its eigenvalues are nonnegative. In particular, the first M eigenvalues of $\boldsymbol{\xi}\boldsymbol{\xi}^T$ are positive, and if, as assumed, $\boldsymbol{\Delta}$ represents distances between N recorded signals in an $(M \leq N)$-dimensional space, all remaining eigenvalues vanish by the spectral theorem. Inversely, if all eigenvalues of $\boldsymbol{\xi}\boldsymbol{\xi}^T$ are positive, then the dissimilarities in $\boldsymbol{\Delta}$ can be identified with distances of points $\boldsymbol{\xi} \in \mathbb{R}^N$ (Havel et al., 1983, Th. 3.1). The coordinates of these points can be readily obtained via singular value decomposition:

$$\boldsymbol{\xi}\boldsymbol{\xi}^T = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T = \left(\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\right)\left(\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\right)^T. \tag{6.10}$$

We sort the eigenvalues $\boldsymbol{\Lambda}$ in descending order and combine the first M columns of the matrix of eigenvectors $\boldsymbol{Q}$ into $\boldsymbol{Q}_{1,\ldots,M}$, spanning a space $\mathbb{R}^M$. Then we have

$$\boldsymbol{\xi} = \boldsymbol{Q}_{1,\ldots,M}\boldsymbol{\Lambda}_{1,\ldots,M}^{1/2}, \tag{6.11}$$

which is the M-dimensional classical (or metric) MDS representation of the N signals from the connectivity matrix $\boldsymbol{\Delta}$. The space $\mathbb{R}^M$, in which the signals are represented as points, has been termed *functional space* by Friston et al. (1996).

[3] This operation is typically referred to as double centering and implies that, whereas distances are invariant under translations, scalar products are not.

Interestingly, this representation is equivalent to a principal component analysis (PCA) of the scalar product matrix $\boldsymbol{\xi}\boldsymbol{\xi}^T$; we note that every real-valued, symmetric matrix that decomposes into such a scalar product is positive semi-definite. In particular, if the connectivity matrix is already of this form and has a positive trace (i.e., is *not* reflexive) as does the similarity measure covariance, then one can apply PCA directly onto $\boldsymbol{\Delta}$ and the functional space $\mathbb{R}^M$ is given as a linear transformation of the original signal space. The general similarity with PCA implies that MDS solutions are nested: if the embedding dimension M is increased to $\tilde{M} > M$, then the first M coordinates of these points in $\mathbb{R}^{\tilde{M}}$ are identical to the M-dimensional reconstruction. On this account, Gower (1966) proposed the term principal coordinate analysis for the MDS transformation, nowadays this is more commonly called metric MDS; for a discussion of non-metric variants of MDS see (Borg and Groenen, 2005). We note that there is indeed a subtle difference between PCA and metric MDS: the double centering operation in the latter usually results in the removal of the first (baseline) principal component (Heiser and Meulman, 1983).

An important advantage of functional space is the possibility of a discriminant analysis of signals. For the connectivity matrix $\boldsymbol{\Delta}$ this is not recommended because of the collinearity problem (Næs and Mevik, 2000), although this may explain the efficiency of PCA in many applications. In particular, covariance matrices should be subject to linear discriminant analysis with great care. Also, cross-validation has been a particular problem because the MDS reconstruction depends on all $\frac{1}{2}N(N+1)$ dissimilarities in $\boldsymbol{\Delta}$. While this does provide for the robustness of the method, for cross-validation it forms a challenge since one needs to compare a single signal with the remaining $N-1$ signals in functional space that is calculated only from their $\frac{1}{2}(N-1)N$ dissimilarities. A recently developed iterative algorithm allows to find the prospective coordinates of the single signal in this space by minimization of an error criterion (Trosset and Priebe, 2008). A major drawback of metric MDS, at least for some applications, is the necessity of obtaining all $N \times N$ connectivity values with equal quality. With respect to this, Spence and Domoney (1974) have shown that in the case of sufficiently small noise levels, even in the absence of 80% of (randomly selected) entries of $\boldsymbol{\Delta}$ the MDS reconstruction is almost identical to the reconstruction that involves all mutual dissimilarities. The important case of noisy signals, however, remains an area of active research; at present, the approach of Singer (2008) appears quite promising.

When the connectivity matrix is asymmetric, i.e., $\boldsymbol{\Delta} = \boldsymbol{S} + \boldsymbol{A}$ with $\boldsymbol{A} \neq 0$, metric MDS needs to be modified to handle the anti-symmetric part $\boldsymbol{A}$. From the currently available methods (Borg and Groenen, 2005), we mention two: in the Gower model, a special form of singular value decomposition is applied, that is,

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{Q}^T. \tag{6.12}$$

The singular values in $\boldsymbol{\Lambda}$ arise in pairs, and $\boldsymbol{R}$ is a permutation-reflection matrix

with diagonal anti-symmetric 2×2 blocks containing 1 and -1 off-diagonal element (Constantine and Gower, 1978). A second, more accessible representation is obtained in the radius-distance model of Okada and Imaizumi (1987), which visualizes both the symmetric and anti-symmetric part of $\boldsymbol{\Delta}$. In brief, each signal is represented as a circle. The symmetric part $\boldsymbol{S}$ is given by the centers of the circles, and the anti-symmetric part $\boldsymbol{A}$ by the radius distance between their centers: for two signals i and j, this distance is computed by subtracting the starting radius from the distance of the respective center points and adding the ending radius; see Figure 6.1 for an example.

6.2.3 Graph-theoretic analysis

Recently, the graph-theoretic analysis of connectivity matrices has attracted a lot of attention (Bullmore and Sporns, 2009). It has revealed interesting properties of large-scale cortico-cortical connectivity and has important applications in a clinical setting, many of which are reviewed by Stam and Reijneveld (2007). Problems with graph-theoretic methods are discussed by Ioannides (2007), who also proposed a nested analysis to integrate functional connectivity obtained for a variety of distinct tasks and conditions. Common practice is to threshold the connectivity matrices so that a connectivity above a certain, fixed value is set to unity and neglected otherwise. The resulting binary connectivity matrices (or adjacency matrices) naturally represent an undirected connectivity graph. Thresholding, however, may discard substantial information and to date there is no general agreement on a criterion for threshold selection. Considering weighted graphs, in which connectivities are interpreted as edge weights, is hence our preferred approach. For the resulting (weighted or undirected) graphs several statistical measures can be investigated. An example are the network participation indices of Kötter and Stephan (2003), who also employ a variant of MDS to derive a two-dimensional visualization of the networks under study. A more recent approach is the so-called motif analysis, in which the frequency of occurrence of small, induced subgraphs (i.e., motifs) is compared with their expected number in randomly generated graphs (Sporns and Kötter, 2004). With respect to phase locking, Bialonski and Lehnertz (2006) proposed to consider the eigenvectors of the phase uniformity matrix. The spectrum of the adjacency matrix was considered by da Costa and Barbosa (2004) for the analysis of anatomical connectivities; this measure characterizes the cycle structure of the connectivity graph and can be employed analogously for functional connectivity matrices.

The first eigenvector of the connectivity matrix offers an elegant interpretation presuming the matrix fulfills a few additional properties. First, the connectivity values are interpreted as transition probabilities, e.g., for the flow of information, then the system of signals can be considered a Markov chain; this implies that the connectivity matrix has to be non-negative and normalized row- or column-wise (possibly

violating the triangle inequality thereby). Second, the connectivity matrix is (almost always) a-periodic and irreducible, that is, for all pairs $(i, j))$ of signals there exists a linking sequence of indices $[i = i_1 \to i_2 \to \cdots \to i_n = j]$ with positive connectivities. In other words, the corresponding connectivity graph is strongly connected and does not have a decomposition into periodic classes. Then, the first eigenvector is the unique, invariant eigenvector of the Markov chain, that is, it has unit eigenvalue and represents a probability distribution that is stable under the transition dynamics (Brémaud, 1999). It can be interpreted as the equilibrium solution of diffusion on the connectivity graph. A famous application of such an invariant eigenvector is the PageRank algorithm which constitutes the core of the internet search engine Google (Brin and Page, 1998; Bianchini et al., 2005). More details of its implementation for large networks[4] are discussed by Langville and Meyer (2004).

The relevance of the first eigenvector originates from the fact that it describes the relative importance of a given signal (a node in the connectivity graph) in the total network. Instead of just considering local information, it is determined by the global connectivity structure. In an application to cortical signals, it represents an underlying distributed background activity or 'functional rest-state'. To our knowledge, determining this eigenvector for functional connectivity matrices has not been considered before, although da Costa and Barbosa (2004) suggest its use for anatomical connectivites; see Figure 6.4 for an example.

6.3 Connectivity measures

To test for distance properties we list several important methods that have found widespread use in encephalography; for approaches to functional magnetic resonance imaging (fMRI) we refer to Li et al. (2009). For the sake of brevity, we abstain from discussing the measures' statistical, practical, or methodological properties (David et al., 2004). We also do not discuss measures derived from model-driven analysis, i.e., structural equation modeling, dynamic causal modeling, or psychophysiological interactions.

6.3.1 Statistical measures

Covariance quantifies the linear relationship between two signals. Given two real-valued signals, x_i and x_j, it is typically defined as the expectation

$$\sigma_{ij} = \mathbb{E}\left[\Big(x_i - \mathbb{E}\big[x_i\big]\Big)\Big(x_j - \mathbb{E}\big[x_j\big]\Big)\right]. \tag{6.13}$$

[4] An implementation for MATLAB (The Mathworks, Natick) is provided through the ConTest toolbox, available from:
http://www.maths.strath.ac.uk/research/groups/numerical_analysis/contest/toolbox.

Covariance is not reflexive, in the sense that the variance of two signals should both represent zero dissimilarity. Normalizing by the individual standard deviations yields the (Pearson) correlation coefficient ρ_{ij}, which lies between -1 and $+1$. As a negative similarity appears useless, this can be transformed to a reflexive dissimilarity measure $\Delta_{ij}^{(\mathrm{corr})}$ by letting

$$\Delta_{ij}^{(\mathrm{corr})} = \sqrt{1 - \rho_{ij}^2}. \tag{6.14}$$

Alternatively, we can use the Pearson absolute dissimilarity

$$\Delta_{ij}^{(\mathrm{Pearson})} = 1 - |\rho_{ij}|. \tag{6.15}$$

Obviously, information about the direction of the connection is lost, as both covariance and correlation are symmetric. Since two uncorrelated signals can both correlate strongly with a third, correlations do not respect the triangle inequality. The dissimilarities $\Delta^{(\mathrm{corr})}$, however, do agree with the triangle inequality (Socolovsky, 2002). For discretely sampled signals of length N this is seen by the well-known representation of correlations as cosines of angles between unit vectors in a N-dimensional space, where the measure $\Delta^{(\mathrm{corr})}$ corresponds to moduli of sines between such vectors.

A non-linear, probabilistic generalization of correlation is mutual information (Shannon and Weaver, 1949; Gray, 1990; Cover and Thomas, 1991),

$$I(X_i, X_j) = \int_{X_j} \int_{X_i} p_{ij}(x_i, x_j) \log \frac{p_{ij}(x_i, y_j)}{p_i(x_i) p_j(x_j)} \; \mathrm{d}x_i \, \mathrm{d}x_j, \tag{6.16}$$

in which p_{ij} is the density of the joint probability distribution of X_i and X_j, and p_i and p_j are its respective marginals. By construction, mutual information is symmetric and non-negative. The triangle inequality, however, is not fulfilled. To show this, mutual information can be identified as the Kullback-Leibler divergence between the joint probability distribution and the product of its marginals (Kullback and Leibler, 1951), which is known to violate the triangle inequality (Cover and Thomas, 1991). However, one can readily modify mutual information to achieve metric properties, when first rephrasing (6.16) in terms of joint and conditional entropies as

$$I(X_i, X_j) = H(X_i, X_j) - H(X_i|X_j) - H(X_j|X_i), \tag{6.17}$$

which are defined as

$$H(X_i, X_j) = -\int_{X_j}\int_{X_i} p_{ij}(x_i, x_j)\log p_{ij}(x_i, x_j)\,\mathrm{d}x_i\,\mathrm{d}x_j,$$
$$H(X_i|X_j) = -\int_{X_j}\int_{X_i} p_{ij}(x_i, x_j)\log p_{i;j}(x_i|x_j)\,\mathrm{d}x_i\,\mathrm{d}x_j. \tag{6.18}$$

In contrast to mutual information these joint and conditional differential entropies alone are not very useful as distance measures, since probability densities can be greater than one point-wise, yielding negative entropies[5].

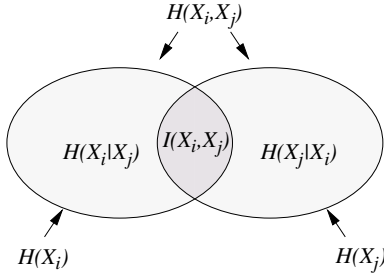


Figure 6.2: Relationship between mutual information and (conditional) entropies. More formally, one finds $I(X_i, lX_j) = H(X_i, X_j) - H(X_i|X_j) - H(X_j|X_i) = H(X_i) - H(X_i|X_j) = H(X_j) - H(X_j|X_i)$; after Cover and Thomas (1991).

On the other hand the relative conditional entropies, $H(X_i|X_j) + H(X_i|X_j)$, do provide a proper distance that fulfills (*i-iii*) (Cover and Thomas, 1991). Using (6.17) and normalizing that distance to the interval $[0, 1]$ yields the definition

$$\Delta_{ij}^{(\mathrm{m.inf})} = 1 - \frac{I(X_i, X_j)}{H(X_i, X_j)}. \tag{6.19}$$

Correlation and mutual information are two well-established quantities that estimate how much a random variable tells us about another. More recently, another related measure, namely Granger causality (Granger, 1969), has become quite popular in neuroscience (Kamiński et al., 2001). For the sake of legibility we do not dwell on the somewhat lengthy definition. Suffice to say that by fitting a parametric, linear auto-regressive model to the data under study, it can be observed whether the additional freedom offered by including x_j terms into the model for the i-th signal x_i does decrease the prediction error. For instance, consider two discretely sampled time series $x_i(t_k)$ and $x_j(t_k)$ and build a linear predictor of the current value of x_i

[5] Yet, there are some interesting links between mutual information and the Kolmogorov-Sinai entropy that is used in the context of classifying complex dynamics (Matsumoto and Tsuda, 1988; Deco et al., 1997); however, this discussion is beyond the scope of the current paper.

from m previous values by means of $x_i(t_n) = [\sum_{k=1}^{m} a_k x_i(t_{n-k})] + \epsilon_n$. The variance of ϵ_n provides an estimate of the resulting prediction error. Equivalently, one can build a linear predictor that includes x_j, that is, $x_i(t_n) = [\sum_{k=1}^{m} \tilde{a}_k x_i(t_{n-k})] + [\sum_{k=1}^{m} b_k x_j(t_{n-k})] + \varepsilon_n$. Again, the variance of ε_n measures the prediction error. If $\mathrm{var}(\varepsilon_n)/\mathrm{var}(\epsilon_n) < 1$, then the prediction of x_i is improved by incorporating x_j. Hence, x_j has a causal influence on x_i in the sense of Granger and $1 - \mathrm{var}(\varepsilon_n)/\mathrm{var}(\epsilon_n)$ is a measure of its strength. In the frequency domain this is quantified by an absolute off-diagonal value of the transfer matrix of the error. Put differently, it is the magnitude of the cross-coefficient of the 'noise' when fitting this residual part to the data; see (Dhamala et al., 2008a,b) for non-parametric versions in the context of neuroscience. Note that Granger causality is not a reflexive measure, as it takes the value unity for identical signals.

A profound problem with Granger causality is that a vanishing Granger causality does not exclude a causal relationship between two signals (Lütkepohl, 2005, Ch. 2.3.1). This measure is also not symmetric, and it does not fulfill the triangle inequality (Eichler, 2007). We note that instantaneous Granger causality, for which only the equal-time value of x_j is used in the prediction of x_i, is symmetric. However, the latter is zero if and only if the noise is uncorrelated, implying that in this case causality equals conventional statistical correlation.

6.3.2 Spectral measures

Since the frequency domain is dual to the time domain, all aforementioned connectivity measures have their spectral counterparts. Neural populations typically exhibit oscillatory activity with important functional roles reflected in distinct frequency bands (Singer, 1993, 1999; Buzsáki, 2006). A controversial hypothesis is the idea that neuronal information is represented by rate-coding, that is, that neuronal activity transmits information by frequency, and not by amplitude (Barlow, 1972; Gray, 1994; Theunissen and Miller, 1995; Friston, 1997; Eggermont, 1998). Convincing results on temporal and other kinds of coding did temper claims for the exclusiveness of rate-coding (Engel et al., 1992), yet it remains an attractive idea that keeps receiving support from spectral connectivity measures.

The covariance (6.13) between two signals is a scalar quantity that can be generalized by introducing finite time lags between the to-be-compared signals. We consider the correlation function of two signals that depend continuously on time, which yields the definition

$$c_{ij}(\tau) = \mathbb{E}\left[x_i(t+\tau)x_j(t)\right]. \tag{6.20}$$

Note that the correlation function is often normalized either via the auto-correlation at zero lag, $c_{ii}(\tau = 0)$, or via the (product of) corresponding standard deviation(s). The correlation theorem shows that the Fourier transform of the cross-covariance can

be identified with the cross-spectrum, i.e., the inner product between the individual Fourier transforms of x_i and x_j:

$$f_{ij}(\omega) = \mathbb{E}\left[\mathcal{F}\left[x_i(t)\right]\mathcal{F}^*\left[x_j(t)\right]\right] = \mathbb{E}\left[\frac{1}{2\pi}\iint x_i(t_1)x_j(t_2)e^{-i\omega(t_1-t_2)}\,\mathrm{d}t_1\,\mathrm{d}t_2\right]. \quad (6.21)$$

It can be considered the complex-valued analogue of the coefficient of variation. If the signals are identical, that is, if we consider the auto-correlation function $c_{ii}(\tau)$, this yields the power spectral density

$$f_{ii}(\omega) = \mathbb{E}\left|\mathcal{F}\left[x_j(t)\right]\right|^2 = \mathbb{E}\left[\frac{1}{2\pi}\iint x_i(t_1)x_i(t_2)e^{-i\omega(t_1-t_2)}\,\mathrm{d}t_1\,\mathrm{d}t_2\right], \quad (6.22)$$

an identity known as Wiener-Khintchine theorem.[6]

Normalizing the cross-spectrum f_{ij} leads to the measure coherency (Brillinger, 1981),

$$R_{ij}(\omega) = \frac{f_{ij}(\omega)}{\left(f_{ii}(\omega)f_{jj}(\omega)\right)^{1/2}}. \quad (6.23)$$

A trivial modification, (6.23) $\rightarrow (1-R_{ij})$, leads to a reflexive, symmetric measure. The complex-valued coherency contains a frequency-dependent amplitude (=conventional (squared) coherence, i.e., $|R_{ij}|^2$) and a phase spectrum. Alternatively, the cross-spectrum can be decomposed into real and imaginary parts, often denoted to as co- and quad-spectrum, respectively. The former is symmetric, the latter is anti-symmetric, and therefore the imaginary part of coherency (i.e., $\mathrm{im}\,R_{ij}$) appears well-suited for the study of directional influences (Nolte et al., 2004, 2008). Note that the quad-spectrum does not contribute to the mean spectral power but only modulates the power spectral density, though in practice numerical limitations in estimating the Fourier transforms may render this separation less strict. Since identical signals result in zero $|\mathrm{im}\,R_{ij}|$, this is also not a reflexive measure.

It is important to note that all the above is strictly speaking only valid for (weakly) stationary signals without discrete frequency components. For non-stationary signals, these definitions have to be modified, as the covariances become explicitly time-dependent. In consequence, the cross-spectrum will also depend on time, but many results can be easily generalized, requiring only slight modifications. An interesting approach in that regard are correlations between Wavelet coefficients (Quyen et al., 2001; Achard and Bullmore, 2007). Various further issues particularly tailored for applications in neuroscience have been discussed in (Nunez et al., 1997).

Granger causality has indeed been defined in the spectral domain by Granger (1969), who considered it a generalization of coherence and proposed for this pur-

[6] Computing the Fourier transform might be problematic because dependent on the explicit form of x_i the integral $\int x_i(t)e^{-i\omega t}\,\mathrm{d}t$ may not exist. A similar problem can also arise for the cross-spectrum of long-range dependent signals, but we cannot consider these interesting issues here.

pose the term causality coherence. This idea was extended to the multivariate case by Geweke (1982). Moreover, it was thereby shown that Granger causality decomposes into an anti-symmetric and a symmetric part, where the latter is the instantaneous Granger causality mentioned above in Section 6.3.1. Similar to the imaginary part of coherency, the study of this anti-symmetric part can provide insight into causal relationships; with all the aforementioned limitations, of course.

6.3.3 Non-linear measures

In contrast to the stochastic approach of Section 6.3.2, the non-linear analysis of functional connectivity generally considers the brain a dynamical system. Again, the oscillatory character of neural activity plays an important role as it is closely related to the mathematical notion of recurrent behavior. It is thus no coincidence that most non-linear connectivity measures are spectral-like measures (Quyen and Bragin, 2007), in particular phase relations are relevant (Sauseng and Klimesch, 2008). Non-linear time series analysis (Schreiber, 1999; Kantz and Schreiber, 2004) is, however, a broader discipline. Its starting point is the reconstruction by delay-vector embedding (Takens, 1981; Stark, 2000), with which one tries to reconstruct the deterministic aspect of the dynamics from its temporal differences $x(t+\tau) - x(t)$, instead of its autocorrelations. The former can be interpreted as finite differences in a Taylor expansion of the flow (vector-field) of the dynamics, as Packard et al. (1980) suggested. Since the Fourier transform of derivatives corresponds to powers, this is topologically equivalent to an approximation by a power series. There is a vast amount of studies on statistical measures derived from this reconstruction, mostly within the physics community. As stated before, Stam (2005) reviewed many of these in the context of their clinical applications. A measure specifically designed to quantify non-linear interdependences of EEG signals based on prediction errors has been described by Breakspear and Terry (2002). More recently, measures derived from recurrence plots have become popular (Webber, Jr. and Zbilut, 1994; Marwan et al., 2007). Synchronization likelihood is a well-known example for a measure that quantifies recurrences in a quite general way (Stam and van Dijk, 2002). Most of these measures are (or can be trivially modified to be) reflexive and symmetric.

Phase relationships are immediately interpretable (Kreuz et al., 2007), as they are based on the notion of conventional synchronization (Boccaletti et al., 2002). The phase uniformity is defined as the length of the resultant vector of the instantaneous phases of signals that dependent continuously on time, or, alternatively, of discretely sampled data (Mardia and Jupp, 2000) and has variously been referred to as mean phase coherence (Mormann et al., 2000) or phase locking value or index (Lachaux et al., 1999; Sazonov et al., 2009). It is usually applied to phase differences (Boonstra

et al., 2006; Houweling et al., 2008), for which it reads,

$$\Delta_{ij}^{(\mathrm{univ})} = \frac{1}{T}\left|\int_T e^{i(\phi_i(t)-\phi_j(t))}\,\mathrm{d}t\right| \quad \text{or} \quad \Delta_{ij}^{(\mathrm{univ})} = \frac{1}{N}\left|\sum_{k=1}^{N} e^{i(\phi_i(t_k)-\phi_j(t_k))}\right|. \tag{6.24}$$

The value $1-\Delta_{ij}^{(\mathrm{univ})}$ is known as phase dispersion or (relative) circular variance (Batschelet, 1981; Mardia and Jupp, 2000). Its distribution under a uniform probability density is used in the Rayleigh test. The uniformity is a symmetric measure, but note that the resultant (mean phase) is reflected, and like conventional uniformity it does not fulfill the triangle inequality; the reasoning is equivalent to the violation of the inequality for the afore-discussed covariance. Variants of phase uniformity are the phase entropy of Tass et al. (1998), the phase lag index (Stam et al., 2007), and the bi-phase locking value (Darvas et al., 2009). The latter is asymmetric. The multivariate case has been discussed in (Hutt et al., 2003).

6.3.4 Wasserstein distances

Since most of the afore-listed connectivity measures are not metric, we finally describe a very general class of connectivity measures, Wasserstein distances, that do respect all properties (*i*)-(*iii*). Apart from being true distances, these measures have remarkable properties that will be briefly sketched.

Wasserstein distances are general distances between probability distributions and can be defined for any probability distribution given on a metric space. Consider two probability measures p_i and p_j that assign probabilities $p_i[U_i] \geq 0$ and $p_j[U_j] \geq 0$ to suitable subsets $U_i \times U_j \subseteq X_i \times X_j \subseteq \mathbb{R}^{2m}$ of some multivariate space. These measures can be absolutely continuous, i.e., represent probability densities, singular, or even fractal.

The Wasserstein distance $\Delta_{ij}^{(\mathrm{Wass;q})}$ of order $q \geq 1$ between p_i and p_j is given by the solution of an optimal transportation problem in the sense of Kantorovich (Villani, 2003). It measures the amount of work, or distance times probability mass transported, that is needed to transform p_i into p_j, weighted according to q. Formally, it is given by the functional

$$\Delta_{ij}^{(\mathrm{Wass;q})} = \left(\inf_{\Pi}\int_{X_i\times X_j} ||x_i - x_j||^q \;\mathrm{d}\pi(x_i,x_j)\right)^{1/q} \tag{6.25}$$

that is optimized over all (joint) probability measures $\pi \in \Pi$ with prescribed marginals p_i and p_j:

$$p_i(U_i) = \int_{X_j} \mathrm{d}\pi[U_i, x_j] \qquad \text{and} \qquad p_j(U_j) = \int_{X_i} \mathrm{d}\pi[x_i, U_j]. \tag{6.26}$$

In the usually encountered case of discrete mass distributions, this definition reduces to a convex optimization problem known as the transportation or transshipment problem. Then, the distributions p_i and p_j can be considered weighted point sets

$$p_i = \sum_{k=1}^{n_1} \alpha_k \delta_{x_k}, \qquad \text{and} \qquad p_j = \sum_{l=1}^{n_2} \beta_l \delta_{y_l}, \tag{6.27}$$

in which the supplies $\alpha_k \in (0, 1]$ and the demands $\beta_l \in (0, 1]$ are such that $\sum_k \alpha_k = \sum_l \beta_l = 1$. Any measure in Π can then be represented as a non-negative matrix $\boldsymbol{G}$ that fulfills so-called source and sink conditions

$$\sum_l G_{kl} = \alpha_k, \quad k = 1, 2, \ldots, n_1, \quad \text{and} \quad \sum_k G_{kl} = \beta_l, \quad l = 1, 2, \ldots, n_2. \tag{6.28}$$

These are in fact discrete analogs of the conditions on the marginals in Eq. 6.26. Finally, the Wasserstein distance $\Delta_{ij}^{(\text{Wass;q})}$ is given by the solution of the transportation problem

$$\Delta_{ij}^{(\text{Wass;q})} = \min \left(\sum_{kl} G_{kl} ||x_l - y_l||^q \right)^{1/q} \tag{6.29}$$

over all matrices $\boldsymbol{G}$. It can be explicitly solved in polynomial time (of complexity about N^3) by a network simplex algorithm (Balakrishnan, 1995; Schrijver, 1998); see Löbel (1996) for a proper implementation.

Remarkably, $\Delta_{ij}^{(\text{Wass;q})}$ is a true distance in the space of all probability measures on X_i and X_j, i.e., it is (strongly) reflexive and symmetric by construction, but the triangle inequality is non-trivial to establish (Villani, 2003). Note that the metric distance $d(x, y) = ||x - y||$ can be replaced by an arbitrary distance function. Although most commonly Euclidean distance is used, when specializing to the discrete distance ($d_0(x, y) = 1$ if and only if $x \neq y$) the corresponding $q = 1$ Wasserstein distance is (one-half of) total variation, i.e., the integrated absolute difference between two probability distributions. The orders $q = 1$ (Rubinstein-Kantorovich distance) and $q = 2$ (quadratic Wasserstein distance) are most often used; the latter has further important properties, e.g., it is possible to interpolate between signals in functional space reconstructed from this distance (Villani, 2003, Ch. 5).

Wasserstein distances have a plenitude of applications in statistics, image registration (Haker et al., 2004), inverse modeling (Frisch et al., 2002), and classification where they are known as the Earth Mover's distance (Rubner et al., 2000). They have also been used to define a distance between (non-linear) time series (Moeckel and Murray, 1997), known as the transportation distance. This distance assumes an underlying dynamical system for each time series and employs the aforementioned delay-vector embedding procedure (Takens, 1981) to map each scalar time series into

the same k-dimensional reconstruction space $\Omega = \mathbb{R}^k$. The time average

$$p_i = \frac{1}{n} \sum_{k=1}^{n} \delta_{x_{i,k}} \tag{6.30}$$

of the indicator function of the points $x_{i,k} \in X_i$ visited by the i-th dynamical system is used as the (empirical) probability measure; here δ_x is the Dirac measure concentrated at the point x. Measuring the similarity of these time averages, which form invariant measures in the limit of infinite time series, is considered in detail in (Muskulus and Verduyn-Lunel, submitted). It has been applied to sensor MEG signals collected during listening to auditory stimulation (Muskulus and Verduyn-Lunel, 2008b), which revealed evidence of hemispheric specialization even in rather simple task circumstances; see below.

However, the mathematical elegance of this measure has its price: when compared with conventional distances, like the ones implemented in spectral analyses, the Wasserstein distances are computationally much more demanding. In fact, for time series longer than a few thousand samples at present one needs to sample smaller subseries and compute the mean Wasserstein distances via bootstrapping techniques (Davison and Hinkley, 1997). Notably, in a different context these distances have already shown superior classification abilities, namely in the analysis of lung diseases (Muskulus and Verduyn-Lunel, 2008a). We believe they can form a major characteristic in quantifying neurophysiological signals and may hence be particularly important to qualify data in neuroscience.

6.4 Example: MEG data during motor performance

To discuss the impact of (the violation of) metric properties (*i*)-(*iii*) for the analysis of neurophysiological signals, we illustrate the above techniques with functional connectivity data obtained from MEG recordings during execution of a bimanual task. An in-depth analysis of the experiment can be found in (Houweling et al., 2008). In brief, subjects performed a so-called 3:2 polyrhythm in which right and left index finger simultaneously produced isometric forces at frequencies of 1.2 Hz and 0.8 Hz, respectively. Brain activity was measured with a 151 channel whole-head MEG (CTF Systems, Vancouver) and bipolar electromyogram (EMG) was assessed from bilateral extensor and flexor digitori with a reference at the left wrist. The MEG signals were mapped to source space using synthetic aperture magnetometry (SAM) beamformers (Vrba and Robinson, 2001). Here we restricted ourselves to the discussion of two SAM sources located in primary motor areas that showed maximal power contrast in the β-frequency band. We further included the bilateral EMGs as we are generally interested in inter-hemispheric and cortico-muscular interactions. All signals were filtered in the lower β-frequency band using a 4-th order bi-directional

Butterworth band-pass filter (20–25 Hz), prior to computation of the instantaneous phases $\phi_i(t)$ of the analytic signals (computed via the Hilbert transform).

Since force production was rhythmic, we defined motor events as instances of maximal increases in left or right force and evaluated two time points, 100 ms before and 250 ms after each of these events, as they coincided with maximal β-power changes; cf. Fig. 7 in (Houweling et al., 2008). Duration of recordings was 30 min (excluding short rest periods and task initiation), implying about 1600 events on the left and about 2000 events on the right side over which measures were estimated. In fact, the design yielded two asymmetric $(4{\cdot}2 \times 4{\cdot}2)$ connectivity matrices for the left and right hand events, respectively. Their elements were either defined via relative circular phases over events or via the Wasserstein distances between the corresponding phase distribution (estimated over events). As said here the data merely serve to illustrate procedure, so that we used data from a single subject, for which we will discuss MDS results in a two-dimensional embedding. The statistical evaluation of results will be addressed elsewhere.

The relative circular variance values at all 24 off-diagonal combinations of the four signals, $\mathrm{M1}_{\mathrm{left}}$, $\mathrm{M1}_{\mathrm{right}}$, $\mathrm{EMG}_{\mathrm{left}}$, and $\mathrm{EMG}_{\mathrm{left}}$, and two time-points, t_{pre} and t_{post}, for both types of motor events (left and right forces) are depicted in Figure 6.3. The most clear-cut difference of variances is found between left and right M1s and between M1s and their contralateral EMG, the latter primarily in the contrast between $(t_{\mathrm{pre}}, t_{\mathrm{pre}})$-$(t_{\mathrm{post}}, t_{\mathrm{post}})$.

Obviously, the pair-wise comparison forms a challenge, but when represented by MDS more structure comes to the fore; see Figure 6.4. In fact, the MDS reconstruction provides a reasonable representation, as 'functionally similar' signals are located close to each other. Obviously, the two M1s are functionally similar as the corresponding circle-centers are close. Considering the equal time points, all four signals are connected by U-shape forms; see the dashed lines in Figure 6.4 linking $\mathrm{EMG}_{\mathrm{right}}$-$\mathrm{M1}_{\mathrm{left}}$-$\mathrm{M1}_{\mathrm{right}}$-$\mathrm{EMG}_{\mathrm{left}}$. Both time points displayed these U-shapes, but for post-event times especially the distances between M1s are increased. Recall that we study the relative phases of the β-band. That is, an increase in functional distance relates to an increased relative circular variance, or in other words, β-desynchronization, which is well-known in the rhythmic isometric force production (Boonstra et al., 2006; Houweling et al., 2008).

Importantly, the triangle inequality was violated ($\epsilon = 0.009$) for the phases from the left event, rendering the interpretation of the centers in the right panel of Figure 6.4 questionable. By the same token, however, this may explain why the left panel provides a much clearer picture. To complement this information, Figure 6.4 also depicts the reconstructed invariant eigenvector (scaled to unit length), for which the functional connectivities are re-interpreted as transition probabilities. For this sake the eigenvector was not computed using the relative circular variance but for the uniformity matrix $\mathbf{\Delta}^{(\mathrm{univ})}$, interpreted as transition probabilities after normal-

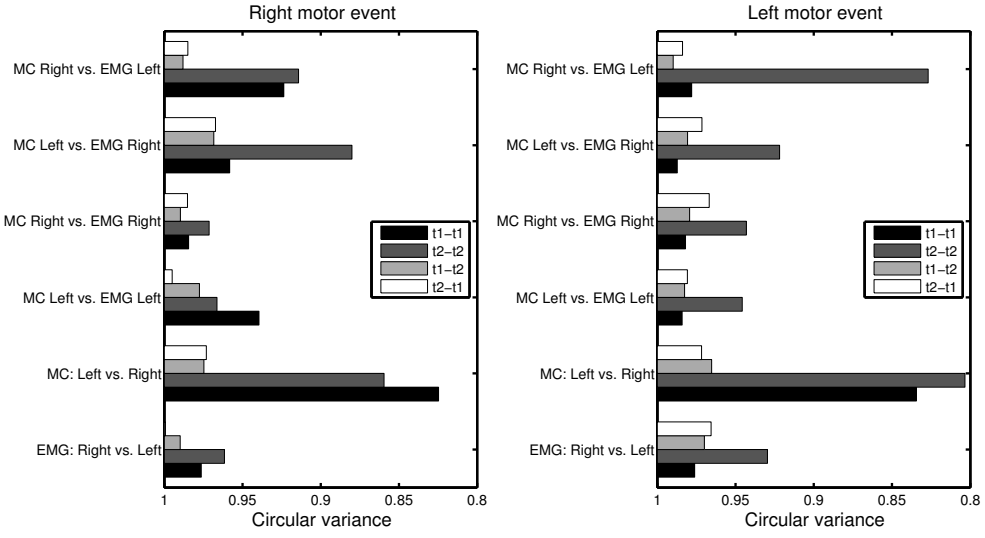


Figure 6.3: Circular variance of experimental data for all combinations of signals (see text). MC: Motor cortex. EMG: Electromyograph. T1: 100 ms before motor event. T2: 250 ms after motor event.

izing its row sums; the diagonal elements were set to zero, $\Delta_{ii}^{(\mathrm{univ})} = 0$, to focus on interactions between signals (or times) rather than incorporating the fact that a signal interacts with itself at the same time. Clearly, the two M1s can be identified as 'functional centers' as the radii of the circle are comparatively large. For the right motor event, the radius around $\mathrm{EMG}_{\mathrm{right}}$ is larger after the event, i.e., at t_{post}, showing its synchrony after maximal force (trivially at t_{pre} the EMG was less important). Even more pronounced is this effect visible for the left motor events, as the $\mathrm{EMG}_{\mathrm{left}}$ has a significant contribution to the invariant eigenvector. Interestingly, the contribution before maximal force matched largely that after the motor event.

As mentioned above, we repeated the entire procedure using the (quadratic) Wasserstein distances from the phase distributions instead of relative circular variances. That is, we constructed the connectivity matrix given in (6.25) using the same samples as described above. Distances between phases are geodesic, i.e., $||\varphi_1 - \varphi_2|| = \min(|\varphi_1 - \varphi_2|, 2\pi - |\varphi_1 - \varphi_2|)$, and the Wasserstein distances quantify differences between the distributions of phases over events. To accelerate computations, we bootstrapped the distances, sampling 512 phases randomly a hundred times. From the mean of these distances the MDS representation, i.e., the circle centers in Figure 6.5, was derived, where links between elements are highlighted equivalent to Figure 6.4.

Interestingly, the Wasserstein distances revealed a markedly different pattern than the circular variances. In particular, the two types of events yielded differ-

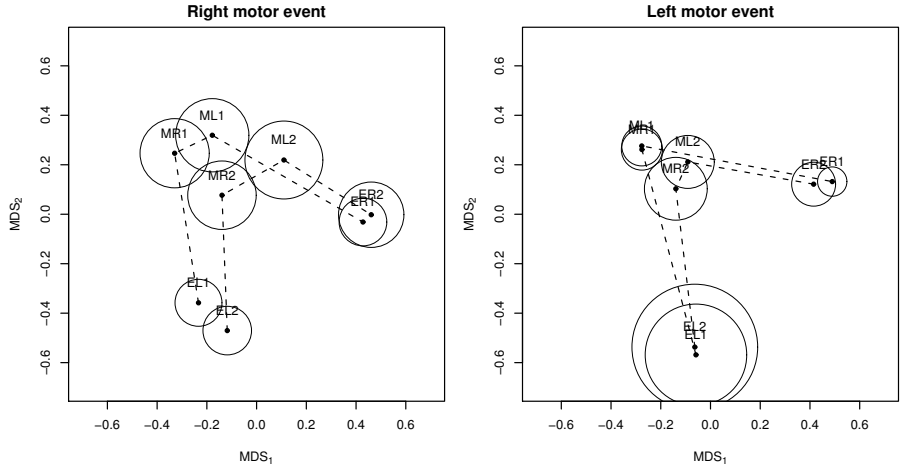


Figure 6.4: MDS results for the circular variances of Fig. 6.3; the four signals at two distinct time steps before and after the left/right motor events are presented as points (circle centers) in a two-dimensional functional space (Friston et al., 1996). The circles around the points indicate the invariant eigenvector for the uniformity (i.e., the radius of a circle is the value of the eigenvector's corresponding component); recall the sketch of Page Ranking in Section 6.2.3. E: EMG, M: Motor cortex, L: Left, R: Right, 1: 100 ms before motor event, 2: 250 ms after motor event; see text for further details.

ent result: the Wasserstein distances unraveled a difference in the phase dynamics between events at which the right force was produced, and the events in which the left force was produced, possibly reflecting differences in functional integration due to handedness (the subject was right-handed). For the latter, we again observed an (almost) U-shaped pattern for equal times. For the right-hand side events, however, the arrangement of the equivalent functional relationships was quite different and formed an X-like shape (recall that we linked $\mathrm{EMG}_{\mathrm{right}}$-$\mathrm{M1}_{\mathrm{left}}$-$\mathrm{M1}_{\mathrm{right}}$-$\mathrm{EMG}_{\mathrm{left}}$). That is, for the right force events the M1s were closer to the ipsilateral EMGs than the contralateral EMGs. As before, this X-shape was present for both time points. This indicates indirect ipsilateral phase synchronization, most probably via a cross-talk between bilateral M1s.

The invariant eigenvector, computed from $(1-\delta_{ij})-\Delta_{ij}^{(\mathrm{Wass};2)}$ after normalizing the row-sums, is again shown via circles, normalized to length $1/8$, consistent with the smaller scale in Figure 6.5. It is almost equally distributed along all signals, which indicates that the magnitude of the Wasserstein distances was more or less comparable for all signals. Therefore, this eigenvector does not supply additional information regarding functional integration in this simple example.

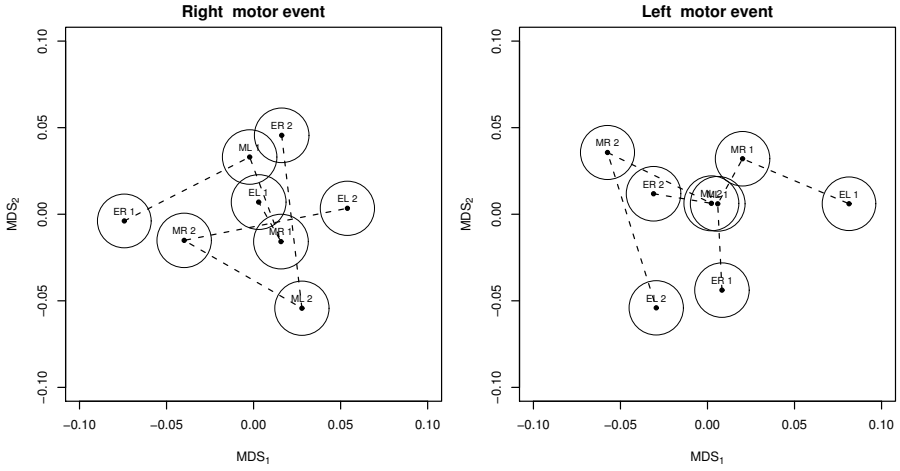


Figure 6.5: The MDS representation of the quadratic Wasserstein distances. E: EMG, M: Motor cortex, L: Left, R: Right, 1: 100 ms before motor event, 2: 250 ms after motor event. The components of the invariant eigenvectors are again given by the size of the surrounding circles; compare with Figure 6.4.

6.5 Example: Auditory stimulus processing[‡]

As an illustrative example for the application of Wasserstein distances to electrophysiological time series, we present results obtained from baseline measurements (resting state) in the experiment by Houweling et al. (2008). An auditory stimulus was presented to the right ear (EARTone 3A, CABOT Safety Corporation) at a pitch of 440Hz, frequency of 1.2Hz, and with a duration of 50ms. Magnetoencephalographic time series were recorded at 1.25kHz sampling frequency over a 20s interval. Downsampling to 250Hz yielded 5000 time points. The left panel of Fig. 6.6 shows a MDS representation of the sensors' Euclidean distances, and the right panel a representation of their averaged distances, when the sensors were grouped in 14 subsets. The latter has been done for visualization purposes, mainly.

For simplicity, only data for a single subject are discussed here. The remaining subjects showed essentially the same features. The MEG time series were normalized and centred, and attractors reconstructed with a lag $q = 10$ and embedding dimension $k = 5$. For each pair of sensors the Wasserstein distances were bootstrapped three times with 500 sample points each. Grouping the sensors into the 24 groups

‡ The contents of this section were originally published in:
Muskulus M, Verduyn-Lunel S — Reconstruction of functional brain networks by Wasserstein distances in a listening task. In: Kakigi R, Yokosawa K, Kurik S (eds): Biomagnetism: Interdisciplinary Research and Exploration. Hokkaido University Press. Sapporo, Japan (2008), pp. 59-61.

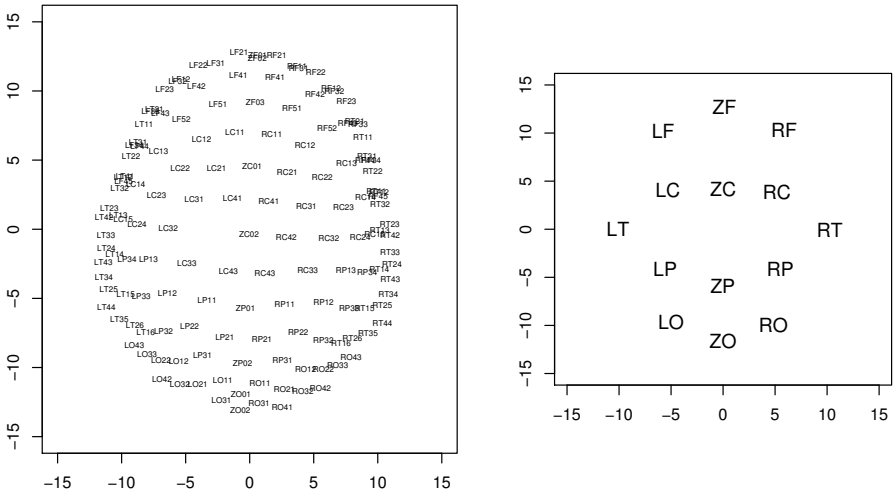


Figure 6.6: Left panel: MDS reconstruction of distances of the 148 sensors considered. Right panel: MDS reconstruction of aggregated distances of the 24 sensor groups. L: left, R: right, F: frontal, C: cranial, T: temporal, Z: central, P: parietal, O: occipital.

shown in the right panel of Fig. 6.6, the corresponding distances were aggregated into a 24-by-24 matrix of average distances between sensor groups. The left panel of Figure 6.7 shows a two-dimensional MDS representation of these distances.

These distances represent the difference in the dynamics of the MEG time series. In the listening task under study, the dynamics of auditory processing should be similar. Indeed, the RT group, where the auditory cortex is located, has a small distance from the LF group, where speech processing takes place (indicated by an arrow in Figure 6.7) (Kandel et al., 2000). This is independent evidence for *hemispheric specialization*: even in the absence of linguistic information, the relevant cognitive areals in the left hemisphere are involved in the processing of auditory signals; a fact that had been previously established and discussed in (Helen et al., 2006).

6.6 Conclusion

We have discussed commonly used functional connectivity measures with respect to their metric properties. Most measures do not fulfill the triangle inequality. This is particularly unfortunate since only proper (pseudo-) metric measures allow for interpreting connectivities as functional distances. For instance, if the triangle in-

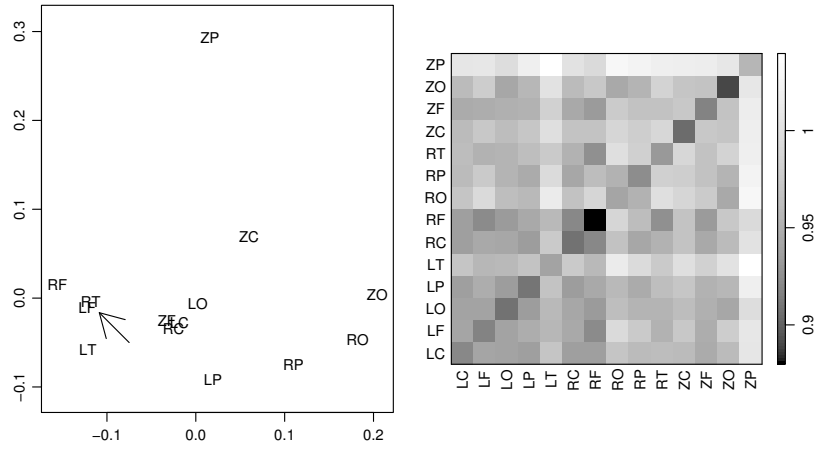


Figure 6.7: Aggregated Wasserstein distances of MEG sensors. Left panel: Two-dimensional MDS reconstruction. Abbreviations as in Fig. 6.6. The arrow indicates the feature of main interest. Right panel: The distance matrix.

equality is not fulfilled, this may hint at the existence of further subsystems that may remain hidden for further analysis. Moreover, this may compromise the analysis by causing spurious results; the Page Ranking algorithm may resolve these under specific conditions, presuming non-trivial similarity measures are employed, but it is preferable to use measures that can be directly interpreted.

If the central properties, symmetry, reflexivity, and the triangular inequality are fulfilled, then MDS can be used to map these distances into a functional space. That representation allows for visualizing the interrelationships between signals and should be considered a very valuable tool for the formulation of hypotheses regarding functional relationships between neurophysiological processes. It also allows for the application of discriminant analysis, and further analysis of functional connectivities by morphometric methods (Small, 1996).

As very few functional connectivity measures are metric, we suggest the use of Wasserstein distances. These distances allow for the general assessment of the coupling strength between two signals, either by considering them dynamical systems, or by quantifying differences in probability distributions, e.g., of their instantaneous phases. For the latter case, the example of MEG data served to illustrates their strength when combined with MDS. A one-step analysis unraveled cortico-muscular synchrony as well as cross-talk between bilateral primary motor areas in the β-band. For the application directly in the time domain, i.e., when considering

the brain a dynamical system, validation of this approach with simulated time series from a forward model will be presented elsewhere.

Epilogue

The question is not what you look at, but what you see.

Henry David Thoreau

Distances & Measurements

The concept of distance is basic to the human mind. We often do not qualitatively compare the objects of our thoughts, but prefer to explicitly express our ideas about how similar or dissimilar we judge them to be. Even with regard to psychological states, we like to distinguish different degrees of involvement, excitement, attachment, etc. The urge to classify and order the phenomena in the world seems to be a basic human need. Aristotle, the great ancient classificator, constructed extensive systems of thought in which phenomena were categorized and assigned to disjoint classes with distinct properties. Although the application of the "Aristotelean knife" (Robert Pirsig) has led to many unnatural and problematic classifications in hindsight, it was nevertheless extremely fruitful in that it imposed order on the world, enabling a large and still ongoing scholarly activity.

The next important step in the scientific enterprise was the shift from a purely mental exercise to actual experimentation. Instead of reasoning about possible causes and relations in the natural world, researchers were actively asking questions and trying to construct theories that were consistent with the facts obtained. Abstracting from the individual reasearcher and his situation, science was founded on the notion of *universality*: Patterns observed under a given experimental situation should be reproducible in a different location and time, even by a different scientist. The basic tool that allows for such an *objective* approach is the notion of a *measurement*. Thereby, the objects of our inquiry are related in a prescribed way to standardized

models, allowing the scientist to extract universal information. This information is conveniently represented in the strict language of mathematics, which is universal by its underlying logical foundation.

The different levels of measurement have been defined by Stanley Smith Stevens in an influential article (Stevens, 1946). *Nominal* measurements correspond to Aristotle's legacy of classification: Objects and their properties are assigned and distinguished by labels. Mathematically, this is the domain of set theory. On the *ordinal* level, objects are ordered, corresponding to a totally (or linearly) ordered set. Next, *interval* measurements allow for quantitative statements. Properties measured are mapped to numbers, and the notion of distance surfaces. An example is the Celsius scale for temperature, where one degree Celsius is defined as one hundredth of the difference in temperature between water at the melting and the boiling point, respectively. Remarkably, measurements on an interval scale are relative, i.e., only distances are well-defined, and there does not exist a designated origin[7]. Mathematically, such measurements correspond to affine spaces. Finally, *ratio* measurements are expressed on a scale that possesses a non-arbitrary zero value.

A great deal of early science was involved with the search for the most elementary properties by which we can compare the objects in our world. This has led to the development of *systems of measurement*, sets of units specifying anything which can be measured. The international system of units (SI) identifies seven distinct kinds of physical quantities that can be measured: length, mass, time, electric current, temperature, luminous intensity and amount of substance. Notwithstanding its great success in the commercial and scientific domain, it can be argued whether this is a complete or natural list.

Looking back

Here we were concerned with more abstract quantities. The objects considered in this thesis are *complex systems* that can not be easily reduced to one or more fundamental properties: the respiratory system, the brain, dynamical systems. Even if it were possible to project the state of such a complex entity to a single number, the great loss in information incurred does not make this an attractive proposal. Therefore, instead of extracting a single property from such a complex system, we have considered ways in which we can *compare* systems quantitatively with each other. Although this did also result in a single numerical quantity, namely, a *distance* between each pair of systems under consideration, *the totality of all such distances contains a much greater amount of information.* This simple fact was the starting point for the

[7] Although it is commonly said that water freezes at 0 degree Celsius, the Celsius scale was not intended to be used for such absolute statements. Rather, the correct way would be to say that "water freezes at a temperature difference of 0 degrees from the temperature where water freezes".

methods developed and applied in the rest of this thesis. It is *not* obvious, and special methods were needed to extract this information from the measured distances.

The central questions considered in this thesis were the following:

- How can we define a useful distance for complex systems?
- What kind of information is obtained from such a distance and how can we analyze it?
- What does this tell us about the original systems?

These questions are increasingly diffult to answer. It is not too difficult to define interesting "distances" between complex systems, although a few pitfalls need to be avoided. In particular, in order to allow for sensible comparisons between more than two distinct systems, a "distance" measure needs to be a true distance (in the mathematical sense), i.e., it needs to exhibit metric properties that allow for a consistent and natural interpretation in such a multivariate setting. This seriously restricts the class of possible "distance" measures, and involves an important principle: Being a true distance allows for a natural representation of complex systems as points in an abstract *functional* space, which is a very powerful way to visualize and analyze differences and commonalities between complex systems. For general "distance" measures such a representation is usually not possible. Indeed, it is well known that bivariate measures (such as "distances") can, and generally do, lead to spurious or even false results when applied in a multivariate setting (Kus et al., 2004). This problem is completely avoided by using a true distance. Of course there is a price to pay for this convenience: It might not be easy to find a suitable, true distance for the systems we want to study. And even if we obtain such a measure, it is not clear that it then also captures the relevant information about a system. Fortunately, the class of *optimal transportation distances*, are general enough to be both applicable in most settings, and in such a way that they capture interesting information.

The geometrical and statistical analysis of distances is also a rather well-developed topic, so we mostly did connect results scattered in the literature and closed a few gaps. However, what is actually measured in such an interval-scale approach is a completely different matter. The first two questions were addressed in a *phenomenological* setting: it is not necessary to know exactly what causes differences in complex systems, if one is primarily interested in the existence of such differences. For example, in the application to the respiratory system, we were interested in distinguishing healthy breathing from breathing with a diseased lung, which is a simple supervised classification task — albeit one of considerable interest. Since such classification was possible, we might now ask why this is the case. Then the question of how to reverse-engineer the information obtained from abstract distances becomes important. This road is mostly unexplored so far.

The future

The examples in this thesis demonstrate that the combination of optimal transportation distances, reconstruction of these distances by multidimensional scaling, and canonical discriminant analysis of the resulting coordinates is a powerful and versatile approach to the classification and study of complex systems. This thesis is finished, but many paths remain still to be explored. Let me mention a few here that have not been discussed in the earlier chapters.

- On the practical side: The calculation of the Wasserstein distances is still too complex (i.e., slow) to handle large datasets (with more than a few hundred to thousand sample points per subject). Bootstrapping smaller subsets helps a long way in reducing the computational complexity, but algorithmic improvements would be preferable. A number of interesting approximation algorithms have been developed in recent years, and implementing these as actually usable software would be desirable.
- How can classification based on nonmetric multidimensional scaling be cross-validated? Since nonmetric reconstructions are usually obtained by an iterative procedure, this is not as simple as it sounds. Optimally matching the resulting point configurations (de Leeuw and Meulman, 1986) might be one possibility to proceed.
- To avoid the quadratic dependence on sample size when computing all pairwise distances between N samples, is it possible to reconstruct Euclidean configurations locally, i.e., by only using the distances of the closest $k \ll N$ points?

Appendices

Appendix A

Distances

Life is like a landscape. You live in the midst of it but can describe it only from the vantage point of distance.

Charles Lindbergh

In this appendix we collect and discuss background information about distances and their statistical analysis. Section A.1 reviews the mathematical foundation and culminates in the characterization of the conditions under which a reconstruction of distances by points in an Euclidean space is possible. Section A.2 discusses how to obtain such reconstructions in practice and introduces various diagnostic measures that help to assess their quality. Section A.3 discusses the statistical analysis of distances and describes linear discriminant analysis in the reconstructed functional space, leave-one-out crossvalidation and permutation tests for group effects.

A.1 Distance geometry

The content of this section is developed in more detail in the standard monograph of Blumenthal (1953) and the article of Havel et al. (1983).

A.1.1 Distance spaces

An *abstract space* is a set of elements S, called *points*, that are endowed with a *topology*. The latter embodies a relation of nearness that results from defining certain subsets as *open*. A topology on S is then a collection $\mathcal{T}$ of all open subsets of S, such that the empty set $\emptyset$ and S are in $\mathcal{T}$, the union of any collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$, and the intersection of any finite collection of sets in $\mathcal{T}$ is also in $\mathcal{T}$.

The main use of a topology is to allow for the definition of limits of sequences of elements. A sequence $(p_1, p_2, \dots)$ of elements $p_i \in S$ has a *limit* $p \in S$ if and only if for each integer $n \in \mathbb{N}$ there exists an open set $U_n \in \mathcal{T}$ such that $p \in U_n$ and $p_m \in U_n$, for all $m \geq n$, which is written as $\lim_{i \to \infty} p_i = p$.

Abstract spaces are too general in practice, since they do not need to have unique limits. For example, endowing a space S with the trivial topology $\mathcal{T} = \{\emptyset, S\}$, every point $p \in S$ is the limit of every sequence. Therefore, we will only consider the subset of abstract spaces that are also Hausdorff spaces. These have the following additional property (restriction): If $p \neq q$ for two points $p, q \in S$, then there exist

open sets $U_p, U_q \in \mathcal{T}$ such that $p \in U_p, q \in U_q$ and $U_p \cap U_q = \emptyset$. Since Hausdorff spaces separate their points, they are also called separated spaces.

Although the above notions are necessary for the study of functions on S, in particular, to define the concept of continuity, as a basis for making measurements in a space S additional structure is needed. This will again be axiomatically prescribed.

Definition 6. A *distance space* is an abstract set S together with a *distance* $d : S \times S \to D$ from an abstract *distance set* D.

We write $d(p, q) \in D$ for the value of the distance between two points $p, q \in S$. The most important case are numerical distances:

Definition 7. A distance space is called *semimetric* if (i) $D \subseteq \mathbb{R}_+$, (ii) $d(p, q) = d(q, p)$, and (iii) $d(p, q) = 0$ if and only if $p = q$.

Here $\mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$ is the set of all non-negative real numbers. We can express the last two conditions in Definition 7 by saying that distances in a distance space are *symmetric* and *positive definite*, or simply by saying that they are *semimetric*.

Definition 8. The distance $d : S \times S \to D$ is *continuous* at $p, q \in S$, if for any two sequences $(p_n)_{n\geq 0}$ and $(q_n)_{n\geq 0}$ with limits $\lim_{n\to\infty} p_n = p$ and $\lim_{n\to\infty} q_n = q$, we have that $\lim_{n\to\infty} d(p_n, q_n) = d(p, q)$.

Continuous distances impose a certain regularity on distance spaces:

Theorem 3 (Blumenthal (1953))**.** A distance space with a continuous distance is Hausdorff.

Although $d(p, q) = 0$ if and only if $p = q$, there nevertheless still exists a potential anomaly in that two distinct points of a semimetric space may be joined by an arc of zero length:

Example 2 (Blumenthal)**.** Let $S = [0, 1]$ be the points of the unit interval and define the distance $d(x, y) = (x - y)^2$ for all points $x, y \in S$. Topologically, this space is equivalent to the space obtained by replacing d by the Euclidean distance $|x - y|$, so its character as a continuous line segment is unchanged, i.e., S is an arc.

Consider the sequence of polygons P_n with vertices

$$0, 1/2^n, 2/2^n, \ldots, (2^n - 1)/2^n, 1.$$

Each pair of consecutive vertices has distance $1/2^{2n}$ and since there are 2^n such pairs, the "length" of P_n is $1/2^n$. In the limit that $n \to \infty$, the length of S approaches zero. □

This anomaly results from the great freedom offered by the distance function, whose values are independent of each other, in the sense that the distance between

any pair of points does not depend on the distance between any other pair. Considering the simplest case of only three points, with three mutual distances, the following property is suggested from a closer look at Euclidean space:

Postulate 1 (Triangle inequality)**.** If p, q, r are any three points of a semimetric space, then

$$d(p,q) \leq d(p,r) + d(r,q). \tag{A.1}$$

Definition 9. A semimetric space in which the triangle inequality holds is called a *metric space*. The distance function of a metric space is called a *metric*.

Remark 6. The triangle inequality can be motivated differently. Let $(a,b),(c,d)$ be two pairs of ordered points in a semimetric space, and define $d(a,c)+d(b,d)$ as the distance of the pairs. When is this distance *uniformly continuous*? By this we mean that for each $\epsilon > 0$ there exists a number $\delta(\epsilon) > 0$ such that for all pairs $(a,b),(c,d)$ the property $d(a,c)+d(b,d) < \delta(\epsilon)$ implies $|d(a,b) - d(c,d)| < \epsilon$.

The *easiest way* to satisfy this requirement is if $|d(a,b) - d(c,d)| \leq d(a,c) + d(b,d)$, since then $\delta(\epsilon)$ may be taken to be equal to ϵ. But if this holds, then consideration of the pair $(a,b),(c,c)$ shows that this implies the triangle inequality, $d(a,b) \leq d(a,c) + d(c,b)$.

On the other hand, if the triangle inequality holds, then

$$|d(a,b) - d(c,d)| \leq |d(a,b) - d(b,c)| + |d(b,c) - d(c,d)| \leq d(a,c) + d(b,d),$$

where the first inequality arises from the triangle inequality of the modulus function, $|a+b| \leq |a| + |b|$.

Note that uniform continuity of a semimetric does not imply the triangle inequality in general. □

Example 3 (The n-dimensional Euclidean space E_n)**.** The points of E_n are all ordered n-tuples $(x_1, x_2, \ldots, x_n)$ of real numbers. The distance is defined for each pair of elements $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by

$$d(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}.$$

The triangle inequality follows from the Cauchy-Schwartz inequality. □

Example 4 (The n-dimensional spherical space S_n)**.** The points of S_n are all ordered $(n+1)$-tuples $x = (x_1, x_2, \ldots, x_{n+1})$ with $||x||^2 = \sum_{i=1}^{n+1} x_i^2 = 1$. Spherical distance is defined for each pair of elements x, y to be the smallest nonnegative number $d(x,y)$ such that

$$\cos(d(x,y)) = \sum_{i=1}^{n+1} x_i y_i.$$

This is an example of a geodesic (shortest-arc) distance. □

Example 5 (The Hausdorff metric)**.** A metric space M is *bounded* provided there exists a constant $K > 0$ such that $d(x, y) < K$ for all elements $x, y \in M$. Let X be the set of all closed, non-empty subsets of a bounded metric space M. Define

$$d(A, B) = \sup_{a \in A} \left(\inf_{b \in B} d(a, b) \right).$$

The function d is not a metric distance since it is not symmetric in general. Moreover, $d(A, B) = 0$ implies $\inf_{b \in B} d(a, b) = 0$ for all $a \in A$, such that $a \in \operatorname{cl} B = B$. Thus $B \subseteq A$, but in general $d(A, B) = 0$ does not imply that $A = B$. Both these shortcomings are fixed by symmetrizing d, and the resulting metric is called the Hausdorff metric:

$$d_H(A, B) = \max[d(A, B), d(B, A)].$$

To prove the triangle inequality, note that if $d(A, B) < \rho$, then $\inf_{b \in B} d(a, b) < \rho$ for all elements $a \in A$, so there exists $a \in A, b \in B$ such that $d(a, b) < \rho$. Let now A, B, C be three distinct elements of S and put $d(A, B) = \rho$ and $d(B, C) = \sigma$. For each $\epsilon > 0$ we have that $d(B, C) < \sigma + \epsilon$, and there exists $b \in B, c \in C$ such that $d(b, c) < \sigma + \epsilon$. Analogously, from $d(A, B) < \rho + \epsilon$ there exists $a \in A$ such that $d(a, b) < \rho + \epsilon$. Since M is metric,

$$d(a, c) \leq d(a, b) + d(b, c) < \rho + \sigma + 2\epsilon.$$

From this it follows that $d(A, C) \leq \rho + \sigma = d(A, B) + d(B, C)$. Similarly, it follows that $d(C, A) \leq d(C, B) + d(B, A)$. Together, the two relations

$$d_H(A, B) + d_H(B, C) \geq d(A, C)$$
$$d_H(A, B) + d_H(B, C) \geq d(C, A)$$

imply that $d_H(A, B) + d_H(B, C) \geq d_H(A, C)$, i.e., the triangle inequality in S. □

A.1.2 Congruence and embeddability

Topology was originally defined as the study of invariants of homeomorphisms, i.e., continuous functions with continuous inverses. Since homeomorphisms form a group, topology fits the definition of a geometry in the way of Felix Klein, as the study of invariants under a selected group of transformations.

The subgroup of homeomorphisms for which the distance of two points is an invariant is the group of *congruences*, and the resulting geometry is referred to as *distance geometry* (or metric topology).

Definition 10. If $p, q \in S$ and $p', q' \in S'$ for two metric spaces S, S' (with distances d, d'), then p, q are *congruent* to p', q' if and only if $d(p, q) = d'(p', q')$. Two subsets

P, Q of the same or different metric spaces are congruent provided there exists a map $f : P \to Q$ such that each pair of points from P is mapped onto a congruent point-pair of Q.

The relation of congruence is symmetric, reflexive and transitive, and therefore constitutes an equivalence relation.

We now consider the *subset problem*: What are necessary and sufficient conditions that an arbitrary distance space must satisfy in order that it may be congruent with a subset of a member of a prescribed class of spaces? In particular we will be interested in isometric embeddings of a finite set of points into Euclidean space E_n.

Definition 11. A set S is *congruently embeddable* (embeddable, for short) in a semimetric space T if S is congruent to a subset of T. A set S is *irreducibly embeddable* in E_n if it is embeddable in E_n, but not in any nontrivial subspace.

Definition 12. The *Gram matrix* of a set of vectors $\{x_i \mid 0 \leq i \leq N\}$ from an inner-product space is the matrix G of inner-products $G_{ij} = \langle x_i, x_j \rangle$. The *metric matrix* of a finite set of N points from a semimetric space, with respect to a reference point (indexed as the 0-th point), is the $(N \times N)$ matrix M with entries

$$M_{ij} = \frac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2), \tag{A.2}$$

where $d_{ij} = d(x_i, x_j)$ is the value of the semimetric for the points indexed by i and j.

In Euclidean space, as a consequence of the *law of cosines*

$$d(x_i, x_j)^2 = d_(x_0, x_i)^2 + d_(x_0, x_j)^2 - 2\langle x_i, x_j \rangle \tag{A.3}$$

in the plane containing each triple x_0, x_i, x_j of points, the metric matrix corresponds to the matrix of scalar products relative to the reference point x_0, with entries $M_{ij} = \langle x_i - x_0, x_j - x_0 \rangle$. It is also clear that the Gram matrix is positive semidefinite; in fact, that each positive semidefinite matrix can be realized as the Gram matrix of a set of vectors. This characterization carries over to the metric matrix, which solves the subset problem for Euclidean spaces:

Theorem 4 (Havel et al. (1983))**.** A configuration of $N+1$ points in a semimetric space is irreducibly embeddable in E_n, for some $n \leq N$, if and only if the corresponding metric matrix from any point is positive semidefinite of rank n. The eigenvalues of this matrix are then the (second) moments of the distribution of points along the n principal coordinate axes, and the eigenvectors, scaled by the square-roots of the corresponding eigenvalues, are the principal coordinate axes of the Euclidean configuration.

Proof. If the points are irreducibly embeddable in E_n, let $(x_0, \ldots, x_N)$ (where $x_i \in E_n$) be any family of vectors that represent them. The vectors x_i are then necessarily linearly independent. The metric matrix (with respect to the 0-th point, without loss of generality) is equal to the Gram matrix of the family $(x_1 - x_0, \ldots, x_N - x_0)$ in E_n, thus positive semidefinite and of rank n (since linear independence does not change under translation). The statement about the principal axes and the eigenvalues follows from the well-known identification of covariances with scalar products (Rodgers and Nicewander, 1988), such that the eigendecomposition of the Gram matrix defines the principal axes.

Conversely, if the $(N \times N)$ metric matrix M (with respect to the 0-th point, without loss of generality) is positive semidefinite of rank n, it can be diagonalized by an orthogonal transformation Y:

$$\Lambda = Y^t M Y. \tag{A.4}$$

The matrix Λ contains n positive eigenvalues and $N - n$ zeros on the diagonal (ordered by decreasing size, without loss of generality), and scaling the eigenvectors by their roots, a matrix $X = \Lambda^{1/2} Y$ is obtained such that $M = X^t X$. The columns of X are the coordinates of the N original points in E_n, centered on the 0-th point (at the origin). It is clear that the eigenvectors define the principal axes of X. □

This theorem solves the embedding problem for a finite set of points. The reference point is identified with the origin of the Euclidean space, and the coordinates of the points are uniquely reconstructed up to symmetries of the eigenspaces (reflections for eigenvalues with multiplicity one, subspace rotations for eigenvalues with larger multiplicities). In practice, these remaining degrees of freedom are fixed by the details of the numerical method used to diagonalize the metric matrix. It is also customary to choose the center of mass as reference point. A simple calculation shows how to obtain the corresponding metric matrix.

Theorem 5 (Havel et al. (1983))**.** The distance to the center of mass of each point i of a configuration of N points in a Euclidean space is given in terms of the remaining distances by

$$d_{0i}^2 = \frac{1}{N} \sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N^2} \sum_{k>j}^{N} d_{jk}^2. \tag{A.5}$$

Let $1_N = (1, 1, \ldots, 1)^t$ be the $(N \times 1)$-vector consisting of ones. Define the *centering operator* $J = I - \frac{1}{N} 1_N 1_N^t$. A short calculation shows that the corresponding metric matrix is obtained by its action on the matrix of squared distances D^2 (with entries $D_{ij}^2 = d_{ij}^2$) of a given family of N points,

$$M = -\frac{1}{2} J D^2 J^t. \tag{A.6}$$

This operation is usually called *double-centering*. In Section A.2 it will be used to derive representations of reduced dimensionality $n \ll N$ from a given set of distances between N points.

For completeness, we end this section with an important result that characterizes embeddability of a space in terms of finite subsets.

Definition 13. A semimetric space T has *congruence order* k with respect to a class $\mathcal{S}$ of spaces provided each space $S \in \mathcal{S}$ is embeddable in T whenever any k-subset $\{x_0, \ldots, x_{k-1} \mid x_i \in S\}$ has that property.

Theorem 6 (Havel et al. (1983))**.** The Euclidean space E_n has congruence order $n+3$ with respect to the class of all semimetric spaces.

In fact, an even stronger property holds:

Theorem 7. A semimetric space S is irreducibly embeddable in E_n if S contains a $(n+1)$-set of points irreducibly embeddable in E_n such that every $(n+3)$-subset of S containing it is embeddable in E_n.

A.2 Multidimensional scaling

The previous section discussed when points with given distances can be realized by an embedding in some Euclidean space. In practice, we are rarely presented with this ideal situation and distances are usually contaminated by noise and discretized, and we cannot expect to find zero eigenvalues numerically. Moreover, it is often *a priori* unclear whether a set of measured distances admits a Euclidean representation at all. If this were impossible, negative eigenvalues will occur in the diagonalization of the metric matrix. Since these can also arise by numerical instabilities and errors in the distances, it can be difficult to decide whether a Euclidean representation is warranted.

The techniques of *multidimensional scaling* therefore focus on the reduction of dimension, and diagnostic measures are used to quantify the goodness of reconstruction.

Similar to principal component analysis, the reduction of dimension is achieved by restricting to the first $n \leq N$ principal axes in Theorem 4. We need to distinguish between the distances actually measured between all N systems, represented by a $(N \times N)$ matrix of squared distances D^2, and the Euclidean distances of a point configuration reconstructed to represent them, represented by a $(N \times N)$ matrix of squared distances Δ^2. Recall that the Frobenius norm of a matrix A is the root sum-of-squares,

$$||A|| = \left(\sum_{ij} |A_{ij}|^2 \right)^{1/2} . \tag{A.7}$$

Box 11. Why reconstruct distances in Euclidean space?

The alternative would be to consider reconstructions in more general metric spaces, e.g., spaces endowed with a Minkowski norm, or to consider nonmetric reconstructions, where the order relations between the distances are preserved as much as possible. In fact, there are good reasons why we only consider reconstructions of points in Euclidean space here:

- The connection between Euclidean norm and scalar products:
  Since Euclidean norm is a quadratic form, we can transform distances into scalar products. These we can consider values of a kernel function, and pattern analysis by kernel methods becomes possible.
- The efficiency of metric multidimensional scaling:
  Metric solutions are easy to calculate by linear algebra.
- The intuitiveness of Euclidean space:
  Euclidean space is simply the space with which we are most familiar with.

Of course, Euclidean distance has additional beneficial properties, e.g., invariance under rotations.

It induces a distance $d(A, B) = ||A - B||$ between two matrices.

Definition 14. The (*raw*) *stress* of a reconstructed configuration is

$$\sigma_{\mathrm{r}}(D^2, \Delta^2) = \frac{1}{2}||D^2 - \Delta^2||^2 = \frac{1}{2}\sum_{ij}\left(D_{ij}^2 - \Delta_{ij}^2\right)^2. \tag{A.8}$$

In the specific context of classical multidimensional scaling raw stress it is also known as the *strain* of a configuration.

Theorem 8 (Gower (1966),Havel et al. (1983))**.** The $(N \times N)$ symmetric matrix of rank n that best approximates any given $(N \times N)$ symmetric matrix of higher rank, in the sense of minimizing the Frobenius distance, is obtained by setting all but the n eigenvalues of largest magnitude to zero (and transforming back).

Recall the eigendecomposition $Y\Lambda Y^t = M$ (A.4), where Λ is a diagonal matrix of eigenvalues sorted by decreasing value, and Y is an orthogonal matrix whose rows contain the respective eigenvectors. Let Λ_n be the diagonal $(n \times n)$ matrix that contains only the largest $n \leq N$ eigenvalues of Λ, and Y_n be the matrix consisting of the first k columns of Y. Then the $(N \times n)$ coordinate matrix of *classical* (or *metric*) *multidimensional scaling* is given by $X_n = Y_n \Lambda_n^{1/2}$. Note that we have assumed here that the magnitude of negative eigenvalues is smaller than the magnitude of the n-th largest (positive) eigenvalue, i.e., we have assumed that errors and misrepresentations of distances are relatively small.

This representation of distances in Euclidean space minimizes the strain and leads to a nested solution: The coordinates in X_{n-1} are the same as the first $n-1$ coordinates of X_n (up to symmetries of the eigenspaces). It is called the *functional* or *behavior representation* of the distances Δ.

A.2.1 Diagnostic measures and distortions

The raw stress (A.8) has the disadvantage that it depends on the global scale of the distances. The following "badness-of-fit" measure is a scale-invariant diagnostic that quantifies the fraction of the sum-of-squares misrepresentation error that is not accounted for by the distances.

Definition 15 (Borg and Groenen (2005))**.** The *normalized stress* of a reconstructed configuration is

$$\sigma_n(D^2, \Delta^2) = \frac{\sum_{ij} \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\sum_{ij} D_{ij}^2}. \tag{A.9}$$

The value of $1-\sigma_n(D^2, \Delta^2)$ is the fraction of distances explained in the Euclidean configuration, i.e., a *coefficient of determination*. Being a global statistic, σ_n is sensitive to outliers, i.e., points with an unusually large misrepresentation error. These can be identified by assessing the local misrepresentation error, and the following two diagnostic measures accomplish this.

Definition 16. The *Shepard diagram* of a reconstructed configuration is the diagram obtained by plotting the $N(N-1)/2$ distances Δ_{ij} of the Euclidean configuration against the measured distances D_{ij}. The (*normalized*) *maximal misrepresentation error* is given by

$$\sigma_{\max} = \frac{\max_{ij} \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\frac{1}{N^2} \sum_{ij} D_{ij}^2}. \tag{A.10}$$

Definition 17. The (*normalized*) *stress per point* of the i-th point in a reconstructed configuration, consisting of N points, is given by

$$\sigma_n^i(D^2, \Delta^2) = \frac{\frac{1}{N} \sum_j \left(D_{ij}^2 - \Delta_{ij}^2\right)^2}{\sum_{ij} D_{ij}^2}. \tag{A.11}$$

Whereas the Shepard diagram visualizes the goodness-of-fit of all distances and can be useful to detect anisotropic distortions in the representation, the stress per point allows to detect suspect points or outliers that should be studied more closely. Raw stress per point, defined as in (A.11) but without the normalization in the denominator, can be conveniently visualized in a reconstructed configuration by plotting circles around each point, with area equal to the average stress of each point.

Note that the definitions have been given for symmetric distance matrices; in the case of (small) asymmetries these need to be changed accordingly.

We conclude this overview of the most important diagnostic measures with two examples.

Example 6. Figure A.1 shows three two-dimensional reconstructions of $N = 50$ points randomly distributed along the unit circle. In the left panel the configuration was obtained by classical multidimensional scaling when the distance matrix was calculated from Euclidean distances. Circles were used to depict the values of raw stress per point. The reconstruction is almost perfect, with misrepresentation errors on the order of the numerical accuracy, i.e., with $\sigma_{\max} \approx 10^{-34}$. This is reflected in the Shepard diagram (left panel of Figure A.2), which shows an almost diagonal line.

When the distance matrix is calculated from geodetic distances (Example 4), misrepresentation errors are introduced. The corresponding configuration is shown in the middle panel of Figure A.1. Stress per point is distributed relatively evenly among all points, with the largest errors accruing where the least points were present, and accounts for about 2 percent of the sum-of-square error ($\sigma_n \approx 0.02$). The Shepard diagram (middle panel of Figure A.2) shows that most distances are slightly overrepresented, whereas a few of the largest distances are underestimated. Note that both eigenvalues were positive (not shown). Changing the reconstruction dimension does also not allow for much leeway in improving the reconstruction. In one dimension the misrepresentation error is very large ($\sigma_n \approx 0.30$), whereas for larger dimensions it is also slightly larger than in two dimensions (left panel of Figure A.3, solid curve). For dimensions above about $N/2$, the first negative eigenvalue is encountered.

The right panels of Figure A.1 and Figure A.2 show results for a reconstruction from Euclidean distances that were contaminated with noise (normal, with unit variance). The misrepresentation error is again distributed relatively evenly, but the shape of the configuration has seriously deteriorated due to the large amount of noise. Its influence can be seen in the Shepard diagram, which shows that errors in the distances are distributed randomly. The dependence on reconstruction dimension (Figure A.3, grey curve) is not qualitatively changed, only shifted to larger errors.

Example 7. A different example is provided by the reconstruction of a torus, shown in Figure A.4. Since the line element of the standard torus, embedded as a two-dimensional surface in three dimensions, can only be evaluated numerically, we resort to the simpler representation of the torus as the quotient of the plane under the identification $(x, y) \sim (x + 2\pi, y) \sim (x, y + 2\pi)$. The torus $T^2 \simeq S^1 \times S^1$ is then identified by a square with opposite boundaries identified. This is a natural representation

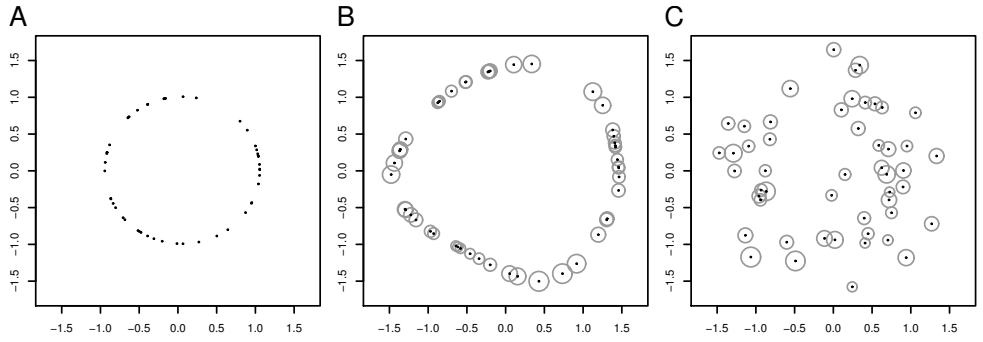


Figure A.1: Reconstruction of the one-dimensional circle S^1 by classical multidimensional scaling from $N = 50$ random samples. A: S^1 with the induced Euclidean metric. B: S^1 with its intrinsic, geodesic metric. C: S^1 with the induced metric, but Gaussian noise (unit variance) added to the distances. Radius of circles indicates (raw) stress-per-point.

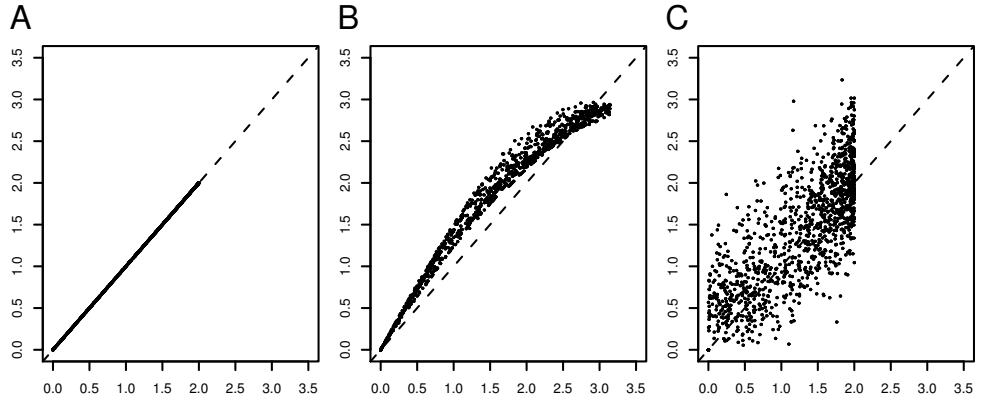


Figure A.2: Shepard diagrams for the reconstructions in Fig. A.1, depicting distances in the reconstructed configuration (vertical axis) against original distances (horizontal axis). A: S^1 with induced Euclidean metric. B: S^1 with intrinsic, geodesic metric. C: S^1 with induced metric under Gaussian noise.

for two simultaneously measured phases, with geodetic distance

$$\left((\min(|y_1 - x_1|, 2\pi - |y_1 - x_1|))^2 + (\min(|y_2 - x_2|, 2\pi - |y_2 - x_2|))^2 \right)^{1/2} \quad \text{(A.12)}$$

between two points $(x_1, x_2), (y_1, y_2) \in T^2$. The left panel of Figure A.4 shows the reconstructed configuration for Euclidean distances, the middle panel the configuration for the geodesic distance, and the right panel was obtained for Euclidean

distances under random noise (normal, unit variance).

The two-dimensional configuration of the geodesic distances approximates a square, with points in its interior exhibiting the largest misrepresentation error. Globally, about 15 percent of the distances cannot be accounted for in this representation ($\sigma_n \approx 0.15$), which drops to a mere 2 percent if the samples are reconstructed in four dimensions. The systemic distortions in the two-dimensional case can be clearly seen in the Shepard diagram (middle panel of Figure A.5), whereas a four-dimensional reconstruction closely approaches the original distances (right panel). The right panel of Figure A.3 shows the normalized stress against the reconstruction dimension (solid curve). The minimal stress is achieved for about four dimensions, and then rises again slightly due to numerical errors.

Reconstruction from the Euclidean distances under noise leads to similar changes as in Example 6. The misrepresentation error shows the same qualitative behavior with respect to the reconstruction dimensionality, only shifted to a higher level (right panel in Figure A.3, grey curve).

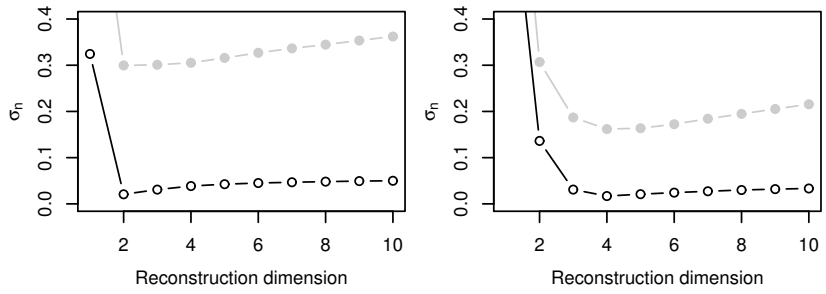


Figure A.3: Normalized stress for different reconstruction dimensions for distances without (dark) and under Gaussian noise (gray). A: S^1 with intrinsic, geodesic metric. B: T^2 with intrinsic, geodesic metric. The configurations were reconstructed from $N = 50$ random points each.

These examples show that stress diagrams as in Figure A.3 can be used to decide which dimension is optimal for the reconstruction from a given distance matrix, and whether misrepresentation errors might be caused by random noise or by systematic distortions due to an intrinsically different geometry. Whereas the effects of noise cannot be reduced by increasing the reconstruction dimension, this is possible (to a great extent) for non-Euclidean distances.

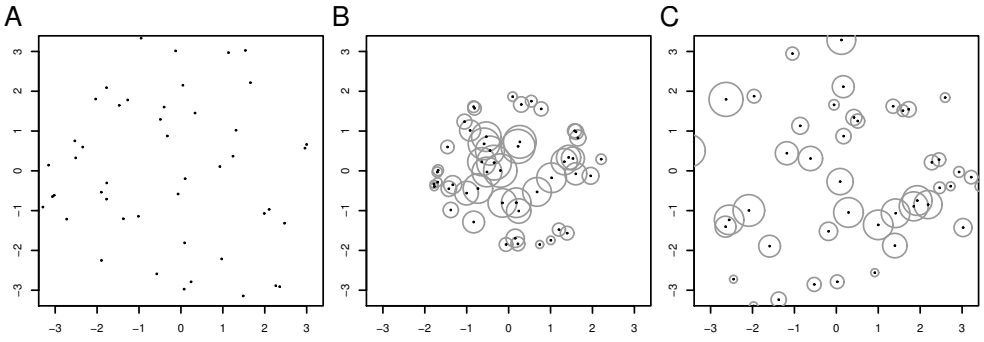


Figure A.4: Reconstruction of the two-dimensional torus T^2 by classical multidimensional scaling from $N = 50$ random samples. A: T^2 with the induced Euclidean metric. B: T^2 with its intrinsic, geodesic metric. C: T^2 with induced metric, but Gaussian noise (unit variance) added to the distances. Radius of circles indicates (raw) stress-per-point.

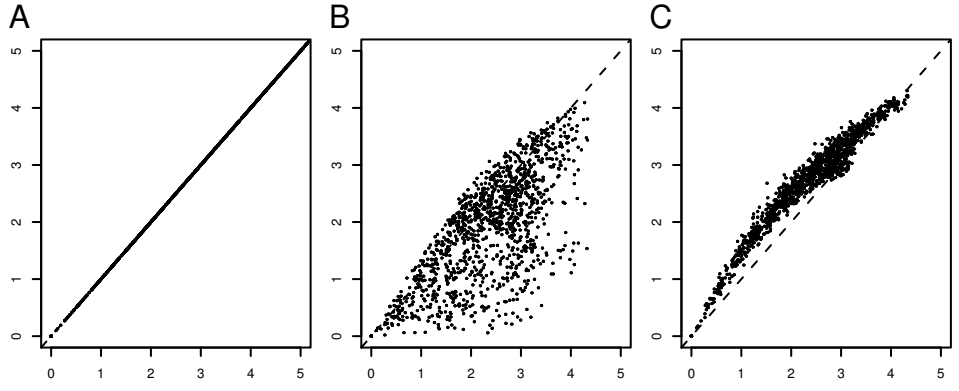


Figure A.5: Shepard diagrams for reconstruction of the torus T^2. A: T^2 with the Euclidean metric in 2D. B: T^2 with its intrinsic metric in 2D. C: T^2 with its intrinsic metric in 4D.

A.2.2 Violations of metric properties and bootstrapping

Different from the effect of non-Euclidean geometries, the influence of noise in measured distances can and usually does destroy the metric properties, for sufficiently large noise levels. Such violations of metric properties are also interesting conceptually. Many bivariate measures commonly used (e.g., in electrophysiology, see Section 6.3) do not respect metric properties, and it is instructive to consider what effect this does have on dissimilarity matrices, where we use the word *dissimilarity* to denote a bivariate measure that does not necessarily fulfill metric properties. Moreover,

these violations occur when the distances are resampled (see below) to reduce bias in measurements, or to improve the speed of computations.

Reflexivity

Reflexivity is the property that the *self-distances* $d(x, x)$ are zero. Conceptually, this embodies the notion of identity, and measures that do not result in reflexive dissimilarities are problematic. The reason is, of course, that such dissimilarities cannot be interpreted in terms of points, but would need to be realized as extended objects — if this is consistently possible at all. Therefore, reflexivity should not be destroyed by even the effect of measurement noise, but since the numerical calculation of distances can introduce round-off errors, reflexivity can be violated in practice. The usual solution is to simply force the diagonal terms of dissimilarity matrices to zero, but there is a situation in which self-distances naturally occur and contain valuable information.

Up to now we have been assuming implicitly that measurements made on systems are ideal, in the sense that the system's behavior is captured in its totality. In practice this is barely the case, however, since measurements are finite and should always be considered approximations of a system. If we consider generalized measurements (Section 1.1) that result in probability measures, these measures are empirical and might differ from the true probability measure that would be obtained under ideal conditions. The variability inherent in these probability measures can be estimated, however, by bootstrapping the empirical measures. Thereby, a random sample (with replacement) is drawn from the measure under consideration, usually of the same size as the original observations on which that measure is based, and interpreted as another (empirical) probability measure. Repeating this process a number of times, a set of probability measures is obtained that represent the variability of the unknown, underlying probability measure. Although this is not an unbiased method, since it takes an empirical realization of a probability measure as its point of departure, such *bootstrapping* obtains an approximation of the original measure that is valid to a great extent, i.e., with largely reduced statistical error (Efron, 1981; Davison and Hinkley, 1997).

A second advantage of resampling the measurements is that one can choose a *smaller* sample size. Although this invariably increases the variance, and a larger number of bootstrap replications is needed to obtain the same reduction in bias, it may speed up computations enormously. We will therefore routinely use this device for the involved calculations of the optimal transportation distances (confer Sections 2.5,3.6.3,6.4). In practice, it will result in not a single distance between two systems, but rather in a set of bootstrap replicates of numerical distances. We will then take the mean of these as an estimate of the "true" distance between two systems.

A special case occurs with the self-distances $d(x, x)$, however, since distances can only be nonnegative. The magnitude of the self-distances under resampling is there-

fore an indication of the *numerical resolution* of our distance measure. Systems that are closer than the average self-distance cannot be resolved properly and appear to be distinct in actual calculations, and distances between two distinct systems should be considered to be influenced by statistical errors of the same order. This state of affairs can also not be remedied by subtracting a constant from all distances, since this might destroy the triangle inequality (see below). It is important to keep this qualification in mind.

Symmetry

Violations of symmetry, where $d(x, y) \neq d(y, x)$ can arise by noise or resampling error (see above), but might also indicate directionality effects. These again lead to representations of systems as extended objects (confer Figure 6.1 in Chapter 6), which is undesirable for further analysis. In the first case, the accepted method is to simply average out the asymmetries. Given a dissimilarity matrix D, it can be decomposed into a symmetric part $S = \frac{1}{2}(D + D^t)$ and an antisymmetric part $A = \frac{1}{2}(D - D^t)$, such that

$$D = A + S. \tag{A.13}$$

The symmetric part S is then used as an estimate of the underlying true distances. However, if the antisymmetric part A is not of negligible size relative to S, this hints at the influence of directionality. General dissimilarity measures (see Section 6.3 for examples) might measure the flow of information between two systems, or the strength of influence one system exerts upon another, which are genuinely asymmetric effects. Due to the decomposition (A.13), however, it is possible to treat the symmetric and antisymmetric part independently. This problem is therefore alleviated to a great extent. Treatment of the antisymmetric part is further discussed in Section 6.2.2, for completeness.

Triangle inequality

The triangle inequality is basic to a representation in Euclidean space. As before, violations of this property hint at directionality effects and suggest that systems might need to be represented by extended objects (Figure 6.1). It is the most common violation for many dissimilarity measures, since reflexivity and symmetry are often easy to accomplish, whereas the triangle inequality is a nontrivial geometric constraint. Violations of the triangle inequality are therefore important conceptually, since they suggest that a geometric representation might be unsuitable. If the triangle inequality is not fulfilled, it is not possible to compare more than two systems in a sensible (multivariate) way without introducing additional, spurious effects that are undesirable. However, adding a constant $c > 0$ to all distances (from a finite set), the triangle

Box 12. How to publish distance matrices?
When publishing research results obtained from or with (measured) distance matrices, the following information should ideally be also given:

- Was the triangle inequality fulfilled? If not, how large was the maximal violation?
- Were all eigenvalues nonnegative? If not, how large was the negative eigenvalue of largest magnitude? How many positive and negative eigenvalues were there?
- Were all diagonal entries zero? If not, how large was the largest diagonal element?

inequality can always be enforced, since for large enough c the equation

$$d(x, y) \leq d(x, z) + d(z, y) + c, \tag{A.14}$$

will be fulfilled.

Bootrapping the distances can break the triangle inequality, and to ensure multivariate comparability we will use the smallest possible constant in (A.14) to fix this, if needed. Of course such violations of the triangle inequality need to be reported.

A.3 Statistical inference

Although multidimensional scaling has a long history, statistical inference about reconstructed point configurations is seldomly encountered in the literature (but see (Anderson and Robinson, 2003)). In this section we will therefore advocate and describe the main methods of statistical analysis used in the rest of this thesis. The starting point for most methods considered here is the reconstructed point configuration of N systems, i.e., their representation as N vectors in a Euclidean space E_n, where $n \leq N$. We call this the *behavior* or *functional space* of the systems. This representation allows for the use of multivariate analysis methods. We are particularly interested in the task of classifying distinct groups of systems. More precisely, we will consider *supervised* classification, in which the true group assignments of all points are assumed to be known perfectly. Let there be $g \in \mathbb{N}$ distinct groups $G_1, \ldots, G_g$, and let the true group label of a point $x \in E_n$ be given by an indicator variable $z = (z_1, \ldots, z_g)$, such that $z_i = 1$ if $x \in G_i$ and $z_i = 0$ if $x \notin G_i$. Let $(x_1, \ldots, x_N)$ denote the points from E_n representing the N systems under study. Denote by $\lambda = (\lambda_1, \ldots, \lambda_N)$ the labelling, such that $g_i = k$ if and only if $z_k = 1$ for the point x_i.

A.3.1 Multiple response permutation testing

The first question about the representation $(x_1, \ldots, x_N)$ of N systems from a priori known g groups is whether this representation does carry information on the group structure, and to what extent.

To assess this, we employ a permutation hypothesis test. Under the null hypothesis of no difference with regard to group association, the labelling λ can be permuted randomly. As a test statistic, we will use the weighted mean of within-group means of pairwise distances among groups. Let $(N_1, \ldots, N_g)$ be the sizes of the g groups, then this is given by

$$\delta_\lambda = \sum_{k=1}^{g} \frac{N_k / \sum_l N_l}{N_k(N_k-1)/2} \sum_{\substack{i<j \\ \lambda_i=\lambda_j=k}} D_{ij}, \tag{A.15}$$

conditional on the group labelling λ and the pairwise distances D_{ij}. Under the null hypothesis the test statistic δ will be invariant under permutations $\pi\lambda$ of the group labelling, and the significance probability of this test is the fraction of values of $\delta_{\pi\lambda}$ obtained that are smaller than the value δ_λ for the original labelling λ:

$$p = \frac{\#\{\delta_{\pi\lambda} < \delta_\lambda\}}{m+1}, \tag{A.16}$$

where m is the number of permutations. Considering all distinct $\binom{N!}{N_1!N_2!\cdots N_g!}$ permutations will be often infeasible, so the value of p is estimated by considering a large enough number of random permutations (typically on the order of 10^5 or larger). This test is called a *multiple response permutation procedure* (MRPP).

Similar as in analysis of variance, the *chance-corrected within-group agreement*

$$A = 1 - \frac{\delta_\lambda}{\mathbb{E}\delta_{\pi\lambda}}, \tag{A.17}$$

where $\mathbb{E}\delta_{\pi\lambda}$ is approximated by the mean of δ under all permutations π considered, is a coefficient of determination that quantifies how much of the group structure is "explained" by the distances.

It is important to stress the difference between these two diagnostic measures. Whereas a small p-value indicates that the structure of the distances is significantly dependent on the group association, it might still be the case (and often will be in practice) that the size of this effect, as measured by A, is rather small. To this extent, the value of A indicates the signal-to-noise ratio of the distances.

The MRPP test is calculated from the distance information only, and can therefore be performed for both the original distance matrix D, and additionally for the distance matrix Δ of the reconstructed points $(x_1, \ldots, x_N)$ that is subject to misrep-

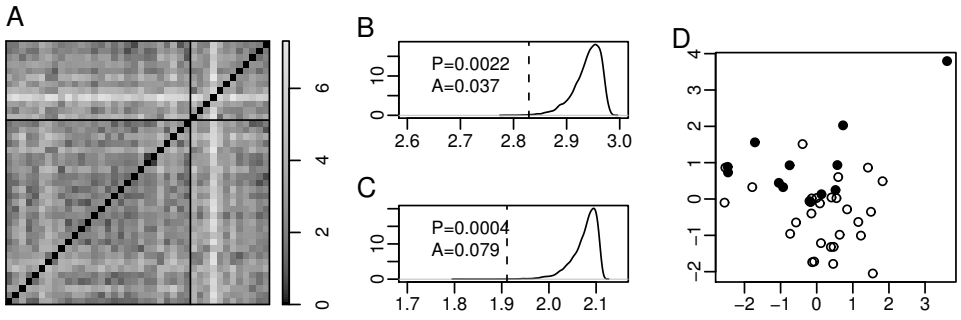


Figure A.6: Example: MRPP test in the Pima dataset. A: Distance matrix between all subjects ($N_1 = 28$ with no diabetes, $N_2 = 12$ with diabetes; separated by dark lines). B: Distribution of MRPP statistic δ for the distances in (A). C: Distribution of δ for the distances of the reconstructed two-dimensional configuration. D: Reconstructed configuration (dark circles: diabetes, open circles: no diabetes).

resentation errors. Both of these tests are of value in practice. The first shows the extent to which measured distances capture the group structure, the second shows how much of this is still preserved under reconstruction. Thereby, it can be judged whether a Euclidean representation is adequate.

Example 8. Let us illustrate the MRPP test with the Pima dataset from the R package MASS (Venables and Ripley, 1999). This dataset contains information collected by the US National Institute of Diabetes and Digestive and Kidney Diseases on diabetes in women of Pima Indian heritage. We will use five variables from the first 40 entries of the training data Pima.tr: plasma glucose concentration, blood pressure, body-mass-index, diabetes pedigree function, and age. The outcome (diabetes or not) is known, with $N_1 = 28$ subjects showing no symptoms of diabetes, and $N_2 = 12$ being diagnosed with diabetes. Distances between subjects were calculated by first centering and scaling the predictor variables to unit variance, and then taking Euclidean distance in the five-dimensional space. Figure A.6 shows the distance matrix, the reconstructed point configuration in two-dimensions, and the distribution of the MRPP statistic δ for both sets of distances. Interestingly, the within-group agreement A of the reconstructed configuration is twice as large as for the original distances, indicating that dimension reduction can improve the classificatory contrast.

A.3.2 Discriminant analysis

In the distance-based approach, discrimination of systems is achieved from their representation in Euclidean space E_n. We advocate the use of robust and conceptually

simple analysis methods, and have therefore chosen *canonical discriminant analysis* as our method of choice for the classification of systems. Canonical discriminant functions are *linear* combinations of variables that best separate the mean vectors of two or more groups of multivariate observations relative to the within-group variance. They are variously known as canonical variates or discriminant coordinates in the literature and generalize the linear discriminant analysis of Fisher (1936) (for the case of $g = 2$ groups). For this reason, the term *linear discriminant analysis* (LDA) is also used for the analysis described here.

Let B_0 be the covariance matrix of the group-wise distributions,

$$B_0 = \frac{1}{g-1} \sum_{i=1}^{g} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^t, \tag{A.18}$$

where $\bar{\mu} = \frac{1}{g} \sum_{i=1}^{g} \mu_i$ is the pooled mean[1]. In practice this will be approximated by the sample between-groups covariance matrix on $g - 1$ degrees of freedoms,

$$B = \frac{1}{g-1} \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^t, \tag{A.19}$$

where $\bar{x}_i$ is the sample mean of the i-th group, and $\bar{x} = \frac{1}{g} \sum_{i=1}^{g} \bar{x}_i = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the mean over the whole sample. The matrix B_0 (and therefore also B) is of rank $b_0 \leq g - 1$, where $b_0 = g - 1$ if and only if the group-means $\mu_1, \ldots, \mu_g$ are linearly independent.

Let Σ_0 be the within-group covariance matrix. The main assumption here is that this is equal for each group (*homoscedasticity* assumption), and it is estimated by the pooled within-group sample covariance matrix. Let $X = [x_1 \cdots x_N]^t$ be the $(N \times n)$ matrix of coordinates and let M be the $(g \times n)$ matrix of group means. Define the $(N \times g)$ matrix of group indicators Z by $Z_{ij} = 1$ if $x_i \in G_i$, and $Z_{ij} = 0$ otherwise. Then

$$\Sigma = \frac{1}{N-g}(X - ZM)^t(X - ZM) \quad \text{and} \quad B = \frac{1}{g-1}(ZM - 1_N \bar{x}^t)^t(ZM - 1_N \bar{x}^t) \tag{A.20}$$

are the two sample covariance matrices in matrix notation.

There exist $r = \min(n, g - 1)$ canonical variates (discriminant "scores"), and for the coordinates in X these are defined by

$$S = XA, \tag{A.21}$$

where $A = [a_1 \cdots a_r]$ is a $(n \times r)$ *scaling* matrix, such that a_1 maximizes the ratio

[1] Using coordinates derived from double centering clearly $\bar{\mu} = 0$, but we prefer to exhibit the general case here.

(generalized Rayleigh quotient)

$$\frac{a_1^t B a_1}{a_1^t \Sigma a_1}. \tag{A.22}$$

The scaling acts on the right, since the coordinates X are in row-order. For $k = 2, \ldots, r$, the variate a_k maximizes the ratio (A.22) subject to the orthogonality constraint $a_k^t \Sigma_0 a_h = 0$ (for $h = 1, \ldots, k-1$). To compute A, choose a preliminary scaling XA_1 of the variables such that they have the identity as their within-group correlation matrix. This is achieved by taking the principal components with respect to Σ, normalized by their variance. On the rescaled variables XA_1, the maximization of (A.22) reduces to the maximization of $a^T B a$ under the constraint $||a|| = 1$. The latter is solved by taking a to be the (normalized) eigenvector of B corresponding to the largest eigenvalue. The eigenvectors corresponding to the next $g - 2$ largest eigenvalues supply the other $g-2$ canonical variates, which are orthogonal as required. In practive we use the lda function in the standard R package MASS (Venables and Ripley, 1999, Ch. 11.4), which employs singular value decomposition (SVD) to find the eigenvectors. Note that this code, as is standard in multivariate analysis, rescales the different coordinates in the reconstruction space E_n to unit variance prior to calculation of the canonical variates. This is one reason why cross-validation (Section A.3.3) is so important: This standardization allows coordinates which contribute very little to the distances (between systems) to influence the discrimination on equal terms with coordinates that contribute much more to the distances. For small sample sizes N the discrimination could then be based on fitting the "noise" in the distances, rather than the "signal".

The allocation of a vector x to a group G_i can be achieved in a number of ways. The simplest way is to choose the group to which the point x has smallest distance. However, this distance should consider the statistical properties of the underlying group conditional distribution, i.e., its spread around its center point. It is therefore common to measure the distance between a vector x and the i-th group, with mean μ_i and covariance matrix Σ, by their Mahalanobis distance,

$$\left((x - \mu_i)^t \Sigma^{-1} (x - \mu_i)\right)^{1/2}. \tag{A.23}$$

If we *assume* that the distribution of the i-th class is multivariate normal with mean μ_i and covariance matrix Σ, then this corresponds to *maximum a posteriori* classification, up to the prior distribution. In detail, the Bayes rule that minimizes the overall misclassification error (under equal misallocation costs) is given by

$$r(x) = i \quad \text{if} \quad \pi_i f_i(x) \geq \pi_j f_j(x) \quad (j = 1, \ldots, g; j \neq i), \tag{A.24}$$

where $\pi = (\pi_1, \ldots, \pi_g)$ is the prior distribution of groups and f_i is the group-conditional probability density of the i-th group. The prior distribution is in practice ap-

proximated by the relative group sizes, and $f_i(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i))$. It is more convenient to work in terms of the log-likelihood, which is given by

$$L_i = -\frac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu_i) + \log|\Sigma| + \log\pi_i. \tag{A.25}$$

Subtracting the constant terms, this simplifies to the maximalization of

$$L_i = x^t\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \log\pi_i. \tag{A.26}$$

In the coordinate system defined by the canonical covariates, the within-group variance is trivial, such that on these variables the Mahalanobis distance is just $||x-\mu_i||$. The log-likelihood further simplifies to

$$L_i = x^t\mu_i - \frac{1}{2}||\mu_i||^2 + \log\pi_i. \tag{A.27}$$

The a posteriori probabilities of group membership are then given by

$$\frac{\exp(-(x^t\mu_i - \min_j x^t\mu_j))}{\sum_k \exp(-(x^t\mu_k - \min_j x^t\mu_j))}. \tag{A.28}$$

Let us summarize. The canonical variates are defined in terms of seoond-order statistical properties (means and covariances) between and within groups of (normalized) coordinates. The main assumption is that the covariances for each group are equal (homoscedasticity assumption). In particular, it is not needed to assume that the group conditional distributions are multivariate normal. Under this assumption, however, the allocation rule (A.24) is optimal, if the total misallocation error is to be minimized. The reasons we routinely employ this normal based classification are summarized in Box 13.

A.3.3 Cross-validation and diagnostic measures in classification

For small to medium sized datasets encountered here, cross-validation of classification accuracies is achieved by a leave-one-out method. This proceeds in the following steps:

1. For the k-th sample point ($k = 1, \ldots, N$) we remove its distance information from the set of original distances D_{ij}, leading to a new $(N-1)$-by-$(N-1)$ matrix of squared distances $D^2_{(k)}$.

2. We reconstruct a Euclidean configuration $X_n^{(k)}$ in n dimensions by metric multidimensional scaling of $D^2_{(k)}$.

Box 13. Why use the homoscedastic normal-based allocation rule?

- Why parametric classification: Although non-parametric alternatives exist, these are much more involved and cannot be routinely used for small sample sizes.
- Why homoscedasticity: Estimation of multivariate covariance matrices is difficult for small sample sizes. The assumption of homoscedasticity allows to only estimate one covariance matrix in place of many, thereby improving the stability of the estimate.
- Why normal based allocation: The multivariate normal model is flexible and computationally efficient, and it is relatively robust. Even if the true distribution is not normal, its approximation by a normal distribution (second-order approximation) is often close, if the distribution has finite second moments and is not too irregular otherwise.

3. We train the classifier on $X_n^{(k)}$, i.e., we estimate the group means and covariance matrix from $X_n^{(k)}$.
4. We estimate the coordinates x' of the i-th sample point in the coordinate system defined by $X_n^{(k)}$ by minimizing an error criterion (Trosset and Priebe, 2008).
5. We predict the group membership of the coordinates x' by the normal-based rule. Additionally, we store the discriminant scores of x'.
6. The above is repeated for all N points. The total number of correct predictions results in the cross-validated accuracy.

The accuracy estimates obtained thereby are almost unbiased. The only parameter needed is the reconstruction dimension $n \leq N$. We will usually determine this by considering each possible choice of $1 \leq n \leq N'$ up to some maximum dimension $N' \leq N$ and choosing the dimension n' that maximizes the cross-validated classification accuracy. Note that this introduces a certain selection bias into the accuracies, but this cannot be avoided for small datasets, and should in fact be negligible.

The cross-validated discriminant scores obtained by the above method provide us with additional diagnostic information. Note however, that these scores are biased due to the different scaling invoked at each step. The main problem here is, that the geometric orientation of the discriminant functions can and will often be different for the distinct $X_n^{(k)}$. For two groups, the sign of the discriminant scores can change, but this problem can be largely avoided: Since the original group membership is known, discriminant scores with the wrong sign can be corrected. Thereby, only a slight bias occurs, as the origin of the coordinate system of the $X_n^{(k)}$ depends on the

points. The discriminant scores will therefore be slightly inaccurate and should be considered with care. As often in statistics, outliers in the data can lead to unexpected results, and it at this point where this could potentially happen.

In a classification task with two groups the classification is achieved by fixing a numerical threshold and predicting all scores to the left of it as *negatives*, and all scores to the right as *positives*. Varying the classification threshold, the number of correctly predicted negatives and positives will change. This can be conveniently visualized in a *receiver-operator-characteristic*, which allows to derive additional diagnostic measures (Hanley and McNeil, 1982).

Let TP denote the number of correctly predicted positives, let FP denote the number of incorrectly predicted positives, and likewise TN and FN for the negatives. The *true positive rate* (TPR) and the *false positive rate* (FPR) are defined by

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \text{and} \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{A.29}$$

respectively. Note that $\mathrm{TP}+\mathrm{FN}$ is the number of positives (known a priori) in the dataset, and $\mathrm{FP}+\mathrm{TN}$ the number of negatives. In the context of a diagnostic test the true positive rate TPR is interpreted as the *sensitivity*, and $1-\mathrm{FPR}$ is interpreted as the *specificity*. The receiver-operator-characteristic depicts the relationship between TPR and FPR.

Example 9. For the Pima dataset of Example 8, classification results are shown in Figure A.7. Receiver-operator characteristics of both the original data (A) and its optimal Euclidean reconstruction (D) are given. The accuracies (both resubstiution and cross-validated) for the reconstruction indicate that resubstitution accuracies tend to overestimate the classification success (B, in gray) for larger reconstruction dimensions. The cross-validated accuracies (B, in black) result in a realistic picture, never rising above the accuracy 0.85 of the original data. Interestingly, for the optimal reconstruction in two dimensions (maximal accuracy), the cross-validated accuracy is almost identical to the resubstitution accuracy, as are the receiver-operator-characteristics (D). Again, this indicates that the distance-based classification can improve classification.

A.3.4 Combining classifiers

In some cases of interest there exists more than one type of measurements of a given family of systems and we will briefly discuss two situations here: (i) if more than one distance matrix is available, and (ii) if more than one classification rule is available.

The first case can arise naturally in the framework of optimal transportation distances (Chapter B), since these distances form a parametric family. Similar to the Minkowski distances $||x-y||_p = (\sum_i |x_i - y_i|^p)^{1/p}$, different distances (for distinct values of $p \geq 1$) stress slightly different aspects of the underlying geometry.

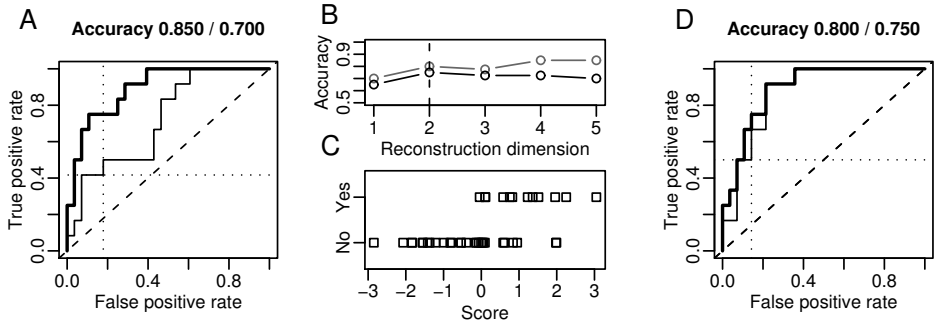


Figure A.7: Classification in the Pima dataset. A: Receiver-operator-characteristic for discriminating negatives (no diabetes) from positives (diabetes). Dark line: resubstitution accuracies. Light line: cross-validated accuracies. The optimal normal-based classification boundary is indicated (stippled lines), leading to the accuracies indicated (above plot). B: Accuracies (grey: resubstitution, dark: cross-validated) against reconstruction dimensions. C: Cross-validated discriminant scores for the optimal two-dimensional reconstruction. D: Corresponding receiver-operator-characteristic.

We encounter a different instance of this problem in Chapter 3, where two distinct time series are available for each subject. There, we will simply normalize both time series to zero mean and unit variance, and combine them into a vector-valued time series. This eventually leads to a multivariate probability distribution from which a single distance matrix is computed.

We recommend to combine distinct measurements into a single distance for practical reasons. Note that squared dissimilarities are additive in the reconstructed Euclidean space, and in the context of multidimensional scaling so-called *three-way scaling* exploits this property, allowing to weight the contributions of distinct distance matrices (Arabie et al., 1987). Since these methods are computationally involved, they will not be considered further here.

For the second situation, there exists a large literature on voting procedures that allow to combine distinct classifiers, and even optimal training rules for this meta-decision problem (Tax et al., 2000).

Appendix B

Optimal transportation distances

Science is what we understand well enough to explain to a computer. Art is verythings else we do.

Donald Knuth

In Section B.1 the general, probabilistic setting is introduced with which we work in the following. Section B.2 introduces the optimal transportation problem which is used to define a distance in Section B.3.

B.1 The setting

Recall the setting introduced in Section 1.1: A complex system S is measured by a measuring device D. The system S is an element of an abstract space of systems $\mathcal{S}$, and a measuring device is a function that maps $S \in \mathcal{S}$ into a space of measurements M. Since we are interested in quantitative measurements, the space M will be a metric space (M, d), equipped with a distance d. For example, we could take (M, d) to be some Euclidean space E_n or, more generally, a manifold with distance induced by geodesics (shortest paths). However, to account for random influences in the measurement process, we will more generally consider spaces of probability measures on M.

Let (M, d) be a metric space. For simplicity of exposition, let us also assume that M is complete, path-connected and has continuous distance function, such that it is Hausdorff in the induced topology. A *curve* on M is a continuous function $\gamma : [0, 1] \to M$. It is a curve from x to y if $\gamma(0) = x$ and $\gamma(1) = y$. The *arc length* of γ is defined by

$$L_\gamma = \sup_{0=t_0<t_1<\cdots<t_n=1} \sum_{i=0}^{n-1} d(\gamma(t_i), \gamma(t_{i+1})), \tag{B.1}$$

where the supremum is taken over all possible partitions of $[0, 1]$, for all $n \in \mathbb{N}$. Note that L_γ can be infinite; the curve γ is then called non-rectifiable.

Let us define a new metric d_I on M, by letting the value of $d_I(x, y)$ be the infimum of the lengths of all paths from x to y. This is called the *induced intrinsic metric* of M. If $d_I(x, y) = d(x, y)$ for all points $x, y \in M$, then (M, d) is a *length space* and d is called *intrinsic*. Euclidean space E_n and Riemannian manifolds are examples of

length spaces. Since M is path-connected, it is a *convex metric space*, i.e., for any two points $x, y \in M$ there exists a point $z \in M$ between x and y in the intrinsic metric.

Let μ be a probability measure on M with σ-algebra $\mathcal{B}$. We will assume μ to be a Radon measure, i.e., a tight locally-finite measure on the Borel σ-algebra of M, and denote the space of all such measures by $\mathcal{P}(M)$. Most of the time, however, we will be working in the much simpler setting of a discrete probability space: Let μ be a singular measure on M that is finitely presentable, i.e., such that there exists a representation

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \tag{B.2}$$

where δ_{x_i} is the Dirac measure at point $x_i \in M$, and the norming constraint $\sum_{i=1}^{n} a_i = 1$ is fulfilled. We further assume that $x_i \neq x_j$ if $i \neq j$, which makes the representation (B.2) unique (up to permutation of indices). Denote the space of all such measures by $\mathcal{P}_{\mathrm{F}}(M)$. Measures in $\mathcal{P}_{\mathrm{F}}$ correspond to the notion of a *weighted point set* from the literature on classification. In our setting they represent a finite amount of information obtained from a complex system.

In particular, let a probability measure $\mu_0 \in \mathcal{P}(M)$ represent the possible measurements on a system S. Each *elementary* measurement corresponds to a point of M, and if the state of the system S is repeatedly measured, we obtain a finite sequence $X_1, X_2, \ldots, X_n$ of iid random variables (with respect to the measure μ_0) taking values in M. These give rise to an *empirical measure*

$$\mu_n[A] = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}[A], \quad A \in \mathcal{B}. \tag{B.3}$$

The measure μ_n is itself a random variable, but fixing the outcomes, i.e., considering a realization $(x_1, x_2, \ldots, x_n) \in M^n$, a measure $\mu \in \mathcal{P}_{\mathrm{F}}(M)$ is obtained,

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}, \tag{B.4}$$

which we call a *realization* of the measure μ_0. Denote the space of all probability measures (B.4) for fixed $n \in \mathbb{N}$ and $\mu_0 \in \mathcal{P}(M)$ by $\mathcal{P}_n(\mu_0)$.

B.2 Discrete optimal transportation

In this secion we will motivate the notion of distance with which we will be concerned in the rest of the thesis. The starting point is the question of how to define a useful distance for the measures in $\mathcal{P}_F$.

Example 10 (Total variation)**.** The *distance in variation* between two measures μ and

ν is

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\in\mathcal{B}} |\mu[A] - \nu[A]|. \tag{B.5}$$

It is obviously reflexive and symmetric. For the triangle inequality, let $\epsilon > 0$ and consider $A \in \mathcal{B}$ such that $d_{\mathrm{TV}}(\mu,\nu) < |\mu[A] - \nu[A]| + \epsilon$. Then

$$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &< |\mu[A] - \rho[A]| + |\rho[A] - \nu[A]| + \epsilon \\ &< \sup_{A\in\mathrm{M}} |\mu[A] - \rho[A]| + \sup_{A\in\mathrm{M}} |\rho[A] - \nu[A]| + 2\epsilon. \end{aligned} \tag{B.6}$$

Since this holds for all ϵ, the triangle inequality is established. Total variation distance metrizes the strong topology on the space of measures, and can be interpreted easily: If two measures μ and ν have total variation $p = d_{\mathrm{TV}}(\mu,\nu)$, then for any set $A \in \mathcal{F}$ the probability assigned to it by μ and ν differs by at most p. For two measures $\mu, \nu \in \mathcal{P}_F$ concentrated on a countable set $x_1, x_2, \ldots$, it simplifies to

$$d_{\mathrm{TV}}(\mu,\nu) = \sum_i |\mu[x_i] - \nu[x_i]|. \tag{B.7}$$

Unfortunately, total variation needs further effort to be usable in practice. Consider an absolutely continuous $\mu_0 \in \mathcal{P}(M)$ with density $f : M \to [0,1]$. For two realizations $\mu, \nu \in \mathcal{P}_n(\mu_0)$ we have that $\mathrm{pr}(\operatorname{supp}\mu \cap \operatorname{supp}\nu \neq \emptyset) = 0$, so $d_{\mathrm{TV}}(\mu,\nu) = 0$ almost surely. In practice, therefore, we will need to use some kind of density estimation to achieve a non-trivial value $d_{\mathrm{TV}}(\mu,\nu)$; confer (Schmid and Schmidt, 2006).

Example 11. The Hausdorff metric is a distance of subsets of a metric space (Example 5). It can be turned into a distance for probability measures by "forgetting" the probabilistic weights, i.e.,

$$d_{\mathrm{HD}}(\mu,\nu) \stackrel{\mathrm{def}}{=} d_{\mathrm{H}}(\operatorname{supp} f, \operatorname{supp} g), \tag{B.8}$$

If M is a normed vector space, then a subset $A \subset M$ and its translation $x + A = \{x + a \mid a \in A\}$ have Hausdorff distance $d_H(A, x + A) = ||x||$, which seems natural. However, Hausdorff distance is unstable against outliers. For example, consider the family of measures defined by $P_0 = \delta_0$ and $P_n = \frac{1}{n}\delta_n + (1 - \frac{1}{n})\delta_0$ for all $n > 0$. Then $d_{\mathrm{HD}}(P_0, P_n) = n$. □

Example 12 (Symmetric pullback distance)**.** Let $f : M^n \to N$ be the projection of an ordered n-tuple from M into a single point of a metric space (N, d'). Call f *symmetric* if its value does not depend on the order of its arguments, i.e., if $f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for all permutations σ from the symmetric group $\Sigma(n)$ on n elements. Then

$$d_f(X,Y) \stackrel{\mathrm{def}}{=} d'(f(X), f(Y)) \tag{B.9}$$

defines a distance between n-element subsets $X, Y \subset M$ (the symmetric pullback of the distance in N).

In particular, if M has the structure of a vector space, then each function $f : M^n \to N$ can be symmetrized, yielding a symmetric function

$$f_\sigma(x_1, \dots, x_n) \stackrel{\text{def}}{=} \frac{1}{n!} \sum_{\sigma \in \Sigma(n)} f(x_{\sigma(1)}, \dots, x_{\sigma(n)}). \tag{B.10}$$

For the projection to the first factor,

$$f : M^n \to M, \quad (x_1, \dots, x_n) \mapsto x_1, \tag{B.11}$$

this yields the *centroid*

$$f_\sigma(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{B.12}$$

with centroid distance $d_f(X, Y) = d(\bar{X}, \bar{Y})$. This construction generalizes in the obvious way to finite probability measures $\mu, \nu \in \mathcal{P}_n(\mu_0)$.

Note however, that the symmetric pullback distance is pseudo-metric: There usually exist many n-subsets X, Y of M with the same pullback distance, i.e., $d_f(X, Y) = 0$ does not imply that $X = Y$.

All the above distances have various shortcomings that are not exhibited by the following distance. Let μ, ν be two probability measures on M and consider a cost function $c : M \times M \to \mathbb{R}_+$. The value $c(x, y)$ represents the cost to transport one unit of (probability) mass from location $x \in M$ to some location $y \in M$. We will model the process of transforming measure μ into ν, relocating probability mass, by a probability measure π on $M \times M$. Informally, $\mathrm{d}\pi(x, y)$ measures the amount of mass transferred from location x to y. To be admissible, the transference plan π has to fulfill the conditions

$$\pi[A \times M] = \mu[A], \quad \pi[M \times B] = \nu[B] \tag{B.13}$$

for all measurable subsets $A, B \subseteq M$. We say that π has marginals μ and ν if (B.13) holds, and denote by $\Pi(\mu, \nu)$ the set of all admissible transference plans.

Kantorovich's *optimal transportation problem* is to minimize the functional

$$I[\pi] = \int_{M \times M} c(x, y) \, \mathrm{d}\pi(x, y) \quad \text{for} \quad \pi \in \Pi(\mu, \nu) \tag{B.14}$$

over all transference plans $\Pi(\mu, \nu)$.

The optimal transportation cost between μ and ν is the value

$$T_c(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} I[\pi], \tag{B.15}$$

and transference plans $\pi \in \Pi(\mu,\nu)$ that realize this optimum are called *optimal transference plans*.

Since (B.14) is a convex optimalization problem it admits a dual formulation. Assume that the cost function c is lower semi-continuous, and define

$$J(\varphi,\psi) = \int_M \varphi \,\mathrm{d}\mu + \int_M \psi \,\mathrm{d}\nu \tag{B.16}$$

for all integrable functions $(\varphi,\psi) \in \mathcal{L} = L^1(\,\mathrm{d}\mu) \times L^1(\,\mathrm{d}\nu)$. Let Φ_c be the set of all measurable functions $(\varphi,\psi) \in \mathcal{L}$ such that

$$\varphi(x) + \psi(y) \le c(x,y) \tag{B.17}$$

for $\,\mathrm{d}\mu$-almost all $x \in M$ and $\,\mathrm{d}\nu$-almost all $y \in M$. Then (Villani, 2003, Th. 1.3)

$$\inf_{\Pi(\mu,\nu)} I[\pi] = \sup_{\Phi_c} J(\varphi,\psi). \tag{B.18}$$

For measures $\mu,\nu \in \mathcal{P}_F$ with representations

$$\mu = \sum_{i=1}^m a_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j=1}^n b_j \delta_{y_j} \tag{B.19}$$

any measure in $\Pi(\mu,\nu)$ can be represented as a bistochastic $m \times n$ matrix $\pi = (\pi_{ij})_{i,j}$, where the source and sink conditions

$$\sum_{i=1}^m \pi_{ij} = b_j, \quad j = 1,2,\dots,n \quad \text{and} \quad \sum_{j=1}^n \pi_{ij} = a_i, \quad i = 1,2,\dots,m, \tag{B.20}$$

are the discrete analog of (B.13), and the problem is to minimize the objective function

$$\sum_{ij} \pi_{ij} c_{ij}, \tag{B.21}$$

where $c_{ij} = c(x_i,y_j)$ is the cost matrix.

Its dual formulation is to maximize

$$\sum_i \varphi_i a_i + \sum_j \psi_j b_j \tag{B.22}$$

under the constraint $\varphi_i + \psi_j \le c_{ij}$.

Example 13 (Discrete distance)**.** Consider the special cost $c(x, y) = 1_{x \neq y}$, i.e., the distance induced by the discrete topology. Then the total transportation cost is

$$T_c(\mu, \nu) = d_{\mathrm{TV}}(\mu, \nu). \tag{B.23}$$

The Kantorovich problem (B.14) is actually a relaxed version of Monge's transportation problem. In the latter, it is further required that no mass be split, so the transference plan π has the special form

$$\mathrm{d}\pi(x, y) = \mathrm{d}\mu(x)\delta[y = T(x)] \tag{B.24}$$

for some measurable map $T : M \to M$. The associated total transportation cost is then

$$I[\pi] = \int_M c(x, T(x))\, \mathrm{d}\mu(x), \tag{B.25}$$

and the condition (B.13) on the marginals translates as

$$\nu[B] = \mu[T^{-1}(B)] \quad \text{for all measurable} \quad B \subseteq M. \tag{B.26}$$

If this condition is satisfied, we call ν the *push-forward* of μ by T, denoted by $\nu = T\#\mu$. For measures $\mu, \nu \in \mathcal{P}_F$, the optimal transference plans in Kantorovich's problem (transportation problem) coincide with solutions to Monge's problem.

A further relaxation is obtained when the cost $c(x, y)$ is a distance. The dual (B.18) of the Kantorovich problem then takes the following form:

Theorem 9 (Kantorovich-Rubinstein (Villani, 2003)[ch. 1.2)**.**] Let $X = Y$ be a Polish space[1], and let c be lower semi-continuous. Then:

$$T_c(\mu, \nu) = \sup \left\{ \int_X \varphi \;\; \mathrm{d}(\mu - \nu); \quad \text{where} \right.$$
$$\left. \varphi \in L^1(d|\mu - \nu|) \quad \text{and} \quad \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{c(x, y)} \leq 1 \right\} \tag{B.27}$$

The Kantorovich-Rubinstein theorem implies that $T_d(\mu + \sigma, \nu + \sigma) = T_d(\mu, \nu)$, i.e., the invariance of the Kantorovich-Rubinstein distance under subtraction of mass (Villani, 2003, Corollary 1.16). In other words, the total cost only depends on the difference $\mu - \nu$. The Kantorovich problem is then equivalent to the Kantorovich-Rubinstein *transshipment problem*: Minimize $I[\pi]$ for all product measures $\pi : M \times M \to \mathbb{R}_+$, such that

$$\pi[A \times M] - \pi[M \times A] = (\mu - \nu)[A]$$

[1] A topological space is a Polish space if it is homeomorphic to a complete metric space that has a countable dense subset. This is a general class of spaces that are convenient to work with. Many spaces of practical interest fall into this category.

for all measureable sets $A \subseteq \mathcal{B}(M)$. This transshipment problem is a strongly relaxed version of the optimal transportation problem. For example, if $p > 1$ then the transshipment problem with cost $c(x, y) = ||x - y||^p$ has optimal cost zero (Villani, 2003). For this reason, the general transshipment problem is not investigated her.e

Example 14 (Assignment and transportation problem)**.** The discrete Kantorovich problem (B.19-B.21) is also known as the (Hitchcock) *transportation problem* in the literature on combinatorial optimization (Korte and Vygen, 2007). The special case where $m = n$ in the representation (B.19) is the *assignment problem*. Interestingly, as a consequence of the Birkhoff theorem, the latter is solved by a permutation σ mapping each source a_i to a unique sink $b_{\sigma(i)}$ $(i = 1, \ldots, n)$; confer (Bapat and Raghavan, 1997).

B.3 Optimal transportation distances

Let (M, d) be a metric space and consider the cost function $c(x, y) = d(x, y)^p$, if $p > 0$ and $c(x, y) = 1_{x \neq y}$ if $p = 0$. Recall that $T_c(\mu, \nu)$ denotes the cost of an optimal transference plan between μ and ν.

Definition 18 (Wasserstein distances)**.** Let $p \geq 0$. The *Wasserstein distance of order p* is $W_p(\mu, \nu) = T_{d^p}(\mu, \nu)^{1/p}$ if $p \in [1, \infty)$, and $W_p(\mu, \nu) = T_{d^p}(\mu, \nu)$ if $p \in [0, 1)$.

Denote by $\mathcal{P}_p$ the space of probability measures with finite moments of order p, i.e., such that

$$\int d(x_0, x)^p \, \mathrm{d}\mu(x) < \infty$$

for some $x_0 \in M$. The following is proved in (Villani, 2003, Th. 7.3):

Theorem 10. The Wasserstein distance W_p, $p \geq 0$, is a metric on $\mathcal{P}_p$.

The Wasserstein distances W_p are ordered: $p \geq q \geq 1$ implies, by Hölder's inequality, that $W_p \geq W_q$. On a normed space, the Wasserstein distances are minorized by the distance in means, such that

$$W_p(\mu, \nu) \geq \left|\left| \int_X x \, \mathrm{d}(\mu - \nu) \right|\right|_p \tag{B.28}$$

and behave well under rescaling:

$$W_p(\alpha\mu, \alpha\nu) = |\alpha| W_p(\mu, \nu),$$

where $\alpha\mu$ indicates the measure $m_\alpha \# \mu$, obtained by push-forward of multiplication by α. If $p = 2$ we have the additional subadditivity property

$$W_2(\alpha_1\mu_1 + \alpha_2\mu_2, \alpha_1\nu_1 + \alpha_2\nu_2) \leq \left(\alpha_1^2 W_2(\mu_1, \nu_1)^2 + \alpha_2^2 W_2(\mu_2, \nu_2)^2\right)^{1/2}.$$

Appendix C

The dts software package

C.1 Implementation and installation

The methods of distance-based analysis can only be used in practice if there exists a reliable computational code. We have therefore implemented a number of algorithms as a software package for the statistical computing environment R (R Development Core Team, 2008).

The main computational routine td solves discrete transportation or assignment problems in $\mathbb{R}^n$ for a variety of ground distances and Wasserstein orders. Distances can be wrapped for phase distributions. One-dimensional problems are efficiently solved by monotone arrangement. All other problems are either solved by a dedicated minimum-cost flow solver or by a general linear programming solver. The first possibility is offered by the MCF code (Löbel, 1996) which is freely available for academic users[1]. Due to license issues, this code was not incorporated into the dts package, but compiles into the library if it is present. The second possibility is offered by the lpSolve package, which needs to be separately installed[2].

The package dts is available as a source-code distribution under the terms of the Creative Commons Attribution-Noncommerical-Share Alike 3.0 Unported License[3]. We will discuss its installation for a general UNIX system here. To install dts when MCF is present, let `MCF_ROOT` be the directory containing the MCF files. Issuing the command

```
R CMD INSTALL dts.tar.gz --configure-vars='MCF_ROOT=${MCF_ROOT}$'
```

configures and installs dts from the source package dts.tar.gz. If MCF is not present, the package can be installed in the usual way, but for computations the package lpSolve is needed, and they will be slower. In this case, installation can be performed by executing

```
R CMD INSTALL dts.tar.gz
```

on the command line. Note that administrative rights might be needed for a global installation.

Further packages that are required to use all features of dts are MASS (for multi-

[1] URL: http://www.zib.de/Optimization/Software/Mcf
[2] URL: http://cran.r-project.org/web/packages/lpSolve/index.html
[3] URL: http://creativecommons.org/licenses/by-nc-sa/3.0/

dimensional scaling), vegan (for MRPP permutation testing), odesolve (for integration of differential equations) and ROCR (for receiver-operator curves).

C.2 Reference

The following is a complete description of all functions available in the dts package, version 1.0-0[4].

`cmdscale.add` *Out-of-sample classical multidimensional scaling*

Description

Obtain the coordinates of an additional point in classical multidimensional scaling.

Usage

```
cmdscale.add(x, m, k = 2, points = NULL, verbose = FALSE, ntries = 10,
   max.iter = 100)
```

Arguments

`x`	Vector of distances between the additional point and all previous points.
`m`	Distance matrix of all previous points.
`k`	Dimension of Euclidean space in which to represent the points.
`points`	Reconstructed coordinates of previous points (optional).
`verbose`	Logical value to indicate whether details of the computation should be shown.
`ntries`	Number of times the solution is attempted.
`max.iter`	Maximal number of iterations in the minimalization problem.

Details

The out-of-sample problem consists in approximating the coordinates of an additional point in a representation of `n` previous points obtained by multidimensional scaling, from its distances with all previous points. In the case of classical multidimensional scaling considered here, the problem can be solved by minimizing a nonlinear error functional (Trosset and Priebe, 2008). The `R` function `optim` is called a number `ntries` of times to perform a simplex search (Nelder and Mead, 1965), and the coordinates that result in the minimal error are returned. For the previous points, coordinates in `points` are used if given; otherwise these are calculated by multidimensional scaling from the distances in `m`. Since the coordinates in multidimensional scaling are unique up to a rotation, this is useful to ensure that the out-of-sample point lies in an already established coordinate system.

[4] The version number follows the major/minor convention, where the first number indicates significant (major) changes, the second number minor changes (with even numbers indicating stable releases and odd numbers indicating developmental versions), and the last number is used to indicate consecutive bug fixes.

Value

A list containing the following components:

`points`	A matrix whose rows contain the coordinates of all points (including the out-of-sample point)
`eig`	A vector of the largest `k` eigenvalues (see `cmdscale` in package `MASS`)
`y`	A vector containing the coordinates of the out-of-sample point
`points0`	A matrix whose rows contain the coordinates of all previous points (excluding the out-of-sample point)

Note

This function uses the `cmdscale` code from the package MASS to obtain the representation of the original points.

Examples

```
library(MooreRayleigh)                # rsphere
x  <- rsphere(10)                     # uniform sample on the sphere
x0 <- rsphere(1)                      # one additional point
m.all <- as.matrix(dist(rbind(x,x0))) # all mutual Euclidean distances
attributes(m.all)$dimnames <- NULL
m     <- m.all[1:10,1:10]
m0    <- m.all[11,1:10]
library(MASS)
par(mfrow=c(1,2))
mds     <- cmdscale(m.all,k=2,eig=TRUE)$points  # project to the plane
eqscplot(mds[,1],mds[,2],xlab="x",ylab="y",
   tol=0.3,main="MDS")
points(mds[11,1],mds[11,2],pch=4)               # mark additional point
mds.add <- cmdscale.add(m0,m,k=2,points=mds[1:10,])$points
eqscplot(mds.add[,1],mds.add[,2],xlab="x",ylab="y",
   tol=0.3,main="Out-of-sample MDS")
points(mds.add[11,1],mds.add[11,2],pch=4)       # mark additional point
cat(paste("Distance between MDS and out-of-sample MDS =",
   round(sqrt(sum(mds.add[11,]-mds[11,])^2),4),"\n"))
```

`ldadist.cv` *Cross-validated linear discriminant analysis of distances matrices*

Description

Linear discriminant analysis for distance matrices.

Usage

```
ldadist.cv(x, classes, pc = NULL, search = FALSE, verbose = TRUE)
```

Arguments

x	A square matrix of mutual distances between n data items.
classes	A factor specifying the class membership of each data item.
pc	The number of components to use in the reconstruction. Can be empty if search is TRUE.
search	A logical value indicating whether to search for the optimal number of components (see details).
verbose	A logical value indicating whether to show details when search is TRUE.

Details

Linear discriminant analysis is performed on coordinates obtained by classical multidimensional scaling. This assumes the data to be represented by points in an n-dimensional Euclidean space. The class membership is estimated, and the fraction of correct classifications defines the *accuracy*. To reliably estimate this, leave-one-out crossvalidation is implemented by the out-of-sample method. In detail, for each data item its distance information is removed from x, the coordinates of the remaining points are calculated by classical multidimensional scaling, and the coordinates of the present point are approximated from its distances to the remaining points (see cmdscale.add). The classification of the single point is obtained by predict.lda from the MASS package, with the remaining points as training data for the linear discriminant function (using lda from the MASS package).

The number of components n can be specified (parameter pc). If search is TRUE, the number of components is searched that results in the best accuracy. In this case, the parameter pc is the maximal number of components to use. If pc is not given, the value n-4 is used (the points usually become collinear for larger numbers of components).

Value

A list containing the following components:

predict, cv.predict
: A vector containing the predicted class membership.

posterior, cv.posterior
: A vector containing the maximum a posteriori classification probabilities for the estimated class membership.

pc, cv.pc
: The number of components used in the representation by multidimensional scaling.

tab, cv.tab
: A two-by-k summary table of the classification, where k is the number of distinct classes.

accuracy, cv.accuracy
: A single number containg the (average) classification accuracy.

correct, cv.correct
: A vector containing ones (correct) and zeros (false) for the classification of each item.

acc, cv.acc
: A vector containg the (average) classification accuracies for each number of components evaluated. Equal to accuracy (cv.accuracy) if search is FALSE.

If search is TRUE, the values are returned for the number of components with the highest accuracy among all searched.

Note

If the number of components is determined by searching, the resulting accuracies are slightly overestimated and have to be interpreted with care. For large sample sizes this bias can be avoided by an additional stage of crossvalidation, but this has not been implemented.

Also note that the number of components should at most be equal to the number of samples from the smallest class.

Examples

```
ndim <- 6
# generate 20 normal variates in "ndim" dimensions
x <- matrix(rnorm(20*ndim),ncol=ndim)
# translate the second half
x[11:20,] <- x[11:20,] + c(0,0,0,rnorm(ndim-3,mean=1))
m <- as.matrix(dist(x))
attributes(m)$dimnames <- NULL
grouping <- as.factor(c(rep(1,10),rep(2,10)))
res <- ldadist.cv(m,grouping,search=TRUE,pc=ndim-1,verbose=TRUE)
```

`mfdfa` *Multifractal detrended fluctuation analysis*

Description

Multifractal detrended fluctuation analysis.

Usage

```
MFDFA(x, detrend = "poly1", q = c(1, 2), sum.order = 0,
   scale.max = trunc(length(x)/4), scale.min = 16, scale.ratio = 2,
   verbose = FALSE)
```

Arguments

`x`	Time series
`detrend`	Detrending method. Can be either '`bridge`' for bridge regression or '`polyN`' for polynomial detrending, where $N > 0$ indicates the order to use.
`q`	A numerical vector that indicates which scaling exponents to extract. Standard detrended fluctuation analysis corresponds to $q = 2$.
`sum.order`	Number of integrations (positive order) or differentiations (negative order) to perform before the analysis.
`scale.max`	Maximal scale to consider.
`scale.min`	Minimal scale to consider.
`scale.ratio`	Ratio between successive scales.
`verbose`	A logical value that indicates whether to show details of the computation.

Details

Deviations of a time series X_i ($i = 1, \ldots, N$) from its mean $\bar{X}$ are first integrated,

$$Y_i = \sum_{j=1}^{N} (X_i - \bar{X}),$$

leading to an unbounded profile. For a given scale $s > 0$, the profile Y_i is then divided into $N_s = \mathrm{int}(N/s)$ nonoverlapping segments of length s. Since the length N of the time series is usually not a multiple of the scale s, a short part at the end of the profile may remain. In order not to disregard this part, the procedure is repeated starting from the opposite end, leading to a total of $2N_s$ segments, denoted by $Y_{k;j}$ ($k = 1, \ldots, 2N_s$; $j = 1, \ldots, s$). For each segment a trend $Z_{k;j}$ (usually linear or quadratic, corresponding to `detrend="poly1"` or `detrend="poly2"`) is individually estimated by least-squares and subtracted. The fluctuation function for a given order $q \geq 0$ is given by

$$F_q(s) = \left(\frac{1}{2N_s} \sum_k \left(\frac{1}{s} \sum_j (Y_{k;j} - Z_{k;j})^2 \right)^{q/2} \right)^{1/q}.$$

This procedure is repeated for a number of scales s and exponents q. The scaling behavior is then assessed by weighted least-squares fitting a line to the scaling function $F_q(s)$ with respect to scale s in a double logarithmic plot, such that $\log F_q(s) \propto \alpha \log s$, with weights proportional to the index of scale (e.g., the third scale is weighted 2/3 relative to the second scale), to compensate for the reduction in data points on which the estimation of the corresponding $F_q(s)$ is based. If the residual error of the fit R^2 is large enough, the estimate $\alpha \geq 0$ is the q-scaling exponent of `x` for each value of q.

Exponents for $q = 0$ can also be evaluated, but need special treatment, with the scaling function given by

$$F_0(s) = \exp\left(\frac{1}{4N_s} \sum_k \ln\left(\frac{1}{s} \sum_j (Y_{k;j} - Z_{k;j})^2 \right) \right).$$

Note that only nonnegative exponents can be evaluted. However, additional integrations (or finite differences) can be performed before the analysis, indicated by `sum.order`. Since each integration increases the exponent by one, the estimated exponent α is corrected by subtracting `sum.order-1` from it at the end. This allows to also resolve negative exponents.

Value

A list containing the following components:

`h`	A vector containing the estimated scaling exponents.
`r.squared`	The residual errors of the linear fits of the scaling relationship.
`scale`	A vector that contains the scales that were considered.
`rmse`	A matrix that contains the residual errors for each exponent at each scale.
`q`	A vector that indicates which scaling exponents were extracted.

Note

This implementation is based heavily on the `DFA` code in the fractal package of W. Constantine and D. Percival (unpublished), correcting an error in the detrending and adding functionality for calculation of multifractal exponents.

Examples

```
x   <- rnorm(2000
foo <- mfdfa(x,q=seq(0,10),detrend="poly2",verbose=TRUE)
plot(res$r.squared,ylim=c(0,1),type="b",xlim="")  # goodness-of-fit
lines(res$h,type="b",pch=2)                       # scaling exponents
legend("bottomleft",pch=c(1,2),legend=c("goodness-of-fit","exponents"))
abline(h=1/2)                                     # theoretical value
```

`mle.pl` *Maximum likelihood estimation for power laws*

Description

Usage

```
mle.pl(x, min.tail = 50, cut = 0, verbose = FALSE, nboot = NULL)
plot.pl(x, ...)
print.pl(x)
test.pl(x, min.tail = 50, cut = 0, verbose = FALSE, nboot = 2500)
```

Arguments

`x`	A numerical vector of positive measurements.
`min.tail`	Minimum number of measurements to use in the estimation.
`cut`	Proportion of measurements to throw away (see details).
`verbose`	A logical value that indicates whether to show details of the estimation procedure.
`...`	Additional arguments for plotting.
`nboot`	Number of bootstrap replicates to use.

Details

Maximum likelihood estimation is implemented in function `mle.pl` to fit the Pareto distribution (see `powerlaw`) to the samples in `x`. The lower-cut off point $x_{\min}$ is determined by the method of Clauset et al. (2009): The `n` samples in `x` are sorted and each of the smallest `n-min.tail` samples is considered as a candidate for the lower cut-off point $x_{\min}$. For each such candidate the Pareto distribution is fitted, resulting in an estimate of the power-law exponent α, and the Kolmogorov-Smirnov statistic KS quantifies the maximal difference between the fitted distribution and the empirical distribution function (for samples from `x` greater or equal to $x_{\min}$). Finally, the value of $x_{\min}$ that minimizes KS is chosen and the corresponding parameter values are returned.

It is recommended to keep at least about 50–100 samples for the estimation, which can be adjusted by the parameter `min.tail`. To speed up the computation, a fraction `cut` of the smallest samples can be discarded before the estimation.

To quantify estimation uncertainty, a number `nboot` of bootstrap replicates can be optionally specified. Each replicate is generated by resampling with replacement from `x` and estimating the parameters $x_{\min}$ and α by the above procedure.

Function `print.pl` provides a summary of the estimation results and `plot.pl` a diagnostic plot. Function `test.pl` performs the simple test of the power-law hypothesis described in Clauset et al. (2009)[Section 4.1]. First the parameters $x_{\min}$ and α are estimated as above. Then a number `nboot` of bootstrap samples are generated where samples are independently drawn either from the fitted power-law model for the tail, or sampled with replacement from the samples of `x` smaller than $x_{\min}$. The probability to choose a sample from the tail is given by its (relative) length. For each of these replicates the estimation procedure in `mle.pl` is repeated, resulting in a set of `nboot` values of the KS statistic (for the best fit of the Pareto distribution). The significance probability for the nullhypothesis of power-law behaviour is given by the fraction of these that are larger than the KS statistic for the fit of the original data `x`. If this is large enough (Clauset et al. (2009) recommend a value of 0.10 for a conservative test), the general alternative is rejected and the power-law hypothesis is accepted.

Value

Function `mle.pl` returns a list with class "`pl`" containing the following components:

`x`	The original data.
`xmin.all,n.all`	All values of $x_{\min}$ considered in the estimation procedure and the length of the remaining tail for which the maximum-likelihood estimation of the exponent α was performed.
`alpha.all,D.all`	Values of α and the KS statistic corresponding to the cut-off points in `xmin.all`.
`ntail`	The length of the tail for the optimal parameter choice.
`xmin, alpha, D`	Values of the parameters and the KS statistic for the optimal parameter choice.
`nboot`	The number of bootstrap replicates used to quantify estimation error.
`alpha.boot,xmin.boot`	If `nboot` is given, these contain the parameter estimates for each bootstrap replicate.

Function `test.pl` returns a list with class "`htest`" containing the following components:

`statistic`	value of the test statistic for the data in `x`
`p.value`	the significance probability for the test
`alternative`	a character string describing the alternative hypothesis ('not symmetric')
`method`	a character string describing the type of test
`data.name`	a character string giving the names of the data
`nboot`	the number of bootstrap replicates used for the test
`xmin.boot,alpha.boot,D.boot`	the values of the estimated parameters and the KS statistic for each replicate
`xmin,alpha,D`	the values of the estimated parameters for the original data in `x`
`x`	the original data

Examples

```
x  <- ppl(1000,2,10)                    # generate synthetic power-law data
pl <- mle.pl(x,cut=0.2,verbose=TRUE,nboot=100)
pl                                      # summary output
plot(pl)                                # diagnostic plot
foo <- test.pl(x,cut=0.2,nboot=100,verbose=TRUE)
foo                                     # should reject the alternative
```

`powerlaw` *The power-law distribution*

Description

Density, distribution function and random generation for the Pareto distribution, i.e., the power-law distribution with lower cut-off point.

Usage

```
dpl(x, alpha, xmin)
ppl(q, alpha, xmin)
rpl(n, alpha, xmin)
```

Arguments

`x, q`	Vector of quantiles.
`n`	Number of observations. If `length(n) > 1`, the length is taken to be the required number.
`alpha`	Exponent $\alpha \geq 1$ of the power-law.
`xmin`	Lower cut-off point $x_{\min} \geq 0$.

Details

The density of the power-law distribution is proportional to $x^{-\alpha}$ with exponent $\alpha \geq 1$. Since this is not integrable for $x \to 0$, it is customary to restrict the power-law distribution to values of x greater than a (lower) cut-off point $x_{\min} > 0$. This is called the Pareto distribution, and its density is given by

$$f(x) = (\alpha - 1) x_{\min}^{\alpha-1} x^{-\alpha}.$$

Its two parameters α and $x_{\min}$ are usually called the shape and scale parameters.

Value

`dpl` gives the density, `ppl` gives the distribution function, and `rpl` generates random variates by the transformation method.

Examples

```
x <- seq(0,100,0.1)
# power-law leads to a straight line in a double logarithmic plot
plot(log10(x),log10(1-ppl(x,3,1)),ylab=expression(log10(1-F(x))))
abline(v=log10(1),lty=2)  # cut-off point
```

`samp.en` *Sample entropy and cross-entropy*

Description

Calculate sample entropy (SampEn) and cross-sample entropy to estimate the rate of information production in dynamical systems.

Usage

```
samp.en(x, y = NULL, r = NULL, r.ratio = NULL, edim = 2, tlag = 1,
  normalize = TRUE, size = NULL, verbose = FALSE)
```

Arguments

`x`	A numerical vector containing the time series for which sample entropy is calculated.
`y`	Optional vector containing a second time series.
`r`	The threshold when to consider two vectors as being neighbours (see details). If not given, use `r.ratio` times the standard deviation of `x`. If `r.ratio` is also not given, use a value of 0.2 for it.
`r.ratio`	If `r` is not given, use `r.ratio` times the standard deviation of `x` as the threshold.
`edim,tlag`	Embedding dimension and time lag to use.
`normalize`	A logical value indicating whether to center and normalize the time series to zero mean and unit standard deviation.
`size`	If present, draw randomly `size` vectors with replacement from the delay vector embedding for the calculation.
`verbose`	A logical value that indicates whether to show details of the calculation.

Details

Calculates the sample entropy (SampEn) introduced by Richman and Moorman (2000) from the time series in `x`. If additional `y` is given, calculates their cross-sample entropy (Cross-SampEn). Both time series are delay embedded with time lag `tlag` in `edim` and `edim+1` dimensions (see `rdelay`). To ensure an equal number of delay vectors in both embeddings, a few vectors at the end of the former are discarded. Then the number of pairs of delay vectors that lie within a distance `r` of each other is counted for both these delay embeddings, resulting in two counts B (in dimension `edim`) and A (in dimension `edim+1`). For computational efficiency the maximum distance is used, and in contrast to the approximate entrop (ApEn) of Pincus (1991) self-matches are not counted. If `y` is given, the distances are calculated between all pairs of delay vectors where one arises from the embedding of `x` and the other from the corresponding one for `y`. Sample entropy is then defined as the negative (natural) logarithm of A/B and is a finite approximation of the Kolmogorov-Sinai entropy (obtained in an appropriate limit of infinite data and vanishing threshold `r`).

Value

A list containing the following components:

`s`	Sample entropy.
`B`	The number of close neighbours in an `edim`-dimensional embedding.
`A`	The number of close neighbours in an `edim+1`-dimensional embedding.

Examples

```
x  <- seq(0,100,0.2)
y1 <- rnorm(length(x))   # a random sample
samp.en(y1)              # large sample entropy
y2 <- sin(x*2*pi/10)     # a deterministic process
samp.en(y2)              # low sample entropy
samp.en(y1,y2)           # large cross-sample entropy
```

`td` *Wasserstein distances for finite distributions of points*

Description

Calculates Wasserstein distance between two sets of multi-dimensional vectors.

Usage

```
td(x, y, wx = NULL, wy = NULL, dist = "l2", order = 2,
   cost.scale = NULL, phases = FALSE, verbosity = 0)
td.1d(x,y, wx = NULL, wy = NULL, dist = "l2", order = 2,
   phases = FALSE, verbosity = 0)
td.lpSolve(x, y, wx = NULL, wy = NULL, dist = "l2", order = 2,
   cost.scale = NULL, phases = FALSE, verbosity = 0)
```

Arguments

`x`	Multi-dimensional point data as a matrix with individual points represented by columns. Can also be numerical vector (for one-dimensional problems).
`y`	Multi-dimensional point data as a matrix with individual points represented by columns. Can also be a numerical vector (for one-dimensional problems).
`wx`	Optional weights for `x`. Should sum to 1.
`wy`	Optional weights for `y`. Should sum to 1.
`dist`	Choose one from "`l1`", "`l2`", "`max`", or a numerical value ≥ 1 for a Minkowski (L_p) distance.
`order`	The order of the Wasserstein distance. Defaults to quadratic Wasserstein distance.
`cost.scale`	Optional scaling factor for weights. Only needed for multidimensional data.
`phases`	Logical value that indicates whether to wrap distances (for phase distributions) or not.
`verbosity`	Verbosity level of output. Higher values result in more diagnostic output.

Details

The Wasserstein distance between to k-dimensional point sets x and y is the cost associated with an optimal transportion problem. Both x and y are interpreted as discrete probability measures on a Euclidean space R^k, and their Wasserstein distance is the minimal total cost when transforming one measure into the other. Each unit of probability mass transported incurs a cost equal to the distance it is moved.

The distance used is in principle arbitrary; however, at present only the most common distances are implemented: 'l1' is the L_1 (Manhattan) distance, 'l2' is L_2 (Euclidean) distance and 'max' is supremum distance. Additionally, the 'order' can be given. Explicitly, the Wasserstein distance of order p is $W_p(x,y) = \inf \left(\int d(x-y)^p \, \mathrm{d}\pi[x,y] \right)^{1/p}$, where the infimum is taken over all probability measures (transportation plans) $\pi[x,y]$ such that $\pi[A,R^k] = x[A]$ and $\pi[R^k,B] = y[B]$ for all subsets $A, B \subseteq R^k$. The quadratic Wasserstein distance (default) has very interesting theoretical properties, in particular, it is possible to interpolate measures. More commonly used, however, is the Wasserstein distance of order 1, also known as the Kantorovich-Rubinstein distance.

In the discrete case considered here, the calculation of the Wasserstein distance is equivalent to solving a so-called (discrete) transportation problem: Let x and y be discrete probability measures, $x = \sum_i a_i \delta_{x_i}$ and $y = \sum_j b_j \delta_{x_k}$, where δ_x is the Dirac measure at the point $x \in R^k$. These can be interpreted as weighted point sets. The supplies $a_i \in (0,1]$ and the demands $b_j \in (0,1]$ need to be normalized, such that $\sum_i a_i = 1$ and $\sum_j b_j = 1$. Each transportation plan can then be represented as a nonnegative matrix f_{ij} that fulfills the source and sink conditions $\sum_j f_{ij} = a_i$ and $\sum_i f_{ij} = b_j$ for all i and j. The Wasserstein distance of order p is then

$$W_p(x,y) = \min \left(\sum_{ij} f_{ij} d(x_i - y_j)^p \right)^{1/p}$$

.

In the one-dimensional case, the problem is solved by monotone arrangement, i.e., by sorting the points of both samples and iteratively matching the largest value of x to its nearest neighbour from y. In the multivariate case, this routine will use the minimum-cost flow solver MCF *(Loebel, 1996)*, if available. Alternatively, the much slower **lpSolve** package is used. In both these cases internally integer arithmetic is used, so the distances and weights will be scaled to the nearest integer in a suitable range (see `discretize.factor`).

Value

The Wasserstein distance of the data.

Note

The computational complexity is high, theoretically on the order of $O(n^5)$ where $n = \max\{|x|, |y|\}$, although in practice often an almost quadratic complexity can be observed. Problems with more than 1000 data points will therefore need to be approximated by resampling smaller point sets a number of times (bootstrapping), or binning the points.

Examples

```
data1 <- c(1,2)
data2 <- c(1,2,3)
td(data1,data2)

data1 <- matrix(c(1,1,0,1,0.5,0.5),nrow=2)
```

```
data2 <- matrix(c(1,1,0,0,0.3,0.3),nrow=2)
tdp(data1,data2) # will be 7/15

data1 <- c(1,2)
data2 <- c(1,2,3)
weights1 <- c(0.9,0.1)
weights2 <- c(0.4,0.5,0.1)
td(data1,data2,weights1,weights2) # will be 0.6
```

`stress` *Diagnostics of misrepresentation error*

Description

Diagnostic measures of misrepresentation error in multidimensional scaling

Usage

```
stress(x, mds.dim)
```

Arguments

`x`	Distance matrix
`mds.dim`	Reconstruction dimension.

Details

The misrepresentation error of metric multidimensional scaling is evaluated. Given a n-by-n distance matrix D_{ij}, raw stress is the total residual square error of the Euclidean distances Δ_{ij} of metric multidimensional scaling in `mds.dim` dimensions,

$$\sigma = \sum_{ij} (D_{ij} - \Delta_{ij})^2.$$

The average of the residual square error with respect to the i-th point,

$$\sigma^{(i)} = \frac{1}{n} \sum_{j=1}^{n} (D_{ij} - \Delta_{ij})^2$$

is called stress-per-point and a useful measure of the local misrepresentation error. To compare these measures between distinct metric spaces, normalized stress is $\sigma_1 = \sigma / \sum_{ij} \Delta_{ij}^2$, and normalized stress-per-point is $\sigma_1^{(i)} = n\sigma_i / \sum_{ij} \Delta_{ij}^2$.

Value

A list containing the following components:

`stress`	Raw total stress
`spp`	A vector containg stress-per-point for each row of `x`
`stress1`	Normalized total stress
`spp1`	A vector containg normalized stress-per-point for each row of `x`

Examples

```
library(MASS)
library(MooreRayleigh)              # rsphere
x   <- rsphere(10)
d   <- as.matrix(dist(x))
attributes(d)$dimnames <- NULL
cmd <- cmdscale(d,k=2,eig=TRUE)
plot(cmd$points[,1],cmd$points[,2])
str <- stress(d,2)
symbols(cmd$points[,1],cmd$points[,2],
   circles=(1/pi*sqrt(str$spp)),inches=FALSE,add=TRUE)
```

`td.interp` *Interpolate two distributions along an optimal transport ray*

Description

Calculates Wasserstein distance between two sets of multi-dimensional vectors and shifts both distributions a fraction along the optimal transport rays.

Usage

```
td.interp(x, y, frac, dist = "l2", order = 2, cost.scale = NULL,
   verbosity = 0)
```

Arguments

`x`	Multi-dimensional point data as a matrix with individual points represented by columns. Can also be numerical vector (for one-dimensional problems).
`y`	Multi-dimensional point data as a matrix with individual points represented by columns. Can also be a numerical vector (for one-dimensional problems).
`frac`	Fraction of distance to move points on the optimal rays.
`dist`	Choose one from "`l1`", "`l2`", "`max`", or a numerical value ≥ 1 for a Minkowski (L_p) distance.
`order`	The order of the Wasserstein distance. Defaults to quadratic Wasserstein distance.
`cost.scale`	Optional scaling factor for weights. Only needed for multidimensional data.
`verbosity`	Verbosity level of output. Higher values result in more diagnostic output.

Details

The calculation of the optimal transport is the same as for the function `td`. The optimal transport mapping is used to shift each point of `x` and `y` a fraction `frac/2` into the direction of its partner (along an optimal transport ray) under the optimal matching.

Value

A list containing the following components:

`td`	The optimal transportation distance between `x` and `y`
`x`	The shifted data points of the original `x`
`y`	The shifted data points of the original `y`

Note

The state of this function is experimental. Currently, the interpolation is only available with the MCF solver and for equal-sized point sets (trivial weights).

Examples

```
library(MooreRayleigh)
x <- rsphere(10,2)              # 10 points on the unit circle
y <- matrix(rep(0,20),ncol=2)
foo <- td.interp(x,y,frac=0.5)
plot(x)
points(y,pch=19)
points(foo$x,col="blue")
points(foo$y,col="blue",pch=19)
```

`ts.delay` *Delay vector embedding of scalar time series*

Description

Delay vector embedding of scalar time series

Usage

```
ts.delay(x, edim, tlag = 1, ofs = 0)
```

Arguments

`x`	A numerical vector (scalar time series).
`edim`	Embedding dimension to use.
`tlag`	Distance between indices of adjacent components of delay vectors. Defaults to 1.
`ofs`	Number of data points to skip from the beginning of `x`. Defaults to 0.

Details

The delay representation of a numerical vector $(x_1, \ldots, x_n)$ with time lag `k` and offset `l` in embedding dimension `d` is the vector-valued series $y = (y_1, y_2, \ldots)$ given by:

$$y_1 = (x_l, x_{l+k}, \ldots, x_{l+k(d-1)})^t,$$

$$y_2 = (x_{l+1}, x_{l+k+1}, \ldots, x_{l+k(d-1)+1})^t,$$

$$\ldots$$

Value

A matrix whose columns contain the delay vectors.

Examples

```
x <- seq(1,9)
ts.delay(x,3)
ts.delay(x,3,3)
```

Appendix D

The MooreRayleigh software package

D.1 Implementation and installation

The Moore-Rayleigh test is a nonparametric test for the spherical symmetry of a sample of vectors under a general alternative. It is introduced and described in Chapter 5. The R package MooreRayleigh has been developed to perform the test in practice and is released under the GNU Public License Version 3.0[1]. The Moore-Rayleigh test has been implemented in R code, the related permutation tests of Diks and Tong (1999) have been implemented in C++ code for speed. Installation is straightforward by invoking

```
R CMD INSTALL MooreRayleigh.tar.gz
```

at the command line.

D.2 Reference

The following is a complete description of all functions available in the MooreRayleigh package, version 1.2-0[2].

`bisect` *Numerical bisection to find the root of a function*

Description

Iteratively bisects an interval to find the root of a continuous function. The initial points are assumed to bracket the root. Only the first zero found is returned.

Usage

```
bisect(x1, x2, f, max.iter = 38, tol = NULL)
```

[1] URL: http://www.gnu.org/licenses/lgpl.html

[2] The version number follows the major/minor convention, where the first number indicates significant (major) changes, the second number minor changes (with even numbers indicating stable releases and odd numbers indicating developmental versions), and the last number is used to indicate consecutive bug fixes.

Arguments

x1	One endpoint of the interval.
x2	The other endpoint of the interval.
f	A continuous function with a root between x1 and x2.
max.iter	Maximum number of bisections to try.
tol	Accuracy of the root. Defaults to (x1+x2)/2.0*.Machine$double.eps.

Details

An initial bracketing is given if f(x1)*f(x2) >= 0.0. The interval is halved every step after evaluating f at the midpoint.

Value

The argument for which f is zero, up to an accuracy of tol.

Examples

```
f <- function(x) { x }
bisect(-1,1,f)  # returns zero
bisect(2,4,sin) # approximates pi
```

---

diks.test *Monte Carlo testing for symmetry of multivariate samples*

Description

Test a multivariate sample for spherical or reflection symmetry.

Usage

```
diksS.test(x, bw = 0.25, n.mc = 1e3, center = FALSE, return.perms = FALSE)
diksU.test(x, bw = 0.25, n.mc = 1e3, center = FALSE, return.perms = FALSE)
```

Arguments

x	Numerical vector or matrix. If a matrix, each row is expected to contain the coordinates of a sample. A vector will be interpreted as a matrix with a single column.
bw	Bandwidth to use in the distance calculation.
n.mc	Number of Monte Carlo samples to use.
center	Center x by subtracting the column means before testing.
return.perms	Numerical value to indicate whether the values of the test statistic are to be returned for all Monte Carlo realizations.

Details

The hypothesis tests of Diks and Tong (1999) are permutation tests. The test statistic T is based on an U-estimator for the squared distance between two multivariate distributions. In terms of a finite sample $X_i \in \mathbb{R}^k$ $(i = 1, 2, \ldots, n)$ it is given by

$$T(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{i,j} K(x_i - x_j),$$

where K is the kernel function $K(X - Y) = \exp(-||X - Y||^2/(4d^2))$ that depends on a bandwidth parameter d (given by the parameter `bw` of the function). Under the nullhypothesis the common distribution of the X_i is invariant under the action of the respective symmetry group. This can be either spherical symmetry (`diksU.test`) or reflection symmetry (`diksS.test`). For each Monte Carlo sample an element of this group is chosen randomly, acts on the sample $(X_1, \ldots, X_n)$, and the corresponding value of T is recorded. The significance probability of the test is the fraction of such "permutations" with a value of T less than or equal to the one for the original sample.

Value

A list with class "`htest`" containing the following components:

`statistic`	value of the test statistic for the original data `x`.
`p.value`	the significance probability for the test.
`alternative`	a character string describing the alternative hypothesis ('not symmetric').
`method`	a character string describing the type of test.
`data.name`	a character string giving the names of the data.
`n.mc`	the number of Monte Carlo samples used.
`bw`	the bandwidth parameter used.
`centered`	a logical value describing whether the columns of `x` were centered.
`statistic.mc`	depending on the value of `return.perms` either `NULL` or a vector containing the values of the test statistic for all Monte Carlo samples.

Examples

```
x <- rsphere(100)          # 100 samples from the unit sphere in 3D
diksU.test(x, n.mc = 1e4)  # should accept the nullhypothesis
y <- F3(100)               # 100 samples from the Fisher distribution
diksU.test(x, n.mc = 1e4)  # should reject the nullhypothesis

x[,1:2] <- 0               # project to a uniform distribution on a line
diksU.test(z, n.mc = 1e4)  # should reject the null hypothesis
diksS.test(z, n.mc = 1e4)  # should accept the null hypothesis
```

`F3` *Fisher distribution*

Description

Generate random variates from the Fisher distribution, also known as the three-dimensional von-Mises distribution.

Usage

```
F3(n = 1, lambda = 1)
```

Arguments

`n`	Number of samples requested.
`lambda`	Concentration parameter. Must be nonnegative.

Details

The Fisher distribution is a singular distribution on the sphere $S^2 \subset \mathbb{R}^3$. Its density $f(x)$ is proportional to $\exp(\lambda\langle x, \xi\rangle)$, where $\xi \in \mathbb{R}^3$ is the mean direction, and $\lambda \geq 0$ is a concentration parameter. In this implementation $\xi = (0, 0, 1)^t$ is fixed as the unit z-vector, and random variates are generated according to the method of Ulrich (1984) and Wood (1994).

Value

A vector if just one numerical sample is requested, otherwise a matrix with one column for each sample.

Examples

```
m <- F3(500, lambda = 5)  # 500 x 3 matrix
library(lattice)          # load package lattice for 3D plotting
cloud(z ~ x*y, data=data.frame(x=m[,1],y=m[,2],z=m[,3]))  # point cloud
```

`lrw` *Properties of random walk with linearly increasing steps*

Description

Calculate the distribution of a symmetric, unbiased one-dimensional random walk with linearly increasing steps.

Usage

```
lrw(N, both.sides = FALSE, nonzero = TRUE)
dlrw(x, N, both.sides = FALSE, scaled = FALSE)
plrw(q, N, both.sides = FALSE, scaled = FALSE)
qlrw(p, N, both.sides = FALSE, scaled = FALSE)
rlrw(n, N, both.sides = FALSE, scaled = FALSE)
```

Arguments

`x, q`	Vector of quantiles
`p`	Vector of probabilities
`n`	Number of observations. If `length(n) > 1`, the length is taken to be the required number.
`N`	Number of steps taken by the walk.
`both.sides`	Logical value indicating whether the distribution is given for both sides of the coordinate system.
`nonzero`	Logical value indicating whether only sites with nonzero probability should be returned.
`scaled`	Logical value indicating whether the argument should be scaled by $N^{3/2}$.

Details

A random walk with `N` steps is considered. At the `n`-th step, the position of the walker either increases or decreases by `n`, with equal probability. The probability distribution of this walk is obtained by iterated convolution. Since it is symmetric, only the positive sites (with nonnegative probability) are returned by default.

Value

Function `lrw` returns various properties of the distribution of this walk:

`pr`	A vector with the probabilities for the walker to be at a certain site after `N` steps.
`counts`	A vector with the counts of how many possibilities there are for the walker to reach a certain site after `N` steps.
`signs`	A vector with the average sign for each site. For each possible distinct walk to finish at a specific site, the sign is the product of the signs of the individual steps taken by the walker.
`pr.signs`	Equal to `pr * signs`.
`dst`	A vector with the sites.

`dlrw` gives the density, `plrw` gives the distribution function, `qlrw` gives the quantile function, and `rlwr` generates random variates.

Examples

```
lrw(N=3,both.sides=TRUE,nonzero=FALSE)

x <- seq(0,100,0.1)
plot(x,dlrw(x,N=10),cex=0.5,pch=20)  # probability funct. after 10 steps
plot(x,plrw(x,N=10),cex=0.5,pch=20)  # distribution funct. after 10 steps

sum(dlrw(seq(1,55),N=10))            # sums to one
```

`mr` *Asymptotic Moore-Rayleigh distribution*

Description

Density, distribution function, quantile function and random generation for the asymptotic form of the Moore-Rayleigh distribution, i.e., for the length of the resultant of a random flight with N linearly growing steps, scaled by $N^{3/2}$, in the limit of $N \to \infty$.

Usage

```
dmr(x, k = 3)
pmr(q, k = 3)
qmr(p, k = 3)
rmr(n, k = 3)
```

Arguments

`x,q`	A vector of quantiles.
`p`	A vector of probabilities.
`n`	The number of variates requested. If `length(n) > 1` use the length as the required number.
`k`	The dimensionality (`>= 2`).

Value

`dmr` gives the density, `pmr` gives the distribution function, `qmr` gives the quantile function, and `rmr` generates random deviates.

Note

The asymptotic Moore-Rayleigh distribution with distribution function $F(x)$ in k dimensions is related to the χ^2 distribution with k degrees of freedom and distribution function $G(x)$ by $F(x) = G(3kx^2)$.

Examples

```
x <- seq(0.02,2.0,0.001)
p <- dmr(x)
plot(x,p,cex=0.5,pch=20,title="Density function in 3D")
d <- pmr(x)
plot(x,d,cex=0.5,pch=20,title="Distribution function in 3D")
```

`mr3` *Exact Moore-Rayleigh distribution in three dimensions*

Description

Density, distribution function, quantile function and random generation for the Moore-Rayleigh distribution in three dimensions, i.e., for the length of the resultant of a random flight with N linearly growing steps, scaled by $N^{3/2}$.

Usage

```
dmr3(x, N)
pmr3(q, N, method = "Borwein")
qmr3(p, N)
rmr3(n, N)
```

Arguments

`x,q`	A vector of quantiles.	
`p`	A vector of probabilities.	
`n`	Number of observations. If `length(n) > 1`, the length is taken to be the required number.	
`N`	Number of steps taken by the walk.	
`method`	This can be one of the following:	
	`Borwein`	Combinatorial Borwein summation (default).
	`Integrate`	Use `integrate` to evaluate the oscillating integrals directly.

Details

Value

`dmr3` gives the density, `pmr3` gives the distribution function, `qmr3` gives the quantile function, and `rmr3` generates random deviates.

Examples

```
x <- seq(0.02,2.0,0.001)
par(mfrow=c(1,2),mar=c(4,4,1,1))  # plot density and distribution
plot(x,dmr(x),type="l",lwd=1/2,lty=2,ylab=expression(pr(R^"*"==r)),
   xlab="r")
lines(x,dmr3(x,N=10),lwd=1/2)
plot(x,pmr(x),type="l",lwd=1/2,lty=2,ylab=expression(pr(R^"*"<=r)),
   xlab="r")
lines(x,pmr3(x,N=10),lwd=1/2)
```

`mr3.test` *Moore-Rayleigh test in three dimensions*

Description

Test a sample of three-dimensional vectors for spherical symmetry.

Usage

```
mr3.test(x, exact = NULL, center = FALSE)
```

Arguments

`x`	A matrix whose rows represent vectors in three dimensions.
`exact`	A logical value that indicates whether to use the exact or the asymptotic distribution.
`center`	A logical value that indicates whether to center the vectors before applying the test.

Details

The Moore-Rayleigh test is a hypothesis test for spherical uniformity under a general alternative. It ranks the `N` vectors in `x` by their lengths. Under the nullhypothesis the vectors are assumed to be distributed uniformly on each hypersphere, and the ranks are randomly realized. The test statistic is the length of the resultant of the vectors in `x`, normalized by their ranks, and corresponds to the distance covered by a uniform random flight with `N` linearly increasing steps under the nullhypothesis. It is scaled by $N^{3/2}$ for asymptotic simplicity. The distribution of the nullhypothesis is available in closed form, and evaluated by a combinatorial sum for an exact test (valid for $N \lesssim 60$ under IEEE 754 arithmetic) or approximated by the asymptotic distribution (see `mr`).

For a two-sample test, the vectors need to be paired (see `pairing`).

Value

A list with class "`htest`" containing the following components:

`statistic`	value of the test statistic for the data in `x`.
`p.value`	the significance probability for the test.
`alternative`	a character string describing the alternative hypothesis ('not symmetric').
`method`	a character string describing the type of test.
`data.name`	a character string giving the names of the data.
`centered`	a logical value describing whether the columns of `x` were centered.

Examples

```
x <- rsphere(10)          # 10 samples from the unit sphere
mr3.test(x,exact=TRUE)    # one-sample test: should reject alternative

y <- rsphere(10)
xy <- pairing.random(x,y)
mr3.test(xy,exact=TRUE)   # two-sample test: should reject alternative

y <- matrix(runif(30),ncol=3)
xy <- pairing.random(x,y)
mr3.test(xy,exact=TRUE)   # two-sample test: should accept alternative
```

`pairing` *Pairing of vectors for two-sample tests*

Description

Pair two set of vectors for a two-sample test.

Usage

```
pairing.transport(x, y, minimize = TRUE, normalize = FALSE)
pairing.ranks(x, y, inverse = FALSE)
pairing.random(x, y)
```

Arguments

`x, y`	Two matrices where each row represents the coordinates of a vector. The number of columns needs to be the same, the number of rows can differ.
`minimize`	Logical value that indicates whether the total cost is minimized or maximized.
`normalize`	Logical value that indicates whether to project all vectors to the unit sphere before matching them.
`inverse`	Logical value that indicates whether to begin matching vectors with the smallest (default) or the largest vectors (if `inverse` is `TRUE`).

Details

The preferred pairing is `pairing.random` in which vectors are randomly matched (by sampling without replacement). Function `pairing.ranks` pairs vectors according to their lengths. The value of `inverse` determines whether this proceeds from the smallest to the largest or the other way around. Function `pairing.transport` pairs vectors by optimal transport under Euclidean distance. The value of `minimize` determines whether the total cost is minimized (default) or maximized.

Value

A list with class "pairing" containing the following components:

x,y	original data.
xy	a matrix representing difference vectors for all pairs.
x.pairing	a vector of indices to index the vectors in x for the pairing.
y.pairing	a vector of indices to index the vectors in y for the pairing.
cost	value of the total cost (only for pairing.transport.
dist	the distance matrix (only for pairing.transport.

Examples

```
x <- rsphere(10)
y <- rsphere(10)
d1 <- numeric(1e3)
d2 <- d1
for (i in seq(along=d)) {
   xy <- pairing.random(x,y)$xy
   d1[i] <- sum(apply(xy^2,1,sum))  # squared lengths of difference
   xy <- pairing.ranks(x,y)$xy
   d2[i] <- sum(apply(xy^2,1,sum))
}
plot(density(d2),main="Sum of squared lengths of difference vectors",
   lty=2)
lines(density(d1),lty=1)
```

---

rsphere *Random variates on the unit sphere*

Description

Generate random vectors on the unit sphere in k-dimensional Euclidean space.

Usage

```
rsphere(n, k = 3)
```

Arguments

n	Number of random variates generated. If length(n) > 1 the length is taken to be the required number.
k	Dimension of the space (default = 3).

Details

Uses the method of Knuth, i.e., the fact that a k-dimensional vector of normal variates is uniformly distributed on the unit sphere S^{k-1} after radial projection.

Value

A matrix, where each row represents the coordinates of one random vector.

Examples

```
rsphere(2)
plot(rsphere(1000,k=2),cex=0.5,pch=20,xlab="",ylab="",
   main="Uniform distribution on the circle")
```

Notes

A great deal more is known than has been proved

Richard Feynman

Section 2.2

Determination of the embedding dimension by the method of false nearest neighbour is described by Kennel et al. (1992). Embedding for non-uniformly sampled time series is covered by Huke and Broomhead (2007).

Section 2.4

Alternative classification by coefficients of global dynamical models is considered in (Kadtke, 1995).

Section 2.5.4

The performance of various synchronization measures is compared by Ansari-Asl et al. (2006) for a number of numerical models of brain activity.

Section 2.7

Recurrence plots are a general tool to analyze dynamical systems, with manifold applications (Webber, Jr. and Zbilut, 1994; Marwan et al., 2007). Joint recurrence plots even allow to compare the dynamics of two dynamical systems defined on different phase spaces, overcoming the main theoretical problem when comparing systems measured with distinct modalities (Marwan et al., 2007). However, quantitative measures defined for recurrence plots do not fulfill metric properties and cannot be used in a multivariate context.

Section 3.1

A good introduction to the physiology of breathing is (Guyton and Hall, 2006). The book by Kulish (2006) contains some advanced topics and corresponding mathematical models.

The branching nature of the bronchial tree is described by Weibel (1963) and by the model of Horsfield et al. (1971). A three-dimensional generalization has been obtained by Kitaoka et al. (1999). Optimality principles that explain the geometry of the lung were considered by Weibel (1991) and criticized by Imre (1999).

The control of breathing is described in (Whipp, 1987; Bronzino, 2006). Mathematical models for this regulatory process include the basic Mackey-Glass model (Keener and Sneyd, 1998) and the famous Grodins model (Grodins et al., 1967a), both incorporating time delays.

General cardiovascular models are described in (Ottesen et al., 2004; Batzel et al., 2006).

Section 3.2

The forced oscillation method is reviewed in (MacLeod and Birch, 2001; Oostveen et al., 2003); the article by Nucci and Cobelli (2001) considers mathematical details. Normal values are described by Landser et al. (1982). Details of the frequency-domain approach and parameter estimation are given by Michaelson et al. (1975). The frequency-dependence of FOT was observed and validated in Jackson et al. (1987).

A few facts about lung mechanics are: The tidal lung volume is about $1\mathrm{L}$, with a dynamic driving pressure of about $-1\mathrm{mmHg}$. Airway resistance is highest at segmental bronchi and lower at higher airway generations. Similarly, resistance decreases nonlinearly with lung volume from 4 (2 L) to about $0.5\mathrm{cmH_2Os/L}$ (6 L); conductance increases linearly from about 0.25 (2 L) to about $2\mathrm{L/cmH_2Os}$ (6 L) (Herman, 2007)[pg. 539f]. Flow is at Reynolds number of about 1600, so mostly laminar. However, turbulence occurs because the walls are not smooth. Interestingly, the work to breathe can take up to 20% of total body energy consumption.

Partitioning of FOT signals has been pioneered by DuBois et al. (1956a). In this model, transfer and input impedance are partitioned as follows,

$$Z_{\mathrm{tr}} = Z_{\mathrm{aw}} + Z_{\mathrm{t}} + \frac{Z_{\mathrm{aw}} Z_{\mathrm{t}}}{Z_{\mathrm{g}}},$$
$$Z_{\mathrm{in}} = Z_{\mathrm{aw}} + \frac{Z_{\mathrm{g}} Z_{\mathrm{t}}}{Z_{\mathrm{g}} + Z_{\mathrm{t}}}$$

where (t = tissue): $Z_t = R_t + i2\pi f L_t - iE_t/(2\pi f)$ and $Z_g = -iE_g/(2\pi f)$ (g = gas, compressibility).

The frequency-dependence of FOT parameters is modelled in the constant-phase model (Hantos et al., 1992; Peták et al., 1997). Thereby, Z_{in} is separated into airway and tissue components, since $R_t = G/(2\pi f)^\alpha$ with a frequency depencence parameter α. Suki et al. (1997) considered tissue nonlinearities in the context of this model.

At frequencies below 2 Hz mainly rheologic properties of the tissues are dominant, as well as mechanical heterogeneities. At frequencies above 100 Hz FOT obtains information about acoustic properties.

Ionescu and Keyser (2008) review other commonly used partioning models. Resistance, compliance and inertance can also be considered in terms of electrical analogs and the math-

ematical theory of the resulting equivalent electrical circuits has been considered quite generally by Smale (1972).

The recent book by Bates (2009) is a general introduction to the modeling of lung mechanics. Viscoelasticity can be incorporated into the simple model (3.7) by an additional quadratic term,

$$P_A(t) - P_0 = R_{\mathrm{rs}}\dot{V}(t) + E_1 V(t) + E_2 V^2(t).$$

Including an inertance term, this becomes a second-order nonlinear model,

$$P_A(t) - P_0 = R_{\mathrm{rs}}\dot{V}(t) + E_1 V(t) + E_2 V^2(t) + L_{\mathrm{rs}}\ddot{V}(t).$$

The inertial pressure is in counterphase with respect to elastic recoil pressure under harmonic forcing and thereby compensates the stiffness of the respiratory system. It becomes dominant over elastance at frequencies greater than 5 to 10 Hz (Peslin and Fredberg, 1986).

Similarly can the linear resistance term be replaced by a flow-dependent resistance (Rohrer, 1915; Wagers et al., 2002). In an actively breathing patient an estimate of pleural pressure P_{pl} is needed to partition between lung and chest wall characteristics. This can be measured approximately by oesophageal pressure (minimally invasive measurement by means of an oesophageal balloon). Another possibility to measure FOT signals is to use the pressure forcing generated by the heart, leading to so-called output impedance (Bijaoui et al., 2001). The problematic upper airways shunt has been studied by (Cauberghs and de Woestijne, 1989) and Bijaoui et al. (1999) discuss how to detect it and estimate impedance in its presence.

In recent years modern multifrequency FOT measurements, e.g., by impulse oscillometry (IOS) have become possible, but necessitate the use of complex, short pressure perturbations (Kuhnle et al., 2000; Klein et al., 80).

Section 3.3

Both asthma and COPD are obstructive lung diseases; under this heading fall also emphysema and chronic bronchitis (excessive mucus production). Narrowing of airways occurs in asthma, due to edema (thicking of airway walls or muscle hypertrophy), which reduces wall springiness and increases compliance.

Prevalence of COPD has been modeled by (Hoogendoorn et al., 2005). Influential projections of disease burden were published in (Murray and Lopez, 1997) and extended in (Mathers and Loncar, 2006).

Clinical application of FOT measurements is reviewed by (Goldman, 2001). LaPrad and Lutchen (2008) is a recent review of FOT with a focus on applications in asthma. (Lutchen et al., 1996) consider disease-related changes in FOT signals and criticize (the use of time averages of) single-frequency FOT in severely diseased lungs. Increased variability of Z_{rs} in asthma was reported by Que et al. (2001), but could not be confirmed later (Diba et al., 2007).

The book by Hamid et al. (2005) covers many further aspects of the physiology of healthy and diseased lungs.

Variability and fluctuation analysis is reviewed in Seely and Macklem (2004). A critical assessment of detrended fluctuation analysis was given by Maraun et al. (2004).

Section 3.5

The idea of "dynamical disease" was popularized by Glass and Mackey (1988).

Section 3.7

A further method that can be considered to highlight differences in FOT time series is bispectrum analysis (Mendel, 1991), which was pioneered in the analysis of EEG data (Barnett et al., 1971). Global dynamical models are reviewed in the recent article of Aguirre and Letellier (2009).

Approaches to detect artifacts in FOT measurements are described by Marchal et al. (2004), who also published considerations specific to measurements in young children (Marchal et al., 2005). Filtering to improve FOT estimates was discussed by Schweitzer et al. (2003).

Fluctuations in the respiratory system are discussed by Suki (2002). The model of Venegas et al. (2007) shows that exacerbations in asthma might be the result of a self-organizing cascade of ventilatory breakdowns.

The work of Bailie et al. (2009); Sassi et al. (2009) shows that heartbeat interinterval series are nonchaotic and multifractal. Wessel et al. (2009) argue that this might be caused by respiratory coupling.

Section 4.1

Since the publication of the first magnetoresonance (MR) image (Lauterbur, 1973), MR imaging has become one of the most important medical imaging methods. The mathematics behind image reconstruction in MRI is described by Twieg (1983); Ljunggren (1983).

Section 4.2

An important alternative to histogram estimates is kernel density estimation (Silverman, 1986), which results in reduced bias away from the interval mid-points. Silverman (1981) described one approaches to bump-hunting, i.e., estimation of the location of the peak of a density. Classification by likelihood ratios in this context has been considered by Silverman (1978). The choice of the bandwidth is still an issue, however. Adaptive bandwidth selection overcomes many of the problems with a single bandwidth (Sain and Scott, 1996; Sain, 2002). Kernel density estimation also offers the possibility to estimate total variation distances. In a different context this has been discussed by Schmid and Schmidt (2006). Since densities are infinite-dimensional objects, multivariate analysis of densities (discriminant analysis, PCA) needs to be based on a finite approximation. The distance-based approach is a natural way to avoid the bias and instability of histograms. A similar, popular approach is offered by kernel discriminant analysis (Shawe-Taylor and Cristianini, 2004).

Section 4.3

Yet another approach to quantitative MRI is offered by finite mixture models of parameter distributions (McLachlan and Peel, 2000; Wehrens et al., 2002), that can be fitted in the Bayesian

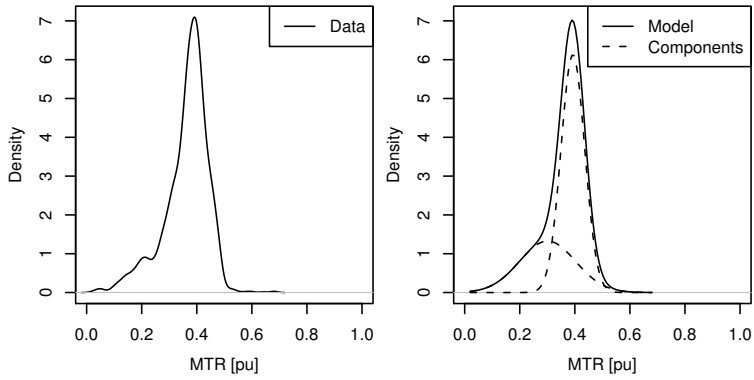


Figure D.1: Example: Fitting a two-component Gaussian mixture to an empirical magnetic transfer ratio distribution.

framework (Fraley and Raftery, 2002). Although computationally involved, this approach allows to work with substantial hypotheses and is a promising direction for future research. Figure D.1 shows an example obtained with the MCLUST software (Fraley and Raftery, 2007). For a true application, the fitting procedure needs to be extended to three-way analysis: One component should describe the background (brain) MTR distribution common to all subjects, whereas the remaining components should describe individual variations in the MTR parameter. This necessitates constrained and hierarchical fitting procedures which are not available at the moment.

The recent work of Oh and Raftery (2007) considers Bayesian clustering of Euclidean representations, and is similar to the distance-based analysis in its philosophy.

Section 4.4

The literature on Alzheimer's Disease is too numerous to review here. Good starting points are Goedert and Spillantini (2006); Roberson and Mucke (2006). Our publication (Muskulus, Scheenstra, Braakman, Dijkstra, Verduyn-Lunel, Alia, de Groot and Reiber, 2009) offers a comprehensive review of small animal models of Alzheimer's disease and mathematical approaches to its genesis and disease severity quantification. Voxel-based relaxometry (Pell et al., 2004) has gained much interest in recent years and is based on estimating significance probabilities for single voxels in a group-wise comparison. Note that the T_2 parameter should be considered multi-exponential (Whittall et al., 1999), but a large (time-consuming) number of echo times is needed to resolve this sensibly and uniquely. Further details can be found in (Whittall et al., 1991).

Section 6.1

A comprehensive introduction into neuronal science is (Kandel et al., 2000). Magnetoencephalography is reviewed by Hämäläinen et al. (1993) and the resourceful book of Nunez

and Srinivasan (2006) is a standard reference for EEG and its problems. Practical aspects of working with electrophysiological signals are covered by Sanei and Chambers (2007). A comprehensive overview of current topics in neuroscience is given by Buzsáki (2006) Approaches to brain connectivity from the viewpoint of complexity theory are reviewed in (Tononi et al., 1998; Tonini, 1998). The book by Kelso (1995) is an introduction to the theory of brain coordination dynamics, where brain activity and interactions are modelled by coupled non-linear oscillators. Recent modeling issues are discussed by Tognoli and Kelso (2009). The basic reference for the modelling of single neurons is Koch (2004), and integrative approaches are discussed in (Koch, 1998). The book by Izhikevich (2006) focusses on nonlinear approaches, i.e., the theory of dynamical systems.

Section 6.3

Functional connectivity of the brain and its relation to anatomical connectivity has been studied by Honey et al. (2007) by functional MRI. Partial coherence is a recent extension of coherence that has been applied in functional MRI to quantify directionality (Sun et al., 2004).

Synchronization approaches to brain dynamics model brain activity as self-sustained coupled non-linear oscillators. A general introduction to all aspects of synchronization is (Pikovsky et al., 2003). The Kuramoto model (Acebrón et al., 2005; Strogatz, 2000) has been very influential in this area (Cumin and Unsworth, 2007). In the time domain, Carmeli et al. (2005) introduced a bivariate synchronization measure that is based on geometrical changes (contractions) in delay embeddings.

A very different approach is considered in Albrecht and Palus (1991), where distance measures between power spectral densities are studied.

Section 6.5

Resting state connectivity in general has been studied by Greicius et al. (2003) and by Beckmann et al. (2005) in functional MRI. Effective connectivity in the context of auditory information processing is discussed by Gonçalves et al. (2001), again for functional MRI.

Section 6.6

The standard reference for the electrophysiological inverse problem is Sarvas (1987). Apart from beamforming, an important alternative approach to source location has been pioneered by Pascal-Marqui (2002). The book by Kaipio and Somersalo (2004) discusses the statistical approach to inverse problems and features a section on MEG.

Section A.1

Alternatively, and in a certain sense dual to the set-theoretic foundation of mathematics, it is possible to base all of mathematics on the notion of functional relationships, i.e., to build mathematics, and thereby also distance geometry, from the notion of a *category* (MacLane, 1985). This approach is not considered here.

Distance geometry started with the works of Menger (Menger, 1928, 1930), and the account of Blumenthal (1953) is still the canonical reference.

Section A.2

Euclidean point representations derived from multidimensional scaling can often be interpreted in terms of substantial properties. Beals et al. (1968) was influential with regard to this interpretative approach, and a more recent discussion is given by Gati and Tversky (1982).

The embedding criterion was already established by Young and Householder (1938). The double centering operation is studied in a more abstract setting by (Critchley, 1988). Gower (1966) put multidimensional scaling on a solid theoretical foundation, stressing the difference between R-matrices (coefficients-of-association between pairs of characteristics) and Q-matrices (coefficients-of-association between pairs of samples).

An introduction to nonmetric scaling procedures is given by (Borg and Groenen, 2005). Recent local embedding algorithms allow to accomodate geodesic distances, confer (Roweis and Saul, 2000; Tenenbaum et al., 2000)

For reconstruction of noisy distances, the method of (Singer, 2008) is available. More generally, the bound smoothing approach of Havel et al. (1983) allows to find optimal representations if lower and upper bounds of all pairwise distances are given. Thereby, the bound smoothing algorithm of Dress and Havel (1988) allows to respect the triangle inequality constraint optimally. In three dimensions, errors and missing entries were considered by Berger et al. (1999), and general distance matrix completion was studied by Trosset (2000).

Large scaling problems can potentially be solved by a divide and join strategy as described by Tzeng et al. (2008).

The interesting article of Laub et al. (2006) considers whether there is a use in classificatory task for negative eigendirections, arising from non-Euclidean pair-wise data.

Multidimensional scaling is essentially based on a Gram matrix (obtained by double centering), so it can be considered in the framework of kernel methods (Shawe-Taylor and Cristianini, 2004). This allows to apply the methods of kernel discriminant analysis, in particular, the recently developed support-vector machine classification strategies (Schölkopf and Smola, 2001; Hastie et al., 2008). However, these methods usually need much larger sample sizes to outperform linear discriminant analysis.

Section A.3

The multiple response permutation test can be applied in a much more general setting; these developments are described by Mielke and Berry (2007).

Although our distance-based approach was developed independently, generalized discriminant analysis on distances was first considered by Anderson and Robinson (2003). The solution of the out-of-sample problem, necessary for leave-one-out cross-validation, was obtained by Trosset and Priebe (2008). Confer (de Leeuw and Meulman, 1986) for a more general solution.

A recent issue of the *Journal of the ICRU* has been devoted entirely to receiver-operator characteristics (*Receiver operating characteristic analysis in medical imaging*, 2008). An advanced

measure of diagnostic accuracy in a probabilistic setting is the Brier score (Spiegelhalter, 1986; Redelmeier et al., 1991).

A seminal article with regard to the combination of distinct classifiers was (Xu et al., 1992); it considers averaging, voting, Bayesian methods, and Dempster-Shafer theory.

Section B.2

The optimal transportation problem was considered by Kantorovich (Kantorovich, 1942b, 1948). The discrete version is variously called the Hitchcock transportation problem (Hitchcock, 1941). In 1D it is efficiently solved by monotone rearrangement (Villani, 2003); Brandt et al. (1991) contains a derivation in the discrete case, where this simple fact is proved by duality (!). For finite points distributed on the circle, see (Cabrelli and Molter, 1995).

The "dictionary of distances" (Deza and Deza, 2006) discusses many other distance measures.

Section B.3

The Wasserstein distance is also called the Hutchinson metric (Hutchinson, 1981). In the image analysis literature it is often referred to as the "Earth mover's distance" (EMD). Its metric properties and the lower bound by the difference in means is given by Rubner et al. (2000) in the discrete case. Note that the EMD is often defined more generally, namely, for two positive measures. Its solution is then given by the optimal transport that matches the smaller measure optimally to the larger. This construction does not result in a distance, however. The behavior of the Earth mover's distance under transformations is considered by Klein and Veltkamp (2005).

The Wasserstein distance has applications as a goodness of fit test in statistics (del Barrio et al., 1999).

An elementary proof of the triangle inequality was recently given by Clement and Desch (2008).

Hoffman (1963) showed that the optimal transportation problem can be solved in linear time $O(m+n)$ if the cost coefficients fulfill the *Monge property*,

$$c_{ij} + c_{rs} \le c_{is} + c_{rj} \quad \text{for} \quad 1 \le i < r \le m, 1 \le j < s \le n.$$

This holds, for example, if $c_{ij} = u_i + v_j$. See the review by Burkard et al. (1996). This property can be generalized to Monge sequences (Alon et al., 1989), and recognition of permuted Monge matrices seems possible in ($O(mn + m \log m)$ time (confer Burkard et al. (1996) for references). It is related to the so-called *quadrangle inequality* that allows significant computational speedups in many algorithms (Yao, 1980).

A feasible solution is always available by the greedy *northwest corner* rule (Burkard, 2007). The fastest algorithm for solving minimum cost flows still seems to be Orlin's algorithm, with a running time of

$$O(n' \log n'(n' \log n' + m')),$$

where $m' = n + m$ and m' is the number of (finite) entries in the cost matrix (Orlin, 1988). The network simplex algorithm is described in detail by Kelly and O'Neill (1991).

Dell'Amico and Toth (2000) compare computational codes for the assingment code.

The Hungarian algorithm (Kuhn, 1955) runs with complexity $O(n^3)$ (Jonker and Volgenant, 1986). In the plane, (Vaidya, 1989; Atkinson and Vaidya, 1995) have obtained improvements to ($O(n^{2.5} \log n)$ for the Euclidean bipartite matching problem and $O(n^{2.5} \log n \log N)$ for the transportation problem, where N is the magnitude of the largest cost. Agarwal et al. (1999) consider an ϵ-approximation algorithm with complexity $O(n^{2+\epsilon})$. A probabilistic algorithm that results in an $1 + \epsilon$ approximation with probability at least $1/2$ has been given by Varadarajan and Agarwal (1999); it has complexity $O((n/\epsilon^3) \log^6 n)$. Kaijser (1998) considers another improvement in the plane.

If many related transportation problems need to be solved, relaxation algorithms should be considered, e.g., the Auction algorithm of Bertsekas and Castanon (1989).

Bibliography

Aban, B., Meerschaert, M. M. and Panorska, A. K.: 2006, Parameter estimation for the truncated Pareto distribution, *J. Am. Stat. Assoc.* **101**, 270–277. 66

Abarbanel, H. D. I., Brown, R., Sidorowich, J. J. and Tsimring, L. S.: 1993, The analysis of observed chaotic data in physical systems, *Rev. Mod. Phys.* **65**, 1331–1392. 15

Able, K. P. and Able, M. A.: 1997, Development of sunset orientation in a migratory bird: no calibration by the magnetic field, *Anim. Behav.* **53**, 363–368. 116

Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J. and Ritort, F.: 2005, The Kuramoto model: A simple paradigm for synchronization phenomena, *Rev. Mod. Phys.* **77**, 137–185. 234

Achard, S. and Bullmore, E.: 2007, Efficiency and cost of economical brain functional networks, *PLoS Comp. Biol.* **3**, 174–183. 151

Agarwal, P. K., Efrat, A. and Sharir, M.: 1999, Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications, *SIAM J. Comput.* **29**, 912–953. 237

Aguirre, L. A. and Letellier, C.: 2009, Modeling nonlinear dynamics and chaos: A review, *Math. Probl. Eng.* **2009**, 238960. 232

Aki, S.: 1987, On nonparametric tests for symmetry in R^m, *Ann. Inst. Statist. Math.* **39**, 457–472. 127

Albrecht, V. and Palus, M.: 1991, The spectral dynamics and its applications in EEG, *Biol. Cybern.* **66**, 71–78. 234

Ali, M. S. and Silvey, D.: 1966, A general class of coefficients of divergence of one distribution from another, *J. R. Stat. Soc. B* **28**, 131–140. 12

Alon, N., Cosares, S., Hochbaum, D. S. and Shamir, R.: 1989, An algorithm for the detection and construction of Monge sequences, *Lin. Alg. Appl.* **114/115**, 669–680. 236

Alongi, J. M. and Nelson, G. S.: 2007, *Recurrence and Topology*, American Mathematical Society, Providence. 12

Anderson, M. J. and Robinson, J.: 2003, Generalized discriminant analysis based on distances, *Aust. N. Z. J. Stat.* **45**, 301–318. 26, 140, 184, 235

Ansari-Asl, K., Senhadji, L., Bellanger, J.-J. and Wendling, F.: 2006, Quantitative analysis of linear and nonlinear methods characterizing interdependencies between brain signals, *Phys. Rev. E* **74**, 031916. 229

Arabie, P., Carroll, J. D. and DeSarbo, W. S.: 1987, *Three-way scaling and clustering*, Sage Publications, Newbury Park. 192

Atkinson, D. S. and Vaidya, P. M.: 1995, Using geometry to solve the transportation problem in the plane, *Algorithmica* **13**, 442–461. 237

Avanzolini, G., Barbini, P., Cappello, A., Cevenini, G. and Chiari, L.: 1997, A new approach for tracking respiratory mechanical parameters in real-time, *Ann. Biomed. Eng.* **25**, 154–163. 61

Bailie, R. T., Cecen, A. A. and Erkal, C.: 2009, Normal heartbeat series are nonchaotic, nonlinear, and multifractal: New evidence from semiparametric and parametric tests, *Chaos* **19**, 028503. 232

Bak, P., Tang, C. and Wiesenfeldt, K.: 1987, Self-organized criticality: An explanation of $1/f$ noise, *Phys. Rev. Lett.* **59**, 381–384. 66, 79

Balakrishnan, V. K.: 1995, *Network Optimization*, Chapman & Hall, London. 18, 154

Bapat, R. B. and Raghavan, T. E. S.: 1997, *Nonnegative matrices and applications*, Cambridge University Press, New York. 199

Barabási, A. L. and Albert, R.: 1999, Emergence of scaling in random networks, *Science* **286**, 509–512. 66

Barlow, H. B.: 1972, Single units and sensation: a neuron doctrine for perceptual psychology?, *Perception* **1**, 371–94. 150

Barnett, T. P., Johnson, L. C., Naitoh, P., Hicks, N. and Nute, C.: 1971, Bispectrum analysis of electroencephalogram signals during waking and sleeping, *Science* **172**, 401–402. 232

Bates, J. H. T.: 2009, *Lung Mechanics: An Inverse Modeling Approach*, Cambridge University Press, New York. 231

Batschelet, E.: 1981, *Circular Statistics in Biology*, Academic Press, London. 153

Batzel, J. J., Kappel, F., Schneditz, D. and Tran, H. T.: 2006, *Cardiovascular and Respiratory Systems: Modeling, Analysis, and Control*, SIAM, Philadelphia. 230

Beals, R., Krantz, D. H. and Tversky, A.: 1968, Foundations of multidimensional scaling, *Psychol. Rev.* **75**, 127–142. 235

Beckmann, C. F., DeLuca, M., Devlin, J. T. and Smith, S. M.: 2005, Investigations into resting-state connectivity using independent component analysis, *Phil. Trans. R. Soc. B* **360**, 1001–1013. 234

Benamou, J.-D., Brenier, Y. and Guittet, K.: 2002, The Monge-Kantorovich mass transfer and its computational fluid mechanics formulation, *Int. J. Numer. Meth. Fluids* **40**, 21–30. 18

Benjamini, Y. and Hochberg, Y.: 1995, Controlling the False Discovery Rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. B* **57**, 289–300. 114

Berger, B., Kleinberg, J. and Leighton, T.: 1999, Reconstructing a three-dimensional model with arbitrary errors, *J. ACM* **46**, 212–235. 235

Bertsekas, D. P.: 1991, *Linear network optimization: algorithms and codes*, The MIT Press, Cambridge. 18

Bertsekas, D. P. and Castanon, D. A.: 1989, The Auction algorithm for the transportation problem, *Annals Operations Res.* **20**, 67–96. 18, 237

Besag, J. and Clifford, P.: 1991, Sequential Monte Carlo p-values, *Biometrika* **78**, 301–304. 136

Beuter, A., Glass, L., Mackey, M. C. and Titcombe, M. S. (eds): 2003, *Nonlinear Dynamics in Physiology and Medicine*, Springer, New York. 37

Bialonski, S. and Lehnertz, K.: 2006, Identifying phase synchronization clusters in spatially extended dynamical systems, *Phys. Rev. E* **74**, 051909. 146

Bianchini, M., Gori, M. and Scarselli, F.: 2005, Inside PageRank, *ACM Trans. Internet Technology* **5**, 92–128. 147

Bijaoui, E., Baconnier, P. F. and Bates, J. H. T.: 2001, Mechanical output impedance of the lung determined from cardiogenic oscillations, *J. Appl. Physiol.* **91**, 859–865. 231

Bijaoui, E., Tuck, S. A., Remmers, J. E. and Bates, J. H. T.: 1999, Estimating respiratory mechanics in the presence of flow limitation, *J. Appl. Physiol.* **86**, 418–426. 231

Blumenthal, L. M.: 1953, *Theory and applications of distance geometry*, Oxford University Press, New York. 143, 169, 170, 235

Boccaletti, S., Kurths, J., Osipov, G., Valladares, D. L. and Zhou, C. S.: 2002, The synchronization of chaotic systems, *Phys. Rep.* **366**, 1–101. 152

Boonstra, T. W., Daffertshofer, A., Peper, C. E. and Beek, P. J.: 2006, Amplitude and phase dynamics associated with acoustically paced finger tapping, *Brain Res.* **1109**, 60–69. 152, 156

Borg, I. and Groenen, P. J. F.: 2005, *Modern Multidimensional Scaling*, Springer, New York. 3, 19, 20, 21, 22, 50, 145, 177, 235

Borwein, D. and Borwein, J. M.: 2001, Some remarkable properties of sinc and related integrals, *Ramanujan J.* **5**, 73–89. 119

Bottomley, P. A., Hardy, C. J., Argersinger, R. E. and Allen-Moore, G.: 1987, A review of ^1h nuclear magnetic resonance relaxation in pathology: Are T_1 and T_2 diagnostic?, *Med. Phys.* **14**, 1–37. 108

Brandt, J., Cabrelli, C. and Molter, U.: 1991, An algorithm for the computation of the Hutchinson distance, *Inf. Proc. Lett.* **40**, 113–117. 236

Breakspear, M. and Terry, J. R.: 2002, Detection and description of non-linear interdependence in normal multichannel human EEG data, *Clin. Neurophysiol.* **113**, 735–753. 152

Brémaud, P.: 1999, *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer, New York. 147

Brillinger, D. R.: 1981, *Time Series. Data Analysis and Theory*, Holden-Day Inc., San Francisco. 143, 151

Brin, S. and Page, L.: 1998, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks & ISDN Systems* **30**, 107–117. 147

Brockwell, P. J. and Davis, R. A.: 1998, *Time Series: Theory and Methods*, Springer, New York. 11

Bronzino, J. D. (ed.): 2006, *Biomedical Engineering Fundamentals*, Taylor & Francis, London. 230

Bullmore, E. and Sporns, O.: 2009, Complex brain networks: graph theoretical analysis of structural and functional systems, *Nature Neurosci.* **10**, 186–198. 146

Burkard, R. E.: 2007, Monge properties, discrete convexity and applications, *Europ. J. Op. Res.* **176**, 1–14. 236

Burkard, R. E., Klinz, B. and Rudolf, R.: 1996, Perspectives of Monge properties in optimization, *Disc. Appl. Math.* **70**, 95–161. 18, 236

Burton, N. J. K.: 2006, Nest orientation and hatching success in the tree pipit *Anthus trivialis*, *J. Avian Biol.* **37**, 312–317. 116

Butler, J., Caro, C. G., Alcala, R. and DuBois, A. B.: 1960, Physiological factors affecting airway resistance in normal subjects and in patients with obstructive respiratory disease, *J. Clin. Invest.* **39**, 584–591. 62

Buzsáki, G.: 2006, *Rhythms of the brain*, Oxford University Press, Oxford. 150, 234

Cabrelli, C. A. and Molter, U. M.: 1995, The Kantorovich metric for probability measures on the circle, *J. Comp. Appl. Math.* **57**, 345–361. 236

Cao, J. and Worsley, K. J.: 1999, The detection of local shape changes via the geometry of Hotelling's T^2 fields, *Annals Stat.* **27**, 925–942. 114

Cao, L., Mees, A. and Judd, K.: 1998, Dynamics from multivariate time series, *Physica D* **121**, 75–88. 15, 47

Carmeli, C., Knyazeva, M. G., Innocenti, G. M. and Feo, O. D.: 2005, Assessment of EEG synchronization based on state-space analysis, *NeuroImage* **25**, 339–354. 234

Cauberghs, M. and de Woestijne, K. P. V.: 1989, Effect of upper airways shunt and series properties on respiratory impedance measurements, *J. Appl. Physiol.* **66**, 2274–2279. 231

Chang, J. and Mosenifar, Z.: 2007, Differentiating COPD from asthma in clinical practice, *J. Intensive Care Med.* **22**, 300–309. 88

Chen, X. J., Kovacevic, N., Lobaugh, N. J., Sled, J. G., Henkelman, R. M. and Henderson, J. T.: 2005, Neuroanatomical differences between mouse strains as shown by high-resolution 3D MRI, *NeuroImage* **29**, 99–105. 114

Chernetsov, N., Chromik, W., Dolata, P. T., Profus, P. and Tryjanowski, P.: 2006, Sex-related natal disperal of white storks (Ciconia Ciconia) in Poland: How far and where to go?, *The Auk* **123**, 1103–1109. 116

Chung, M. K., Worsley, K. J., Paus, T., Cherif, C., Collins, D. L., Giedd, N., Rapoport, J. L. and Evans, A. C.: 2001, A unified statistical approach to deformation-based morphometry, *NeuroImage* **14**, 595–606. 114

Clauset, A., Shalizi, C. R. and Newman, M. E. J.: 2009, Power-law distributions in empirical data, *SIAM Rev.* **51**, 661–703. 66, 68, 85, 207, 208

Clement, P. and Desch, W.: 2008, An elementary proof of the triangle inequality for the Wasserstein metric, *Proc. Amer. Math. Soc.* **136**, 333–339. 44, 236

Constantine, A. G. and Gower, J. C.: 1978, Graphical representation of asymmetric matrices, *Appl. Stat.* **27**, 297–304. 146

Cover, T. M. and Thomas, J. A.: 1991, *Elements of Information Theory*, 1st edn, John Wiley & Sons, Chichester. 143, 148, 149

Critchley, F.: 1988, On certain linear mappings between inner-product and squared-distance matrices, *Lin. Alg. Appl.* **105**, 91–107. 23, 235

Cumin, D. and Unsworth, C. P.: 2007, Generalising the Kuramoto model for the study of neuronal synchronisation in the brain, *Physica D* **226**, 181–196. 234

da Costa, L. F. and Barbosa, M. S.: 2004, An analytical approach to neuron connectivity, *Eur. Phys. J. B* **42**, 573–580. 146, 147

Daróczy, B. and Hantos, Z.: 1982, An improved forced oscillatory estimation of respiratory impedance, *Int. J. Bio. Med. Comp.* **13**, 221–235. 62

Darvas, F., Ojemann, J. G. and Sorensen, L. B.: 2009, Bi-phase locking — a tool for probing non-linear interactions in the human brain, *NeuroImage* **46**, 123–132. 153

David, O., Cosmellia, D. and Friston, K. J.: 2004, Evaluation of different measures of functional connectivity using a neural mass model, *NeuroImage* **21**, 659–673. 147

Davidson, R. N., Greig, C. A., Hussain, A. and Saunders, K. B.: 1986a, Withing-breath changes of airway calibre in patients with airflow obstruction by continuous measurement of respiratory impedance, *Br. J. Dis. Chest* **80**, 335–352. 62

Davison, A. C. and Hinkley, D. V.: 1997, *Bootstrap Methods and their Applications*, Cambridge University Press, New York. 19, 114, 128, 155, 182

de Leeuw, J. and Meulman, J.: 1986, A special jackknife for multidimensional scaling, *J. Classification* **3**, 97–112. 166, 235

Deco, G., Schittenkopf, C. and Schürmann, B.: 1997, Determining the information flow of dynamical systems from continuous probability distributions, *Phys. Rev. Lett.* **78**, 2345–2348. 149

Dehmeshki, J., Buchem, M. A. V., Bosma, G. P. T., Huizinga, T. W. J. and Tofts, P. S.: 2002, Systemic lupus erythematosus: diagnostic application of magnetization transfer ratio histograms in patients with neuropsychiatric symptoms — initial results, *Radiology* **222**, 722–728. 96, 105

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M.: 1999, Tests of goodness of fit based on the l_2-Wasserstein distance, *Ann. Stat.* **27**, 1230–1239. 236

Delavault, E., Saumon, G. and Georges, R.: 1980, Characterization and validation of forced input methods for respiratory measurement, *Resp. Physiol.* **40**, 119–136. 62

Dell'Amico, M. and Toth, P.: 2000, Algorithms and codes for dense assignment problems: the state of the art, *Disc. Appl. Math.* **100**, 17–48. 236

Dellnitz, M. and Junge, O.: 1999, On the approximation of complicated dynamical behavior, *SIAM J. Numer. Anal.* **36**, 491–515. 16

Deza, E. and Deza, M. M.: 2006, *Dictionary of distances*, Elsevier, Amsterdam. 236

Dhamala, M., Rangarajan, G. and Ding, M.: 2008a, Analyzing information flow in brain networks with nonparametric Granger causality, *Neuroimage* **41**, 354–62. 150

Dhamala, M., Rangarajan, G. and Ding, M.: 2008b, Estimating Granger causality from Fourier and wavelet transforms of time series data, *Phys. Rev. Lett.* **100**, 018701. 150

Diba, C., Salome, C. M., Reddel, H. K., Thorpe, C. W., Toelle, B. and King, G. G.: 2007, Short-term variability of airway caliber – a marker of asthma?, *J. Appl. Physiol.* **103**, 296–304. 75, 231

Diks, C. and Tong, H.: 1999, A test for symmetries of multivariate probability distributions, *Biometrika* **86**, 605–614. 121, 127, 128, 217, 219

Diks, C., van Zwet, W. R., Takens, F. and DeGoede, J.: 1996, Detecting differences between delay vector distributions, *Phys. Rev. E* **53**, 2169–2176. 12

Dress, A. W. M. and Havel, T. F.: 1988, Shortest-path problems and molecular conformation, *Disc. Appl. Math.* **19**, 129–144. 235

DuBois, A. B., Brody, A. W., Lewis, D. H. and Burgess, B. F.: 1956a, Oscillation mechanics of lungs and chest in man, *J. Appl. Physiol.* **8**, 587–594. 230

Dutka, J.: 1985, On the problem of random flights, *Arch. Hist. Exact Sci.* **32**, 351–375. 115, 119

Dvorák, S.: 1972, Treolar's distribution and its numerical implementation, *J. Phys. A* **5**, 78–84. 119

Eckmann, J.-P. and Ruelle, D.: 1985, Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.* **57**, 617–656. 73

Efron, B.: 1981, The jackknife, the bootstrap, and other resampling plans, *Biometrika* **68**, 589–599. 182

Eggermont, J. J.: 1998, Is there a neural code?, *Neurosci. Biobehav. Rev.* **22**, 355–70. 150

Eichler, M.: 2007, Granger causality and path diagrams for multivariate time series, *J. Econometrics* **137**, 334–354. 150

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B. and Singer, W.: 1992, Temporal coding in the visual cortex: new vistas on integration in the nervous system, *TINS* **15**, 218–226. 150

Fadel, P. J., Barman, S. M., Phillips, S. W. and Gebber, G. L.: 2004, Fractal fluctuations in human respiration, *J. Appl. Physiol.* **97**, 2056–2064. 87

Falangola, M. F., Ardekani, B. A., Lee, S.-P., Babb, J. S., Bogart, A., Dyakin, V. V. et al.: 2005, Application of a non-linear image registration algorithm to quantitative analysis of T_2 relaxation time in transgenic mouse models of AD pathology, *J. Neurosci. Meth.* **144**, 91–97. 109

Fay, M. P., Kim, H.-J. and Hachey, M.: 2007, On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests, *J. Comput. Graph. Stat.* **16**, 946–967. 136

Fernández, V. A., Garmero, M. D. J. and Garía, J. M.: 2008, A test for the two-sample problem based on empirical characteristic functions, *Comput. Stat. Data Anal.* **52**, 3730–3748. 127

Fingelkurts, A. A., Fingelkurts, A. A. and Kähkönen, S.: 2005, Functional connectivity in the brain — is it an elusive concept?, *Neurosci. Biobehav. Rev.* **28**, 827–836. 140

Fisher, R. A.: 1936, The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188. 187

Fraley, C. and Raftery, A. E.: 2002, Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Stat. Assoc.* **97**, 611–631. 233

Fraley, C. and Raftery, A. E.: 2007, Model-based methods of classification: Using the mclust software in chemometrics, *J. Stat. Software* **18**, 1–13. 233

Frankel, T.: 1997, *The Geometry of Physics*, Cambridge University Press, New York. 50

Fraser, A. M. and Swinney, H. L.: 1986, Independent coordinates for strange attractors from mutual information, *Phys. Rev. A* **33**, 1134–1140. 15

Frey, U., Brodbeck, T., Majumdar, A., Taylor, D. R., Town, G. I., Silverman, M. and Suki, B.: 2005, Risk of severe asthma episodes predicted from fluctuation analysis of airway function, *Nature* **438**, 667–670. 65

Frey, U. and Suki, B.: 2008, Complexity of chronic asthma and chronic obstructive pulmonary disease: implications for risk assessment, and disease progression and control, *The Lancet* **372**, 1088–1099. 65

Friedrich, R. and Peinke, J.: 1997, Description of a turbulent cascade by a Fokker-Planck equation, *Phys. Rev. Lett.* **78**, 863–866. 88

Frigyik, B. A., Srivastava, S. and Gupta, M. R.: 2008, Functional Bregman divergence and Bayesian estimation of distributions, *IEEE Trans. Inf. Theory* **54**, 5130–5139. 12

Frisch, U., Matarrese, S., Mohayaee, R. and Sobolevski, A.: 2002, A reconstruction of the initial conditions of the universe by optimal mass transportation, *Nature* **417**, 260–262. 17, 154

Friston, K. J.: 1997, Another neural code?, *Neuroimage* **5**, 213–20. 150

Friston, K. J., Frith, C. D., Fletcher, P., Liddle, P. F. and Frackowiak, R. S. J.: 1996, Functional topography: Multidimensional scaling and functional connectivity in the brain, *Cereb. Cortex* **6**, 156–164. 140, 144, 158

Friston, K. J., Frith, C. D., Liddle, P. F. and Frackowiak, R. S. J.: 1993, Time-dependent changes in effective connectivity measures with PET, *Hum. Brain Mapp.* **1**, 69–80. 139

Gangbo, W. and McCann, R.: 2000, Shape recognition via Wasserstein distance, *Quarterly Appl. Math.* **4**, 705–737. 17

Gastwirth, J. L.: 1965, Percentile modifications of two sample rank tests, *J. Amer. Stat. Assoc.* **60**, 1127–1141. 116

Gati, I. and Tversky, A.: 1982, Representations of qualitative and quantitative dimensions, *J. Exp. Psychol.* **8**, 325–340. 235

Geweke, J.: 1982, Measurement of linear dependence and feedback between multiple time series, *J. Am. Stat. Assoc.* **77**, 304–313. 143, 152

Gibson, P. G. and Simpson, J. L.: 2009, The overlap syndrome of asthma and COPD: what are its features and how important is it?, *Thorax* **64**, 728–735. 88

Gill, P. M. W.: 2007, Efficient calculation of p-values in linear-statistic permutation significance tests, *J. Stat. Comput. Sim.* **77**, 55–61. 136

Gillis, H. L. and Lutchen, K. R.: 1999, Airway remodelling in asthma amplifies heterogeneities in smooth muscle shortening causing hyperresponsiveness, *J. Appl. Physiol.* **86**, 2001–2012. 62

Glass, L. and Mackey, M. C.: 1988, *From Clocks to Chaos*, Princeton University Press, Princeton. 37, 232

Global Initiative for Asthma: 2009, Gina report, Global strategy for asthma management and prevention. Online [accessed September 2009].
URL: *http://www.ginasthma.com/* 63, 64

Global Initiative for Chronic Obstructive Pulmonary Disease: 2009, Global strategy for diagnosis, management, and prevention of COPD. Online [accessed September 2009].
URL: *http://www.goldcopd.com* 63, 64

Goedert, M. and Spillantini, M. G.: 2006, A century of Alzheimer's disease, *Science* **314**, 777–781. 233

Goldberg, D.: 1991, What every computer scientist should know about floating-point arithmetic, *ACM Comp.* **23**, 5–48. 136

Goldberger, A. L., Amaral, L. A., Hausdorff, J. M., Ivanov, P. C., Peng, C. K. and Stanley, H. E.: 2002, Fractal dynamics in physiology: alterations with disease and aging, *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 1**, 2466–2472. 87

Goldman, M. D.: 2001, Clinical applications of forced oscillation, *Pulm. Pharmacol. Ther.* **14**, 341–350. 231

Gonçalves, M. S., Hall, D. A., Johnsrude, I. S. and Haggard, M. P.: 2001, Can meaningful effective connectivites be obtained between auditory cortical regions?, *NeuroImage* **14**, 1353–1360. 234

Goodhill, G. J., Simmen, M. W. and Willshaw, D. J.: 1995, An evaluation of the use of multidimensional scaling for understanding brain connectivity, *Phil. Trans. R. Soc. Lond. B* **348**, 265–280. 140

Gouesbet, G. and Maquet, J.: 1992, Construction of phenomenological models from numerical scalar time series, *Physica D* **58**, 202–215. 89

Gower, J. C.: 1966, Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338. 145, 176, 235

Granger, C. W. J.: 1969, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**, 424–438. 143, 149, 151

Grassberger, P. and Procaccia, I.: 1983, Estimating the Kolmogorov entropy from a chaotic signal, *Phys. Rev. A* **28**, 2591–2593. 73

Gray, C. M.: 1994, Synchronous oscillations in neuronal systems: Mechanisms and functions, *J. Comp. Neurosci.* **1**, 11–38. 150

Gray, R. M.: 1990, *Entropy and Information Theory*, Springer, New York. 143, 148

Greicius, M. D., Krasnow, B., Reiss, A. L. and Menon, V.: 2003, Functional connectivity in the resting brain: A network analysis of the default mode hypothesis, *Proc. Natl. Acad. Sci. U. S. A.* **100**, 253–258. 234

Grodins, F. S., Buell, J. and Bart, A. J.: 1967a, Mathematical analysis and digital simulation of the respiratory control system, *J. Appl. Physiol.* **22**, 260–276. 230

Guerra, S.: 2005, Overlap of asthma and chronic obstructive pulmonary disease, *Curr. Opin. Pulm. Med.* **11**, 7–13. 64, 88

Guyton, A. C. and Hall, J. E.: 2006, *Textbook of Medical Physiology*, 11th edn, Elsevier, Philadelphia. 58, 230

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. and Lounasmaa, O. V.: 1993, Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain, *Rev. Mod. Phys.* **65**, 413–505. 233

Haker, S., Zhu, L., Tannenbaum, A. and Angenent, S.: 2004, Optimal mass transport for registration and warping, *Int. J. Computer Vision* **60**, 225–240. 17, 18, 154

Hamid, Q., Shannon, J. and Martin, J. (eds): 2005, *Physiologic Basis of Respiratory Disease*, BC Decker, Hamilton. 231

Hanley, J. A. and McNeil, B. J.: 1982, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143**, 29–36. 191

Hantos, Z., Darōczy, B., Suki, B., Nagy, S. and Fredberg, J. J.: 1992, Input impedance and peripheral inhomogeneity of dog lungs, *J Appl Physiol* **72**, 168–178. 230

Härdle, W. and Simar, L.: 2003, *Applied Multivariate Statistical Analysis*, Springer, New York. 5, 21

Hastie, T., Tibshirani, R. and Friedman, J.: 2008, *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn, Springer, New York. 235

Havel, T. F., Kuntz, I. D. and Crippen, G. M.: 1983, The theory and practive of distance geometry, *Bull. Math. Biol.* **45**, 665–720. 24, 144, 169, 173, 174, 175, 176, 235

Heiser, W. J. and Meulman, J.: 1983, Analyzing rectangular tables by joint and constrained multidimensional scaling, *J. Econometrics* **22**, 139–167. 24, 145

Helen, J. L., Watkins, K. E., Bishop, D. V. M. and Matthews, P. M.: 2006, Hemispheric specialization for processing auditory nonspeech stimuli, *Cereb. Cortex* **16**, 1266–1275. 160

Hénon, M.: 1976, A two-dimensional mapping with a strange attractor, *Commun. Math. Physics* **50**, 69–77. 28

Henze, N., Klar, B. and Meintanis, S. G.: 2003, Invariant tests for symmetry about an unspecified point based on the empirical characteristic function, *J. Multivariate Anal.* **87**, 275–297. 127

Herman, I. P.: 2007, *Physics of the Human Body*, Springer, New York. 59, 230

Hill, B. M.: 1975, A simple general approach to inference about the tail of a distribution, *Ann. Stat.* **3**, 1163–1174. 66

Hitchcock, F. L.: 1941, The distribution of a product from several sources to numerous localities, *J. Math. Phys.* **20**, 224–230. 236

Hively, L. M., Gailey, P. C. and Protopopescu, V. A.: 1999, Detecting dynamical change in nonlinear time

series, *Phys. Lett. A* **258**, 103–114. 12

Hoffman, A. J.: 1963, On simple linear programming problems, *in* V. Klee (ed.), *Proceedings of Symposia in Pure Mathematics, Convexity*, Vol. VII, American Mathematical Society, Providence, pp. 317–327. 236

Honey, C. J., Kötter, R., Breakspear, M. and Sporns, O.: 2007, Network structure of cerebral cortex shapes functional connectivity on multiple time scales, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10240–10245. 234

Hoogendoorn, M., van Mölken, M. P. M. H. R., Hoogenveen, R. T., van Genugten, M. L. L., Buist, A. S., Wouters, E. F. M. and Feenstra, T. L.: 2005, A dynamic population model of disease progression in COPD, *Eur. Respir. J.* **26**, 223–233. 231

Horsfield, K., Dart, G., Olson, D. E., Filley, G. F. and Cumming, G.: 1971, Models of the human bronchial tree, *J. Appl. Physiol.* **31**, 207–217. 230

Horwitz, B.: 2003, The elusive concept of brain connectivity, *NeuroImage* **19**, 466–470. 140

Hotelling, H.: 1931, The generalization of Student's ratio, *Ann. Math. Statist.* **2**, 360–378. 121

Houweling, S., Daffertshofer, A., van Dijk, B. and Beek, P.: 2008, Neural changes induced by learning a challenging perceptual-motor task, *NeuroImage* **41**, 1395–1407. 153, 155, 156, 159

Huberty, C. J. and Olejnik, S.: 2006, *Applied MANOVA and discriminant analysis*, John Wiley & Sons, Hoboken. 27

Huke, J. P. and Broomhead, D. S.: 2007, Embedding theorems for non-uniformly sampled dynamical systems, *Nonlinearity* **20**, 2205–2244. 229

Hutchinson, J.: 1981, Fractals and self-similarity, *Indiana Univ. Math. J.* **30**, 713–747. 236

Hutt, A., Daffertshofer, A. and Steinmetz, U.: 2003, Detection of mutual phase synchronization in multivariate signals and application to phase ensembles and chaotic data, *Phys. Rev. E* **68**, 036219. 153

Imre, A.: 1999, Ideas in theoretical biology — comment about the fractality of the lung, *Acta Biotheor.* **47**, 79–81. 230

Ioannides, A. A.: 2007, Dynamic functional connectivity, *Curr. Opin. Neurobiol.* **17**, 161–170. 146

Ioannidis, J. P. A.: 2005, Why most published research findings are false, *PLoS Medicine* **2**, e124. xv

Ionescu, C. and Keyser, R.: 2008, Parametric models for characterizing respiratory input impedance, *J. Med. Eng. & Technology* **32**, 315–324. 230

Izhikevich, E. M.: 2006, *Dynamical systems in neuroscience: The geometry of excitability and bursting*, MIT Press, Cambridge. 234

Jackson, A. C., Lutchen, K. R. and Dorkin, H. L.: 1987, Inverse modeling of dog airway and respiratory system impedances, *J. Appl. Physiol.* **62**, 2273–2282. 230

Jonker, R. and Volgenant, T.: 1986, Improving the Hungarian assignment algorithm, *Op. Res. Lett.* **5**, 171–175. 237

Jupp, P. E.: 1987, A nonparametric correlation coefficient and a two-sample test for random vectors or directions, *Biometrika* **4**, 887–890. 127

Kadtke, J.: 1995, Classification of highly noisy signals using global dynamical models, *Phys. Lett. A* **203**, 196–202. 229

Kaijser, T.: 1998, Computing the Kantorovich distance for images, *J. Math. Imag. Vision* **9**, 173–191. 237

Kaipio, J. and Somersalo, E.: 2004, *Statistical and computational inverse problems*, Springer, New York. 234

Kajikawa, Y. and Hackett, T. A.: 2005, Entropy analysis of neuronal spike train synchrony, *J. Neurosci. Meth.* **149**, 90–93. 116

Kamiński, M., Ding, M., Truccolo, W. A. and Bressler, S. L.: 2001, Evaluating causal relations in neural systems: Granger causal, directed transfer functions and statistical assessment of significance, *Biol. Cybern.* **85**, 145–157. 149

Kandel, E. R., Schwartz, J. H. and Jessell, T. M.: 2000, *Principles of Neural Science*, McGraw-Hill, Columbus.

160, 233

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A. and Stanley, H. F.: 2002, Multifractal detrended fluctuation analysis of nonstationary time series, *Physica A* **316**, 87–114. 69

Kantorovich, L. V.: 1942a, On the transfer of masses, *C. R. (Dokl.) Acad. Sci. URSS* **37**, 227–229. 16

Kantorovich, L. V.: 1942b, On the translocation of masses, *C. R. (Dokl.) Acad. Sci. URSS* **37**, 199–201. 16, 236

Kantorovich, L. V.: 1948, On a problem of Monge, *Uspekhi Mathematicheskikh Nauk* **3**, 225–226. 236

Kantz, H. and Schreiber, T.: 2004, *Nonlinear Time Series Analysis*, Cambridge University Press, New York. 15, 88, 152

Keener, J. and Sneyd, J.: 1998, *Mathematical Physiology*, Springer, New York. 230

Kelly, D. J. and O'Neill, G. M.: 1991, *The Minimum Cost Flow Problem and The Network Simplex Solution Method*, Dissertation, University College, Dublin. 236

Kelso, J. A. S.: 1995, *Dynamic patterns: The self-organization of brain and behavior complex adaptive systems*, MIT Press, Cambridge. 234

Kennel, M. B., Brown, R. and Abarbanel, H. D. I.: 1992, Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Phys. Rev. A* **45**, 3403. 73, 229

Kingman, J. F. C.: 1963, Random walks with spherical symmetry, *Acta Math.* **63**, 11–53. 127

Kitaoka, H., Takaki, R. and Suki, B.: 1999, A three-dimensional model of the human airway tree, *J. Appl. Physiol.* **87**, 2207–2217. 230

Klein, C., Smith, H.-J. and Reinhold, P.: 80, The use of impulse oscillometry for separate analysis of inspiratory and expiratory impedance parameters in horses: Effects of sedation with xylazine, *Res. Vet. Sci.* **2006**, 201–208. 231

Klein, O. and Veltkamp, R. C.: 2005, Approximation algorithms for the Earth Mover's Distance under transformations using reference points, *Technical Report UU-CS-2005-003*, Institute of Information and Computing Sciences, Utrecht University. 45, 46, 236

Knottnerus, J. A. and Muris, J. W.: 2003, Assessment of the accuracy of diagnostic tests: the cross-sectional study, *J Clin Epidemiol* **56**, 1118–1128. 88, 89

Koch, C.: 1998, *Methods in Neuronal Modeling: From Ions to Networks*, MIT Press, Cambridge. 234

Koch, C.: 2004, *Biophysics of computation: Information processing in single neurons*, Oxford University Press, Oxford. 234

Korte, B. and Vygen, J.: 2007, *Combinatorial optimization: Theory and algorithms*, Springer, New York. 199

Kötter, R. and Stephan, K. E.: 2003, Network participation indices: characterizing component roles for information processing in neural networks, *Neural Networks* **16**, 1261–1275. 146

Kovacevic, N., Chen, J., Sled, J. G., Henderson, J. T. and Henkelman, R. M.: 2004, Deformation based representation of groupwise average and variability, *in* C. Barillot, D. R. Haynor and P. Hellier (eds), *MICCAI (1)*, Vol. 3216 of *Lect. Notes Comput. Sci.*, Springer, New York, pp. 616–622. 114

Kraskov, A., Stögbauer, H. and Grassberger, P.: 2004, Estimating mutual information, *Phys. Rev. E* **69**, 066138. 143

Kreuz, T., Mormann, F., Andrzejak, R. G., Kraskov, A., Lehnertz, K. and Grassberger, P.: 2007, Measuring synchronization in coupled model systems: a comparison of different approaches, *Physica D* **225**, 29–42. 152

Kruskal, J. B.: 1964, Nonmetric multidimensional scaling: A numerical method, *Psychometrika* **29**, 115–129. 52

Kuhn, H.: 1955, The Hungarian method for the assignment problem, *Naval Research Logistics Quarterly* **2**, 83–97. 237

Kuhnle, G. E. H., Brandt, T., Roth, U., Goetz, A. E., Smith, H.-J. and Peter, K.: 2000, Measurement of

respiratory impedance by Impulse Oscillometry — effects of endotracheal tubes, *Res. Exp. Med.* **200**, 17–26. 231

Kulish, V. (ed.): 2006, *Human Respiration: Anatomy and Physiology, Mathematical Modeling, Numerical Simulation and Applications*, WIT Press, Ashurst. 230

Kullback, S. and Leibler, R. A.: 1951, On information and sufficiency, *Ann. Math. Stat.* **22**, 79–86. 148

Kus, R., Kaminski, M. and Blinowska, K. J.: 2004, Determination of EEG activity propagation: pair-wise versus multichannel estimate., *IEEE Trans. Biomed. Eng.* **51**, 1501–1510. 140, 165

Lachaux, J.-P., Rodriguez, E., Martinerie, J. and Varela, F. J.: 1999, Measuring phase synchrony in brain signals, *Hum. Brain Mapp.* **8**, 194–208. 143, 152

Landser, F. J., Clement, J. and de Woestijne, K. P. V.: 1982, Normal values of total respiratory resistance and reactance determined by forced oscillations: influence of smoking, *Chest* **81**, 586–591. 230

Langville, A. N. and Meyer, C. D.: 2004, Deeper inside PageRank, *Internet Math.* **1**, 335–380. 147

LaPrad, A. S. and Lutchen, K. R.: 2008, Respiratory impedance measurements for assessment of lung mechanics: Focus on asthma, *Respir. Physiol. Neurobiol.* **163**, 64–73. 231

Lasota, A. and Mackey, M. C.: 1997, *Chaos, Fractals and Noise — Stochastic Aspects of Dynamics*, 2nd edn, Springer, New York. 16

Laub, J., Roth, V., Buhrmann, J. M. and Müller, K.-R.: 2006, On the information and representation of non-Euclidean pairwise data, *Patt. Recog.* **36**, 1815–1826. 20, 235

Lauterbur, P. C.: 1973, Image formation by induced local interactions: Examples employing nuclear magnetic resonance, *Nature* **242**, 190–191. 232

Lee, L., Harrison, L. M. and Mechelli, A.: 2003, A report of the functional connectivity workshop, Dusseldorf 2002, *NeuroImage* **19**, 457–465. 139

Legendre, P. and Legendre, L.: 1998, *Numerical Ecology*, Elsevier, Amsterdam. 3, 21

Lehmann, E. L. and Romano, J. P.: 2005, *Testing statistical hypotheses*, Springer, New York. 114, 128

Li, K., Guo, L., Nie, J., Li, G. and Liu, T.: 2009, Review of methods for functional brain connectivity detection using fMRI, *Comp. Med. Imag. Graph.* **33**, 131–139. 147

Ljunggren, S.: 1983, A simple graphical representation of fourier-based imaging methods, *J. Mag. Res.* **54**, 338–343. 232

Löbel, A.: 1996, Solving large-scale real-world minimum-cost flow problems by a network simplex method, *Technical report*, Konrad-Zuse Zentrum für Informationstechnik Berlin (ZIB). Software available at `http://www.zib.de/Optimization/Software/Mcf/`. 18, 19, 154, 201

Lord, R. D.: 1954, The use of Hankel transformations in statistics. I. General theory and examples, *Biometrika* **41**, 44–55. 117

Lutchen, K. R., Greenstein, J. L. and Suki, B.: 1996, How inhomogeneities and airway walls affect frequency dependence and separation of airway and tissue properties, *J. Appl. Physiol.* **80**, 1696–1707. 231

Lutchen, K. R., Sullivan, A., Arbogast, F. T., Celli, B. R. and Jackson, A. C.: 1998, Use of transfer impedance measurements for clinical assessment of lung mechanics, *Am. J. Respir. Crit. Care Med.* **157**, 435–446. 38, 61

Lütkepohl, H.: 2005, *New Introduction to Multiple Time Series Analysis*, Springer, New York. 150

Mackey, M. C. and Milton, J. G.: 1987, Dynamical diseases, *Ann. N. Y. Acad. Sci.* **504**, 16–32. 37

MacLane, S.: 1985, *Mathematics: Form and Function*, Springer, New York. 234

MacLeod, D. and Birch, M.: 2001, Respiratory input impedance measurement: forced oscillation methods, *Med. Bio. Eng. Comput.* **39**, 505–516. 230

Maindonald, J. and Braun, J.: 2003, *Data Analysis and Graphics Using R. An Example-based Approach*, Cambridge University Press, New York. 25

Maraun, D., Rust, H. W. and Timmer, J.: 2004, Tempting long-memory — on the interpretation of DFA results, *Nonlin. Proc. Geophys.* **11**, 495–504. 231

Marchal, F., Schweitzer, C., Demoulin, B., Choné, C. and Peslin, R.: 2004, Filtering artifacts in measurements of forced oscillation respiratory impedance in young children, *Physiol. Meas.* **25**, 1–14. 232

Marchal, F., Schweitzer, C. and Thuy, L. V. T.: 2005, Forced oscillations, interruptor technique and body plethysmography in the preschool child, *Paediatric Resp. Rev.* **6**, 278–284. 232

Mardia, K. V. and Jupp, P. E.: 2000, *Directional statistics*, John Wiley & Sons, New York. 116, 122, 152, 153

Marwan, N., Romand, M. C., Thiel, M. and Kurths, J.: 2007, Recurrence plots for the analysis of complex systems, *Phys. Rep.* **438**, 237–239. 152, 229

Mathers, C. D. and Loncar, D.: 2006, Projections of global mortality and burden of disease from 2002 to 2030, *PLoS Medicine* **3**, e442. 231

Matsumoto, K. and Tsuda, I.: 1988, Calculation of information flow rate from mutual information, *J. Phys. A* **21**, 1405–1414. 149

McCall, C. B., Hyatt, R. E., Noble, F. W. and Fry, D. L.: 1957, Harmonic content of certain respiratory flow phenomena of normal individuals, *J. Appl. Physiol.* **10**, 215–218. 62

McLachlan, G. J. and Peel, D.: 2000, *Finite Mixture Models*, Wiley, New York. 232

Mcnaught, M. K. and Owens, I. P. F.: 20002, Interspecific variation in plumage colour among birds: species recognition or light enviroment?, *J. Evolution. Biol.* **15**, 505–514. 116

Meho, L. I.: 2007, The rise and rise of citation analysis, *Physics World* **20**, 32–36. xv

Mendel, J. M.: 1991, Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications, *Proc. IEEE* **79**, 278–305. 232

Menger, K.: 1928, Untersuchungen über allgemeine Metrik, *Mathematische Annalen* **100**, 75–163. 235

Menger, K.: 1930, Untersuchungen über allgemeine Metrik. Vierte Untersuchung. Zur Metrik der Kurven., *Mathematische Annalen* **103**, 466–501. 235

Meulman, J., Heiser, W. J. and Leeuw, J. D.: 1983, Progress notes on SMACOF-II, *Technical Report UG-83*, Department of Data Theory, Leiden University. 52

Michaelson, E. D., Grassman, E. D. and Peters, W. R.: 1975, Pulmonary mechanics by spectral analysis of forced random noise, *J. Clin. Invest.* **56**, 1210–1230. 230

Mielke, P. W. and Berry, K. J.: 2007, *Permutation methods: A distance function approach*, 2nd edn, Springer, New York. 17, 27, 235

Milnor, J.: 1985, On the concept of attractor, *Commun. Math. Phys.* **99**, 177–195. 12

Mladineo, R. H.: 1986, An algorithm for finding the global maximum of a multimodal, multivariate function, *Math. Program.* **34**, 188–200. 50

Moeckel, R. and Murray, B.: 1997, Measuring the distance between time series, *Physica D* **102**, 187–194. 12, 13, 16, 19, 143, 154

Moler, C. and Loan, C. V.: 1978, Nineteen dubious ways to compute the exponential of a matrix, *SIAM Rev.* **20**, 801–836. 50

Moore, B. R.: 1980, A modification of the Rayleigh test for vector data, *Biometrika* **67**, 175–180. 115, 116, 127, 128

Mormann, F., Lehnertz, K., David, P. and Elger, C. E.: 2000, Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients, *Physica D* **144**, 358–369. 143, 152

Murray, C. J. L. and Lopez, A. D.: 1997, Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study, *Lancet* **349**, 1498–1504. 231

Muskulus, M., Houweling, S., Verduyn-Lunel, S. and Daffertshofer, A.: 2009, Functional similarities and distance properties, *J. Neurosci. Meth.* **183**, 31–41. 52

Muskulus, M., Scheenstra, A. E. H., Braakman, N., Dijkstra, J., Verduyn-Lunel, S., Alia, A., de Groot, H. J. M. and Reiber, J. H. C.: 2009, Prospects for early detection of Alzheimer's disease from serial MR images in transgenic mouse models, *Curr. Alz. Res.* **6**, 503–518. 52, 108, 233

Muskulus, M. and Verduyn-Lunel, S.: 2008a, The analysis of dynamical diseases by optimal transportation distances, *ERCIM News* **73**, 16–17. 37, 155

Muskulus, M. and Verduyn-Lunel, S.: 2008b, Reconstruction of functional brain networks by Wasserstein distances in a listening task, *in* R. Kakigi, K. Yokosawa and S. Kuriki (eds), *Biomagnetism: Interdisciplinary Research and Exploration*, Hokkaido University Press, Sapporo, pp. 59–61. 143, 155

Næs, T. and Mevik, B.: 2000, Understanding the collinearity problem in regression and discriminant analysis, *J. Chemomet.* **15**, 413–426. 25, 145

Nawroth, A. P., Peinke, J., Kleinhans, D. and Friedrich, R.: 2007, Improved estimation of Fokker-Planck equations through optimization, *Phys. Rev. E* **76**, 056102. 88

Nelder, J. A. and Mead, R.: 1965, A simplex method for function minimization, *Comput. J.* **7**, 308–313. 27, 49, 202

Newman, M. E. J.: 2005, Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.* **46**, 323–351. 66

Ngatchou-Wandji, J.: 2009, Testing for symmetry in multivariate distributions, *Stat. Methodol.* **6**, 230–250. 127

Nichols, T. and Holmes, A.: 2007, Non-parametric procedures, *in* K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols and W. D. Penny (eds), *Statistical parametric mapping. The analysis of functional brain images*, Elsevier, Amsterdam, chapter 21, pp. 253–272. 114, 137

Nolan, J. P.: 2010, *Stable distributions — models for heavy tailed data*, Birkhäuser, New York. In progress, Ch. 1 online [accessed 03-12-2009].
URL: *http://academic2.american.edu/ jpnolan* 102

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S. and Hallett, M.: 2004, Identifying true brain interactions from EEG data using the imaginary part of coherency, *Clin. Neurophysiol.* **115**, 2292–2307. 143, 151

Nolte, G., Ziehe, A., Nikulin, V., Schlögl, A., Krämer, N., Brismar, T. and Müller, K. R.: 2008, Robustly estimating the flow direction of information in complex physical systems, *Phys. Rev. Lett.* **100**, 234101. 143, 151

Nordhausen, K., Sirkia, S., Oja, H. and Tyler, D. E.: 2007, IC]SNP: tools for multivariate nonparametrics, R package version 1.0-2. Online [accessed September 2009].
URL: *http://cran.r-project.org/web/packages/ICSNP/index.html* 121

Nucci, G. and Cobelli, C.: 2001, Mathematical models of respiratory mechanics, *in* E. Carson and C. Cobelli (eds), *Modelling Methodology for Physiology and Medicine*, Academic Press, London, chapter 10, pp. 279–304. 230

Nunez, P. L. and Srinivasan, R.: 2006, *Electric fields of the brain: the neurophysics of EEG*, Oxford University Press, New York. 233

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silverstein, R. B. and Cadusch, P. J.: 1997, EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales, *Electrograph. Clin. Neurophysiol.* **103**, 499–515. 151

Oh, M.-S. and Raftery, A. E.: 2007, Model-based clustering with dissimilarities, *J. Comp. Graphical Stat.* **16**, 559–585. 233

Okada, A. and Imaizumi, T.: 1987, Geometric models for asymmetric similarity data, *Behaviormetrika* **21**, 81–96. 146

Oostveen, E., MacLeod, D., Lorina, H., Farre, R., Hantos, Z., Desager, K. and Marchal, F.: 2003, The forced oscillation technique in clinical practice: methodology, recommendations and future developments, *Eur. Respir. J.* **22**, 1026–1041. 38, 230

Oostveen, E., Peslin, R., Gallina, C. and Zwart, A.: 1986, Flow and volume dependence of respiratory mechanical properties studied by forced oscillation, *J. Appl. Physiol.* **67**, 2212–2218. 62

Orlin, J. B.: 1988, A faster strongly polynomial minimum cost flow algorithm, *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, ACM Press, New York, pp. 377–387. 236

Ottesen, J. T., Olufsen, M. S. and Larsen, J. K.: 2004, *Applied Mathematical Models in Human Physiology*, SIAM, Philadelphia. 230

Packard, N. H., Crutchfield, J. P., Farmer, J. D. and Shaw, R. S.: 1980, Geometry from a time series, *Phys. Rev. Lett.* **45**, 712–716. 14, 43, 152

Pascal-Marqui, R. D.: 2002, Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details, *Methods & Findings in Experimental & Clinical Pharmacology* **24D**, 5–12. 234

Pell, G. S., Briellmann, R. S., Waites, A. B., Abbott, D. F. and Jackson, G. D.: 2004, Voxel-based relaxometry: a new approach for analysis of T2 relaxometry changes in epilepsy, *NeuroImage* **21**, 707–713. 233

Peng, C. K., Havlin, S., Stanley, H. E. and Goldberger, A. L.: 1995, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat series, *Chaos* **5**, 82–87. 68, 87

Peng, C. K., Mietus, J. E., Liu, Y., Lee, C., Hausdorff, J. M., Stanley, H. E., Goldberger, A. L. and Lipsitz, L. A.: 2002, Quantifying fractal dynamics of human respiration: Age and gender effects, *Ann. Biomed. Eng.* **30**, 683–692. 79

Pereda, E., Quiroga, R. Q. and Bhattacharya, J.: 2005, Nonlinear multivariate analysis of neurophysiological signals, *Prog. Neurobiol.* **77**, 1–37. 140, 141

Peslin, R. and Fredberg, J.: 1986, Oscillation mechanics of the respiratory system, *Handbook of Physiology. The Respiratory System. III.*, American Physiological Society, Bethesda, chapter 11, pp. 145–177. 231

Peslin, R., Ying, Y., Gallina, C. and Duvivier, C.: 1992, Withing-breath variations of forced oscillation resistance in healthy subjects, *Eur. Respir. J.* **5**, 86–92. 63

Peták, F., Hayden, M. J., Hantos, Z. and Sly, P. D.: 1997, Volume dependence of respiratory impedance in infants, *Am. J. Resp. Crit. Care Med.* **156**, 1172–. 230

Pikovsky, A., Rosenblum, M. and Kurths, J.: 2003, *Synchronization: A universal concept in nonlinear sciences*, Cambridge University Press, Cambridge. 32, 234

Pincus, S. M.: 1991, Approximate entropy as a measure of system complexity, *Proc. Natl. Acad. Sci. U. S. A.* **88**, 2297–2301. 74, 210

Poon, C.-S. and Barahona, M.: 2001, Titration of chaos with added noise, *Proc. Natl. Acad. Sci. U S A* **98**, 7107–7112. 30

Que, C. L., Kenyon, C. M., Olivenstein, R., Macklem, P. T. and Maksyn, G. N.: 2001, Homeokinesis and short-term variability of human airway caliber, *J. Appl. Physiol.* **91**, 1131–1141. 231

Que, C. L., Maksym, G. and Macklem, P. T.: 2000, Deciphering the homeokinetic code of airway smooth muscle, *Am. J. Respir. Crit. Care Med.* **161**, S161–163. 78, 86

Quiroga, R. Q., Kraskov, A., Kreuz, T. and Grassberger, P.: 2002, Performance of different synchronization measures in real data: A case study on electroencephalographic signals, *Phys. Rev. E* **65**, 041903. 141

Quyen, M. L. V. and Bragin, A.: 2007, Analysis of dynamic brain oscillations: methodological advances, *TINS* **30**, 365–373. 152

Quyen, M. L. V., Foucher, J., Lachaux, J.-P., Rodriguez, E., Lutz, A., Martinerie, J. and Varela, F. J.: 2001, Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony, *J. Neurosci. Meth.* **111**, 83–98. 151

R Development Core Team: 2008, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org* 201

Rahman, A. and Isenberg, D. A.: 2008, Systemic lupus erythematosus, *N. Engl. J. Med.* **358**, 929–939. 95

Ramnani, N., Behrens, T. E. J., Penny, W. and Matthews, P. M.: 2004, New approaches for exploring

anatomical and functional connectivity in the human brain, *Biol. Psychiatry* **56**, 613–619. 139

Ramsay, J. O. and Silverman, B. W.: 1997, *Functional Data Analysis*, Springer, New York. 44

Rapp, P. E., Albano, A. M., Schmah, T. I. and Farwell, L. A.: 1993, Filtered noise can mimic low dimensional chaotic attractors, *Phys. Rev. E* **47**, 2289–2297. 12

Receiver operating characteristic analysis in medical imaging: 2008, *Report 79*, JİCRU **8**. 235

Redelmeier, D. A., Bloch, D. A. and Hickam, D. H.: 1991, Assessing predictive accuracy: How to compare Brier scores, *J. Clin. Epidemiol.* **44**, 1141–1146. 236

Reed, W. J. and Hughes, B. D.: 2002, From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature, *Phys. Rev. E* **66**, 067103. 66

Richardson, A. G., Lassi-Tucci, G., Padoa-Schioppa, C. and Bizzi, E.: 2008, Neuronal activity in the cingulate motor areas during adaption to a new dynamic environment, *J. Neurophysiol.* **99**, 1253–1266. 116

Richman, J. S. and Moorman, J. R.: 2000, Physiological time-series analysis using approximate entropy and sample entropy, *Am. J. Physiol. Heart Circ. Physiol.* **278**, H2039–2049. 74, 210

Roberson, E. D. and Mucke, L.: 2006, 100 years and counting: Prospects for defeating Alzheimer's disease, *Science* **314**, 781–784. 233

Robinson, J.: 1982, Saddlepoint approximations for permutation tests and confidence intervals, *J. R. Statist. Soc. B* **44**, 91–101. 136

Rodgers, J. L. and Nicewander, W. A.: 1988, Thirteen ways to look at the correlation coefficient, *Am. Statist.* **42**, 59–66. 174

Rohrer, F.: 1915, Flow resistance in human air passages and the effect of irregular branching of the bronchial system on the respiratory process in various regions of the lung, *Arch. Ges. Physiol.* **162**, 225–299. 231

Roweis, S. T. and Saul, L. K.: 2000, Nonlinear dimensionality reduction by locally linear embedding, *Science* **22**, 2323–2326. 235

Rubner, Y., Tomasi, C. and Guibas, L. J.: 2000, The Earth Mover's Distance as a metric for image retrieval, *Int. J. Comp. Vision* **40**, 99–121. 17, 49, 154, 236

Ruelle, D.: 1981, Small random perturbations of dynamical systems and the definition of attractors, *Commun. Math. Phys.* **82**, 137–151. 12, 16

Sain, S. R.: 2002, Multivariate locally adaptive density estimation, *Comp. Stat. Data Anal.* **39**, 165–186. 232

Sain, S. R. and Scott, D. W.: 1996, On locally adaptive density estimation, *J. Am. Statist. Assoc.* **91**, 1525–1534. 232

Sanei, S. and Chambers, J. A.: 2007, *EEG signal processing*, John Wiley & Sons, New York. 234

Sarvas, J.: 1987, Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem, *Phys. Med. Biol.* **32**, 11–22. 234

Sassi, R., Signorini, M. G. and Cerutti, S.: 2009, Multifractality and heart rate variability, *Chaos* **19**, 028507. 232

Sauseng, P. and Klimesch, W.: 2008, What does phase information of oscillatory brain activity tell us about cognitive processes?, *Neurosci. Biobehav. Rev.* **32**, 1001–1013. 152

Sazonov, A. V., Ho, C. K., Bergmans, J. W. M., Arends, J. B. A. M., Griep, P. A. M., Verbitskiy, E. A. et al.: 2009, An investigation of the phase locking index for measuring of interdependency of cortical source signals recorded in the EEG, *Biol. Cybern.* **100**, 129–146. 152

Schölkopf, B. and Smola, A. J.: 2001, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge. 235

Scheenstra, A. E. H., Muskulus, M., Staring, M., van den Maagdenberg, A. M. J. M., Verduyn-Lunel, S., Reiber, J. H. C., van der Weerd, L. and Dijkstra, J.: 2009, The 3D Moore-Rayleigh test for the quantitative groupwise comparison of MR brain images, *in* J. L. Prince, D. L. Pham and K. J. Meyers (eds),

Information Processing in Medical Imaging. 21st International Conference, IPMI 2009, Vol. 5636 of *Lect. Notes Comp. Sci.*, Springer, New York, pp. 564–575. 127, 137

Schmid, F. and Schmidt, A.: 2006, Nonparametric estimation of the coefficient of overlapping — theory and empirical application, *Comp. Stat. Data Anal.* **50**, 1583–1596. 195, 232

Schreiber, T.: 1999, Interdisciplinary application of nonlinear time series methods, *Phys. Rep.* **308**, 1–64. 152

Schrijver, A.: 1998, *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester. 18, 154

Schwartzman, A., Dougherty, R. F., Lee, J., Ghahremani, D. and Taylor, J. E.: 2009, Empirical null and false discovery rate analysis in neuroimaging, *NeuroImage* **44**, 71–82. 114

Schweitzer, C., Chone, C. and Marchal, F.: 2003, Influence of data filtering on reliability of respiratory impedance and derived parameters in children, *Pediatric Pulm.* **36**, 502–508. 232

Seely, A. J. E. and Macklem, P. T.: 2004, Complex systems and the technology of variability analysis, *Critical Care* **8**, R367–R384. 231

Shannon, C. E. and Weaver, W.: 1949, *A mathematical theory of communication*, Illinois University Press, Urbana. 148

Shawe-Taylor, J. and Cristianini, N.: 2004, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York. 25, 232, 235

Silverman, B. W.: 1978, Density ratios, empirical likelihood and cot death, *Appl. Stat.* **27**, 26–33. 232

Silverman, B. W.: 1981, Using kernel density estimates to investigate multimodality, *J. R. Stat. Soc. B* **43**, 97–99. 232

Silverman, B. W.: 1986, *Density estimation*, Chapman & Hall, London. 232

Singer, A.: 2008, A remark on global positioning from local distances, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9507–9511. 20, 145, 235

Singer, W.: 1993, Synchronization of cortical activity and its putative role in information processing and learning, *Annu. Rev. Physiol.* **55**, 349–374. 150

Singer, W.: 1999, Neuronal synchrony: A versatile code for the definition of relations?, *Neuron* **24**, 49–65. 150

Slats, A. M., Janssen, K., van Schadewijk, A., van der Plas, D. T. et al.: 2007, Bronchial inflammation and airway responses to deep inspiration in asthma and chronic obstructive pulmonary disease, *Am. J. Respir. Crit. Care Med.* **176**, 121–128. 13, 39, 61, 64

Smale, S.: 1972, On the mathematical foundations of electrical circuit theory, *J. Diff. Geom.* **7**, 193–210. 231

Small, C. G.: 1996, *The Statistical Theory of Shape*, Springer, New York. 21, 161

Socolovsky, E. A.: 2002, A dissimilarity measure for clustering high- and infinite dimensional data that satisfies the triangle inequality, *Technical Report NASA/CR-2002-212136*, Institute for Computer Applications to Science and Engineering. NASA Langley Research Center, Hampton. 148

Spence, I. and Domoney, D. W.: 1974, Single subject incomplete designs for nonmetric multidimensional scaling, *Psychometrika* **39**, 469–490. 20, 145

Spiegelhalter, D. J.: 1986, Probabilistic prediction in patient management and clinical trials, *Stat. Med.* **5**, 421–433. 236

Sporns, O. and Kötter, R.: 2004, Motifs in brain networks, *PLOS Biol.* **2**, 1910–1918. 146

Stam, C. J.: 2005, Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field, *Clin. Neurophysiol.* **116**, 2266–2301. 139, 152

Stam, C. J., Nolte, G. and Daffertshofer, A.: 2007, Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources, *Hum. Brain Mapp.* **28**, 1178–1193. 153

Stam, C. J. and Reijneveld, J. C.: 2007, Graph theoretical analysis of complex networks in the brain., *Nonlin.*

Biomed. Phys. pp. 3–21. 146

Stam, C. J. and van Dijk, B. W.: 2002, Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets, *Physica D* **163**, 236–251. 32, 143, 152

Stark, J.: 2000, Observing complexity, seeing simplicity, *Phil. Trans. R. Soc. Lond. A* **358**, 41–61. 14, 152

Stark, J., Broomhead, D. S., Davies, M. E. and Huke, J.: 1997, Takens embeddings theorems for forced and stochastic systems, *Nonlin. Anal.* **30**, 5303–5314. 15

Stephan, K. E., Riera, J. J., Deco, G. and Horwitz, B.: 2008, The brain connectivity workshops: moving the frontiers of computational neuroscience, *NeuroImage* **42**, 1–9. 139

Stephens, M. A.: 1962, Exact and approximate tests for directions. I., *Biometrika* **49**, 463–477. 116

Stevens, S. S.: 1946, On the theory of scales of measurement, *Science* **103**, 677–680. 164

Strogatz, S. H.: 2000, From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators, *Physica D* **143**, 1–20. 234

Suki, B.: 1993, Nonlinear phenomena in respiratory mechanical measurements, *J. Appl. Physiol.* **74**, 2574–2584. 88

Suki, B.: 2002, Fluctuations and power laws in pulmonary physiology, *Am. J. Respir. Crit. Care Med.* **166**, 133–137. 232

Suki, B., Yuan, H., Zhang, Q. and Lutchen, K. R.: 1997, Partitioning of lung tissue response and inhomogeneous airway constriction at the airway opening, *J. Appl. Physiol.* **82**, 1349–1359. 230

Sun, F. T., Miller, L. M. and D'Esposito, M.: 2004, Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data, *NeuroImage* **21**, 647–658. 234

Takens, F.: 1981, Detecting strange attractors in turbulence, *in* D. A. Rand and L. S. Young (eds), *Dynamical Systems and Turbulence*, Vol. 898 of *Lect. Notes Math.*, Springer, New York, pp. 366–381. 14, 15, 152, 154

Tan, E. M., Cohen, A. S., Fries, J. F., Masi, A. T., McShane, D. J., Rothfield, N. F. et al.: 1982, The 1982 revised criteria for the classification of systemic lupus erythematosus, *Arthritis & Rheumatism* **25**, 1271–1277. 95

Taqqu, M. S., Teverosvsky, V. and Willinger, W.: 1995, Estimators for long-range dependence: An empirical study, *Fractals* **3**, 785–798. 71

Tass, P., Rosenblum, M. G., Weule, J., Kurths, J., Pikovsky, A., Volkmann, J., Schnitzler, A. and Freund, H.-J.: 1998, Detection of n:m phase locking from noisy data: Application to magnetoencephalography, *Phys. Rev. Lett.* **81**, 3291–3294. 153

Tax, D. M. J., van Breukelen, M., Duin, R. P. W. and Kittler, J.: 2000, Combining multiple classifiers by averaging or by multiplying?, *Patt. Recog.* **33**, 1475–1485. 192

Tenenbaum, J. B., de Silva, V. and Langford, J. C.: 2000, A global geometric framework for nonlinear dimensionality reduction, *Science* **22**, 2319–2323. 235

Theunissen, F. and Miller, J. P.: 1995, Temporal encoding in nervous systems: a rigorous definition, *J. Comp. Neurosci.* **2**, 149–62. 150

Thirion, J.-P.: 1998, Image matching as a diffusion process: an analogy with Maxwell's demons, *Med. Image Anal.* **2**, 243–260. 134

Tofts, P. (ed.): 2004, *Quantitative MRI of the brain: measuring changes caused by disease*, John Wiley & Sons, New York. 7, 93

Tognoli, E. and Kelso, J. A. S.: 2009, Brain coordination dynamics: True and false faces of phase synchrony and metastability, *Prog. Neurobiol.* **87**, 31–40. 234

Tonini, G.: 1998, Consciousness and complexity, *Science* **282**, 1846–1851. 234

Tononi, G., Edelman, G. M. and Sporns, O.: 1998, Complexity and coherency: integrating information in the brain, *Trends Cogn. Sci.* **2**, 474–484. 234

Trosset, M. W.: 2000, Distance matrix completion by numerical optimization, *Comp. Optim. Appl.* **17**, 11–22. 235

Trosset, M. W. and Priebe, C. E.: 2008, The out-of-sample problem for classical multidimensional scaling, *Comp. Stat. Data Anal.* **52**, 4635–4642. 26, 140, 145, 190, 202, 235

Tukker, J. J., Fuentealba, P., Hartwich, K., Somogyi, P. and Klausberger, T.: 2007, Cell type-specific tuning of hippocampal interneuron firing during gamma oscillations *in vivo*, *J. Neurosci.* **27**, 8184–8189. 116

Twieg, D. B.: 1983, The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods, *Med. Phys.* **10**, 610–621. 232

Tzeng, J., Lu, H. H.-S. and Li, W.-H.: 2008, Multidimensional scaling for large genomic data sets, *BMC Bioinf.* **9**, 179. 235

Ulrich, G.: 1984, Computer generation of distributions on the sphere, *Appl. Stat.* **33**, 158–163. 122, 220

Vaidya, P. M.: 1989, Geometry helps in matching, *SIAM J. Comput.* **18**, 1201–1225. 237

van Beers, R. J., Haggard, P. and Wolpert, D. M.: 2004, The role of execution noise in movement variability, *J. Neurophysiol.* **91**, 1050–1063. 116

van der Putten, W. J. M., MacLeod, D. and Prichard, J. S.: 1993, Within-breath measurement of respiratory impedance, *Physiol. Meas.* **14**, 393–400. 62

Varadarajan, K. R. and Agarwal, P. K.: 1999, Approximation algorithms for bipartite and non-bipartite matching in the plane, *Proceedings of the tenth annual ACM-SIAM symposium on discrete algorithms*, SIAM, Philadelphia, pp. 805–814. 237

Venables, W. N. and Ripley, B. D.: 1999, *Modern Applied Statistics with S-PLUS*, Springer, New York. 186, 188

Venegas, J. G., Winkler, T., Musch, G., Melo, M. F. V., Layfield, D., Tgavalekos, N., Fischman, A. J., Callahan, R. J., Bellani, G. and Harris, S.: 2007, Self-organized patchiness in asthma as a prelude to catastrophic shifts, *Nature* **434**, 777–781. 232

Villani, C.: 2003, *Topics in Optimal Transportation*, American Mathematical Society, Providence. 3, 13, 16, 17, 19, 52, 89, 140, 153, 154, 197, 198, 199, 236

Vincent, N. J., Knudson, R., Leith, D. E., Macklem, P. T. and Mead, J.: 1970, Factors influencing pulmonary resistance, *J. Appl. Physiol.* **29**, 236–243. 62

Vrba, J. and Robinson, S. E.: 2001, Signal processing in magnetoencephalography, *Methods* **25**, 249–271. 155

Wagers, S., Lundblad, L., Moriya, H. T., Bates, J. H. T. and Irvin, C. G.: 2002, Nonlinearity of respiratory mechanics during bronchoconstriction in mice with airway inflammation, *J. Appl. Physiol.* **92**, 1802–1807. 231

Webb, A. R.: 2002, *Statistical Pattern Recognition*, John Wiley & Sons, New York. 5

Webber, Jr., C. L. and Zbilut, J. P.: 1994, Dynamical assessment of physiological systems and states using recurrence plot strategies, *J. Appl. Physiol.* **76**, 965–973. 152, 229

Wehrens, R., Simonetti, A. W. and Buydens, L. M. C.: 2002, Mixture modelling of medical magnetic resonance data, *J. Chemometrics* **16**, 274–282. 232

Weibel, E. R.: 1963, *Morphometry of the human lung*, Academic Press, London. 57, 230

Weibel, E. R.: 1991, Fractal geometry: a design principle for living organisms, *Am. J. Physiol. Lung Cell. Mol. Physiol.* **261**, L361–369. 57, 230

Wessel, N., Riedl, M. and Kurths, J.: 2009, Is the normal heart rate 'chaotic' due to respiration?, *Chaos* **19**, 028508. 232

Whipp, J. (ed.): 1987, *The Control of Breathing in Man*, Manchester University Press, Manchester. 230

White, E. P., Enquist, B. J. and Green, J. L.: 2008, On estimating the exponent of power-law frequency distributions, *Ecology* **89**, 905–912. 66, 87

Whittall, K. P., Bronskill, M. J. and Henkelman, R. M.: 1991, Investigation of analysis techniques for complicated NMR relaxation data, *J. Magn. Res.* **95**, 221–234. 233

Whittall, K. P., MacKay, A. L. and Li, D. K. B.: 1999, Are mone-exponential fits to a few echoes sufficient to determine T_2 relaxation for in vivo human brain?, *Magn. Res. Med.* **1255**, 1255–1257. 233

Wilcoxon, F.: 1945, Individual comparisons by ranking methods, *Biometrics Bull.* **1**, 80–83. 118

Wood, A. T. A.: 1994, Simulation of the von Mises Fisher distribution, *Comm. Stat. - Sim. and Comp.* **23**, 157–164. 122, 220

Xu, L., Krzyżak, A. and Suen, C. Y.: 1992, Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Trans. Systems Man Cyb.* **22**, 418–435. 52, 236

Yao, F. F.: 1980, Efficient dynamic programming using quadrangle inequalities, *Proceedings of the twelth annual ACM symposium on theory of computing*, ACM, New York, pp. 429–435. 236

Yoo, T. S.: 2004, *Insight into images: Principles and practice for segmentation, registration, and image analysis*, A. K. Peters Ltd., Natick. 134

Young, G. and Householder, A. S.: 1938, Discussion of a set of points in terms of their mutual distances, *Psychometrika* **3**, 19–22. 24, 235

Young, M. P., Scannell, J. W., O'Neill, M. A., Hilgetag, C. C., Burns, G. and Blakemore, C.: 1995, Non-metric multidimensional scaling in the analysis of neuroanatomical connection data and the organization of the primate cortical visual system, *Phil. Trans. R. Soc. Lond. B* **348**, 281–308. 140

Źółtowski, M.: 2000, An adaptive reconstruction of chaotic attractors out of their single trajectories, *Signal Proc.* **80**, 1099–1133. 89

Nederlandse Samenvatting

Het concept afstand is een fundamenteel begrip dat een belangrijke basis vormt voor de ruimtelijke oriëntatie. Het heeft een directe relatie met het natuurwetenschappelijke meetproces: kwalitatieve metingen resulteren in numerieke gegevens, en deze laten zich meteen vertalen in afstanden. Vanuit dit perspectief lijkt het alsof afstanden, net als in het gewone leven, een afgeleid concept zijn: in een bestaande ruimte wiens meetkundige eigenschappen al vast staan dienen afstanden als hulpmiddel om relaties tussen objecten aan te kunnen geven.

Echter, dit verband laat zich ook omkeren. Met een gegeven meetinstrument dat afstanden bepaalt, kan een abstracte ruimte gedefinieerd worden. Deze ruimte wordt niet alleen door de te meten objecten gedefinieerd, maar ook door de eigenschappen van het meetproces zelf. Als het meetinstrument interessante eigenschappen meet, zijn deze terug te vinden als interessante patronen in de ruimte.

In dit proefschrift wordt dit idee toegepast op complexe systemen: het ademhalingsproces, de structuur en activiteit van de hersenen, en dynamische systemen in meer algemene zin. Om in al deze verschillende situaties een afstand tussen twee systemen te berekenen maken wij gebruik van zogenaamde optimal transport afstanden. Deze afstanden zijn gedefinieerd tussen kansverdelingen en geven aan hoeveel "werk" er nodig is om een gegeven kansverdeling in een andere te transformeren. Daardoor meten deze afstanden, die ook wel Wasserstein afstanden genoemd worden, subtiele verschillen in de vorm van kansverdelingen. Deze kansverdelingen kunnen gewone metingen zijn (met het resultaat van de meting als zekere uitkomst) of van meer ingewikkelde aard. In de specifieke toepassing voor dynamische systemen, kunnen wij het statistische gedrag in de toestandsruimte van het systeem als kansverdeling opvatten, die het lange-termijn gedrag van het systeem benadert (de zogenoemde invariante maat van het systeem).

Deze Wasserstein afstanden worden dan omgezet in een representatie, waarbij elk systeem door een punt weergegeven wordt in een abstracte ruimte, zodanig dat de Euclidische afstanden de gemeten afstanden zo goed mogelijk benaderen. Deze techniek staat bekend als multidimensional scaling en wordt met name gebruikt in de sociale wetenschappen en in de psychologie. Omdat de punt-configuraties

onderliggende eigenschappen van de systemen weergeven, noemen wij deze representatie een functionele representatie, en de corresponderende ruimte de functionele of gedragsruimte.

Deze techniek, die in haar meest voor de hand liggende vorm principale componenten analyse van scalaire producten is, maakt het mogelijk voor complexe systemen met veel (zelfs oneindig veel) vrijheidsgraden een laag-dimensionale representatie te vinden, die de essentiële verschillen in hun gedrag (zoals bepaald door de afstanden) weergeeft en onbeduidende verschillen (bijvoorbeeld veroorzaakt door ruis) onderdrukt. Doordat voor N systemen alle $N(N-1)/2$ afstanden bij deze reconstructie gebruikt worden, is deze methode bijzonder robuust.

In de zo gereconstrueerde functionele ruimte kunnen de methodes uit de multivariate analyse gebruikt worden. In dit proefschrift bekijken wij daarbij vooral de classificatie van verschillende groepen van systemen. Canonische discriminant analyse is een klassieke techniek die een lineaire transformatie van de ruimte zoekt zodanig dat de verschillen tussen de groepen optimaal in kaart worden gebracht.

Deze "afstand-gebaseerde" analyse van complexen systemen wordt in dit proefschrift gebruikt om verschillende longziektes (astma and COPD) van elkaar te scheiden. Dit is een moeilijk probleem, maar wij laten zien dat deze methode grote diagnostische nauwkeurigheid bereikt (Hoofdstuk 3).

Ook in de toepassing tot de activiteit (Hoofdstuk 6) en de structuur (Hoofdstuk 4) van de hersenen blijken de Wasserstein afstanden veelbelovende hulpmiddelen te zijn. In het eerste geval worden overeenkomsten in de dynamica van verschillende hersengebieden in kaart gebracht, en in het tweede geval worden subtiele pathologische veranderingen in de weefseleigenschappen van de hersenen gevonden. Deze maken het mogelijk om bepaalde aandoeningen, zoals de auto-immuunziekte systemic lupus erythematosus en de ziekte van Alzheimer, vroegtijdig te herkennen en te kwantificeren.

Voor algemene dynamische systemen meten de Wasserstein afstanden veranderingen in het lange termijn gedrag. Dit biedt een nieuwe mogelijkheid tot numerieke bifurcatieanalyse, en maakt het mogelijk om synchronisatie tussen systemen te kwantificeren.

Een nadeel van deze methode is helaas dat de berekening van een Wasserstein afstand de oplossing van een optimalisatie probleem vereist, wat tijdrovend kan zijn, vooral wanneer een groot aantal systemen met elkaar vergeleken worden. Toekomstig onderzoek zou zich kunnen richten op snelle benaderingen van Wasserstein afstanden, zodat het mogelijk wordt grotere problemen aan te pakken. Ook is de aanpak in dit proefschrift vooral fenomenologisch: systemen worden succesvol geclassificeerd, maar wat de gemeten verschillen betekenen en door welke processen zij veroorzaakt worden is tot nu toe onvoldoende onderzocht.

Curriculum vitae

Michael Muskulus was born on November 4th, 1974, in Sorengo, Switzerland. After obtaining his A-levels ("Abitur") in the quiet city of Laupheim in the south of Germany, he spent 12 out of 15 months of compulsory community service as an ambulance worker on the island of Wangerooge in the North Sea, where he took courses from the FernUniversität in Hagen in mathematics and philosophy. After that he worked in a summercamp in Wales, participated in a controlled drug experiment, picked garlic in New Zealand, slept in trains in Japan, got very sick in India, and enjoyed Yak cheese in Nepal. After his parents gave up all hope, he finally started to study physics at the University of Hamburg, with geophysics and oceanography as subsidiary subjects. After the two-year Vordiploma he had participated in scientific cruises in the North Sea and China, financed his studies by analyzing oceanographical data and tutoring students in mathematics, and switched his subsidiary studies to philosophy. His master's thesis ("Diplomarbeit") was written on a noncommutative approach to quantum field theory in the group of Prof. Klaus Fredenhagen, while working part time at the Max Planck Institute for Meteorology. Disillusioned by the horrible computations he had encountered, the author of this thesis welcomed a position as researcher at this institute after obtaining his degree as "Diplom-Physiker", where he worked in the Regional Climate Modelling group of Dr. Daniela Jacob, tracking cyclones in long-term simulations of Europe's climate.

In the beginning of 2004 the author had a difficult choice to make: either to work 15 months at the Georg von Neumayer station in Antarctica, or to begin a PhD at the Mathematical Institute in Leiden, The Netherlands, in the group of Prof. dr. S.M. Verduyn Lunel (Analysis and Dynamical Systems). The reader holds the result of this choice in his hands. The author's PhD was initially funded under the Computational Life Sciences initiative of the Netherlands Organization for Scientific Research (NWO), and after a promising start that led to quite a few publications in the interdisciplinary VIEWS project, the author stumbled upon an article with an elegant method to compare time series with each other. The constant enthusiasm of his su-

pervisor ensured that an interesting extension of this method, which subsequently grew in detail and complexity, finally evolved into the topic covered in this thesis. Originally intended as a small project with an estimated three weeks of duration, the author has now realized that developing a novel method and writing accompanying software is an evolutionary process that does never end.

The author has received additional training during summer schools and workshops in Milano (Italy), Graz (Austria), Heeze (The Netherlands), Toronto (USA), Oberwolfach (Germany), and quite a few times at the Lorentz centre in Leiden (The Netherlands). He has presented his work in the form of papers or posters at international meetings and conferences in Sevilla (Spain), Edinburgh (Great Britain), Prague (Czech Republic), Sapporo (Japan), Toronto (USA) and San Diego (USA), and has received a Young Investigator Award for Best Paper (Sapporo, Japan). A joint publication with Alize Scheenstra has won a Best Paper Award (Williamsburg, USA).

The author has also been a regular participant of the Study Group Mathematics with Industry, which he enjoyed so much that he considers becoming a mathematical consultant in the distant future. He has taught courses in mathematics, both basic (1st year students) and targeted at the explanation of biological phenomena (2nd year students), and in the context of a winter school in nonlinear dynamics (PhD students). He has reviewed manuscripts for a number of distinct journals, some of which still send him Christmas greetings.

At present he is living in Trondheim, Norway, where he has recently started a Postdoctoral position in the field of wind energy and wind turbine dynamics under Prof. Geir Moe at the Norwegian University of Science and Technology (NTNU). In his freetime, his greatest ambition is to solve the $L(m, n)$ problem.

ISBN 978-90-5335-254-0