PAPER

A comparative study of estimation methods in quantum tomography

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

A comparative study of estimation methods in quantum tomography

Anirudh Acharya, Theodore Kypraios and Mădălin Guţă

School of Mathematical Sciences, University of Nottingham, University Park,
NG7 2RD Nottingham, United Kingdom

E-mail: madalin.guta@nottingham.ac.uk

Abstract

As quantum tomography is becoming a key component of the quantum engineering toolbox, there is a need for a deeper understanding of the multitude of estimation methods available. Here we investigate and compare several such methods: maximum likelihood, least squares, generalised least squares, positive least squares, thresholded least squares and projected least squares. The common thread of the analysis is that each estimator projects the measurement data onto a parameter space with respect to a specific metric, thus allowing us to study the relationships between different estimators.

The asymptotic behaviour of the least squares and the projected least squares estimators is studied in detail for the case of the covariant measurement and a family of states of varying ranks. This gives insight into the rank-dependent risk reduction for the projected estimator, and uncovers an interesting non-monotonic behaviour of the Bures risk. These asymptotic results complement recent non-asymptotic concentration bounds of Guta *et al* (2018 (arXiv:1809.11162)) which point to strong optimality properties, and high computational efficiency of the projected linear estimators.

To illustrate the theoretical methods we present results of an extensive simulation study. An app running the different estimators has been made available online.

Keywords: quantum tomography, statistical methods, maximum likelihood, least squares tomography, projected least squares, low rank states

(Some figures may appear in colour only in the online journal)

1. Introduction

The problem of estimating unknown parameters of quantum systems has been at the forefront of early investigations into the statistical nature of quantum information [13, 16, 42, 43, 78]. Traditionally, key research topics have been the design of optimal measurements and estimation procedures [8, 9, 21, 31, 41, 47, 56, 76], and theoretical aspects of quantum Fisher information and asymptotic estimation [6, 11, 28, 30, 41, 45, 57, 61], see also the monographs [40, 52, 59].

More recently, quantum tomography has become a crucial validation tool current quantum technology applications [29, 38, 58, 66]. The experimental challenges have stimulated research in new directions such as compressed sensing [3, 5, 20, 27, 35, 54, 69], estimation of permutationally invariant states [72], adaptive and selflearning tomography [26, 34, 39, 55, 60, 62], incomplete tomography [73], minimax bounds [4, 25], Bayesian estimation [15, 33], and confidence regions [7, 18, 24, 53, 70]. Since 'full tomography' becomes impossible for systems composed of even a moderate number of qubits, research has focused on the estimation of states which belong to smaller dimensional models which nevertheless capture relevant physical properties, such as low rank states [17, 36, 37, 48, 49] and matrix product states [12, 20, 50].

In this paper we analyse and compare several estimation methods for fixed (non-adaptive) measurement designs, with a focus on risks (mean errors) for different loss functions, asymptotic theory, relationships between estimators and low rank behaviour. The measurement scenarios include repeated measurements with respect to Pauli bases (as customary in multiple qubits experiments), random bases measurements, and the covariant measurement. The loss functions are given by the Frobenius, trace-norm, operator-norm, Bures and Hellinger distances. Each section deals with one class of estimators and the results of a comparative simulations study are presented at the end of the paper. While most estimators have been previously considered in the literature, our aim is to investigate them from a common perspective, as projections of data onto the parameter space. Another aim is to understand and quantify the reduction in statistical error between an unconstrained estimator such as least squares, and an estimator which take into account the physical properties of the parameter space, such as the projected least squares estimator. Among the original results, we derive the asymptotic error rates of these estimators on a class of low rank states in the covariant measurement scenario. Finally, we discuss the computational efficiency of different methods.

Below we summarise the paper using figure 1 for illustration. A measurement $\mathcal{M}$ on d-dimensional system is a positive affine map from the convex set of states $\mathcal{S}_d$ (constrained parameter space) onto the space of outcome probabilities $\mathcal{P}_d \subset \mathbb{R}^z$, where z is the number outcomes. The image of the set of trace-one selfadjoint matrices $M_{sa}^1(\mathbb{C}^d)$ (unconstrained parameter space) is a hyperplane $\mathcal{L}_d$ of $\mathbb{R}^z$, which contains $\mathcal{P}_d$. For informationally complete measurements the map $\mathcal{M}$ is injective and we can identify all matrix estimators (whether positive or not) with their images in $\mathcal{L}_d$.

In section 3 we review the use of Fisher information in asymptotic normality theory and discuss to what extent it is applicable to the study of the maximum likelihood (ML) estimator in quantum tomography [10, 14, 65, 74]. We distinguish between two ML estimators: the *unconstrained* estimator $\hat{\rho}_{uML}$, and the 'physical' estimator $\hat{\rho}_{ML}$. The former is the maximiser of the likelihood, seen as a function over $M_{sa}^1(\mathbb{C}^d)$, and may not be a density matrix. The latter performs the same optimisation but restricted to the physical space of density matrices $\mathcal{S}_d \subset M_{sa}^1(\mathbb{C}^d)$. For large sample size, the unconstrained estimator is asymptotically normal, while the constrained estimator is equal to its projection onto $\mathcal{S}_d$ with respect to the metric defined by the classical Fisher information [65].
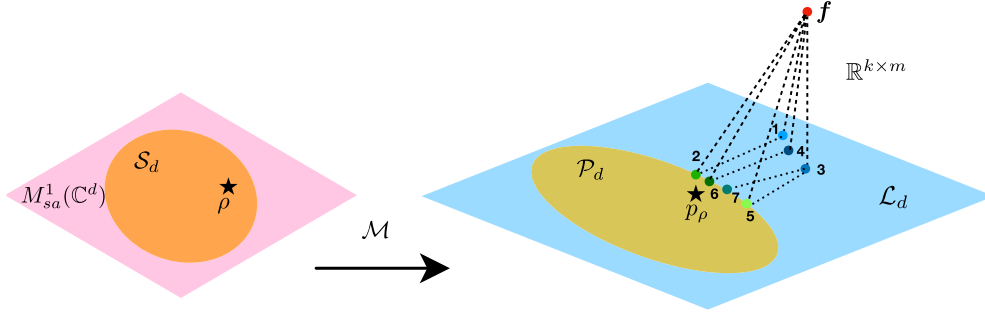
Figure 1. A measurement $\mathcal{M}$ maps the set of states $\mathcal{S}_d$ onto outcome probabilities $\mathcal{P}_d$. Estimators are represented via their images through $\mathcal{M}$, which are obtained by projecting the empirical frequencies $\boldsymbol{f}$ onto the hyperplane $\mathcal{L}_d = \mathcal{M}(M_{\mathrm{sa}}^1(\mathbb{C}^d))$ or the convex set $\mathcal{P}_d$ with respect to a metric. The uML (1) and ML (2) estimators are projections with respect to relative entropy. The GLS (4) and posGLS (6) estimators are projections with respect to the covariance metric, and are asymptotically equivalent to the uML and ML. The LS (3) and posLS (5) estimators project $\boldsymbol{f}$ with respect to the euclidian distance. The PLS (7) estimator projects LS onto $\mathcal{P}_d$ with respect to the Frobenius distance inherited from $M(\mathbb{C}^d)$.

In section 4 we analyse the least squares (LS) estimator $\hat{\rho}_{\mathrm{LS}}$. This exploits the linear dependence between probabilities and states which translates into a linear regression problem

$$\boldsymbol{f} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{f}$ denotes the vector of observed frequencies of measurement outcomes, X is a fixed measurement design matrix, $\boldsymbol{\beta}$ is a vectorisation of the unknown state ρ, and $\boldsymbol{\epsilon}$ is the statistical noise. The LS estimator minimises the prediction error $\|X\hat{\boldsymbol{\beta}} - \boldsymbol{f}\|^2$ over all vectors $\hat{\boldsymbol{\beta}}$, and is the simplest and fastest estimator to compute. However, LS is significantly less accurate than ML, especially for low rank states and does not produce a physical state. Nevertheless, the LS has been the focus of recent investigations [17, 36, 71] as a first step towards more accurate estimators. Here we review some of the *non-asymptotic* concentration bounds for the operator-norm error of the LS estimator. We then study the *asymptotic* properties of the LS estimator in the context of covariant measurements, and rank r states ρ_r with equal non-zero eigenvalues. By exploiting the symmetry of the measurement we compute the explicit expression of the risk with respect to the Frobenius distance. Furthermore, we show that for large d, the eigenvalues of the 'error matrix' $\hat{\rho}_{\mathrm{LS}} - \rho_r$ are approximately distributed according to the Wigner semicircle law on the interval $[-2\sqrt{d/N}, 2\sqrt{d/N}]$, see figure 2. This provides the asymptotic estimates of the operator-norm, and the trace-norm errors of $2\sqrt{d/N}$ and respectively $8d^{3/2}/(3\pi\sqrt{N})$, which complement the non-asymptotic bounds of [36].

In section 5 we discuss the generalised least squares (GLS) estimator. Unlike the LS estimator which minimises the prediction error, the GLS aims to optimise the estimation error, e.g. $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ and needs to take into account the covariance matrix of the multinomial noise $\boldsymbol{\epsilon}$. We show that for large samplesize N, the GLS estimator is asymptotically normal and is equivalent to the uML estimator.

Section 6 reviews the thresholded least squares (TLS) estimator proposed in [17]. This is obtained by projecting the LS estimator onto the set of states whose non-zero eigenvalues are above a certain 'noise level' chosen in a data-driven fashion. The projection can be computed efficiently in terms of the eigenvalues of the LS estimator, and the truncation leads
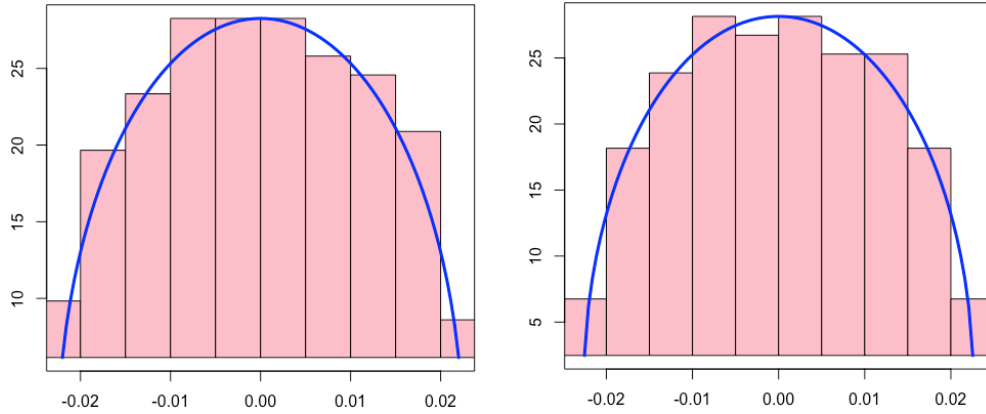
Figure 2. Left panel: histogram of eigenvalues of the 'error block' $C/\sqrt{N}$ versus the corresponding Wigner distribution, for a $n = 7$ atoms, rank $r = 1$ state and $N = 10^6$ samples. Right: histogram of eigenvalues of error matrix for a rank $r = 128$ state.

to significant error reduction for low rank states. In practice, an additional improvement is achieved by using the GLS estimator as starting point.

Section 7 discusses the positive least squares (posLS) estimator which optimises the prediction error over the physical space of density matrices, rather than over the selfadjoint matrices as is the case for the LS estimator [46]. This leads to higher accuracy, but its computational complexity is similar to that of ML. However, in the case of the covariant measurements we find that posLS is equivalent to the projected least squares estimator discussed below, which *can* be computed efficiently. By restricting the GLS optimisation over density matrices we obtain the positive generalised least squares estimator (posGLS), which is shown to be asymptotically equivalent to the ML estimator.

Section 8 deals with the projected least squares (PLS) estimator. This is obtained by projecting the LS estimator onto the space of density matrices with respect to the Frobenius distance [36, 71]. It is faster than TLS, and has similar statistical properties. In [36] is was shown that the PLS estimator satisfies rank-dependent concentration bounds and its trace-norm error scales as $O(r^2 \cdot d \cdot \log d/\sqrt{N})$ for a broad class of two-design measurements including mutually unbiased bases, stabiliser states, and symmetric informationally complete measurements, and as $O(r^2 \cdot d^{1.6} \cdot \log d/\sqrt{N})$ for Pauli bases measurements. In this paper we focus on the asymptotic behaviour of the PLS estimator in the case of covariant measurements. Inspired by the techniques developed in [65] we show how the random matrix properties of the LS can be used to derive the asymptotic Frobenius and Bures risks of the PLS estimator for low rank states and large dimension d and samplesize N. In particular we uncover an interesting behaviour of the Bures risk which (unlike the Frobenius or the trace norm risks) increases steeply with rank for low rank states and then decreases towards full rank, see figure 3.

Section 9 presents the results of a comparative simulations study of the proposed estimators. We simulated data for a range of states of 3 and 4 atoms of different ranks, with different samplesizes, and measurement setups. For each choice we produced 100 datasets in order to estimate risks of different estimators. The measurements are chosen to be either Pauli bases measurements (as standard in ion trap experiments) or random basis measurements. The different estimators and their risks corresponding to Frobenius, Bures, Hellinger, and trace-norm distances were computed and the results are illustrated and discussed, see figures 4–11. A complete set of simulation results available online via an interactive Rshiny
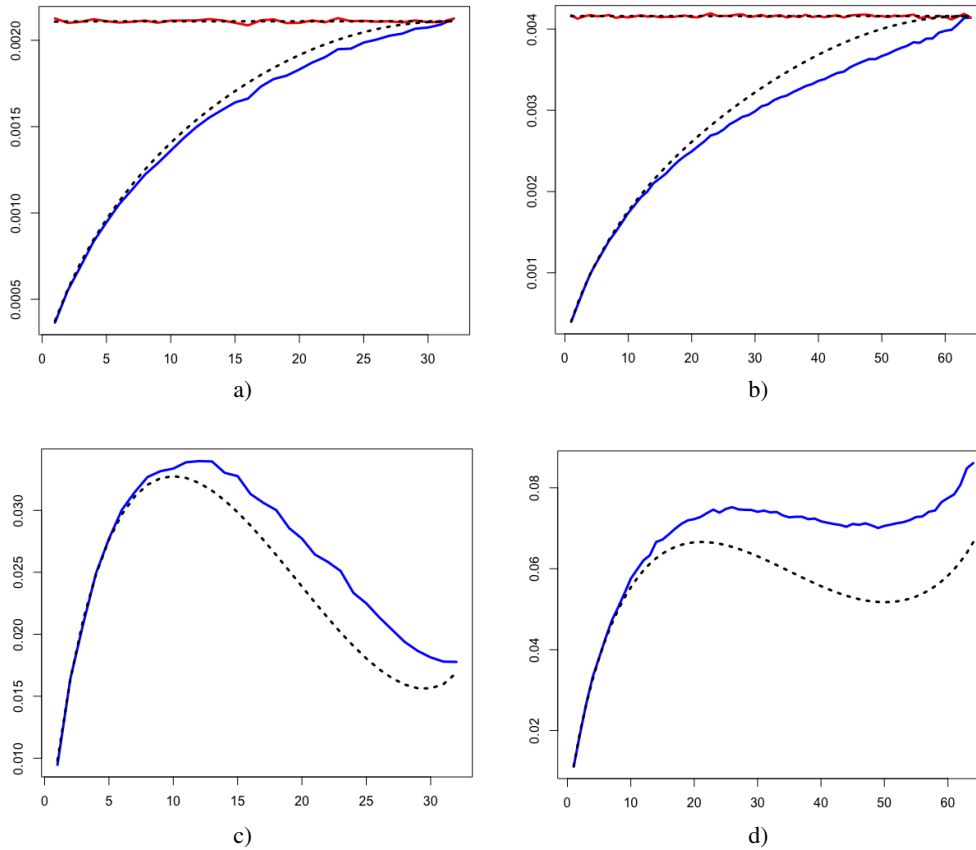
Figure 3. Frobenius and Bures risks for LS (red line) and PLS (blue line) versus the asymptotic predictions (black dotted lines). (a) Frobenius risks for $n = 5$, $N = 5 \cdot 10^5$. (b) Frobenius risks for $n = 6$, $N = 10^6$. (c) Bures risks for $n = 5$, $N = 5 \cdot 10^5$. (d) Bures risks for $n = 6$, $N = 10^6$.

application: https://rudhacharya.shinyapps.io/plots/. We also created an online estimation app which computes the key estimators for a range of states of certain characteristics, or user uploaded states (https://shiny.maths.nottingham.ac.uk/shiny/qt_dashboard/).

2. Quantum tomography

In quantum tomography, the aim is to estimate an unknown state from outcomes of measurements performed on an ensemble of identical prepared systems. Although a large part of our theoretical considerations hold in a general setting, we choose to formulate the tomography problem in the specific context of a system consisting of multiple qubits and projective measurements with respect to several bases. This keeps the discussion closer to realistic experimental procedures, and facilitates the understanding of our simulation results.

We consider composite systems consisting of n qubits, with associated Hilbert space $\mathbb{C}^d$, where $d = 2^n$. The state ρ belongs to the space of $d \times d$ density matrices $\mathcal{S}_d \subset M(\mathbb{C}^d)$. In our analysis we will distinguish between the constrained parameter space $\mathcal{S}_d$ whose elements are trace-one positive matrices, and the unconstrained parameter space $M_{\mathrm{sa}}^1(\mathbb{C}^d)$ consisting of

trace-one selfadjoint matrices. This will allow us to consider procedure which produce constrained, or unconstrained estimators.

The measurement strategies consist of performing standard, von Neumann projective measurements with respect to a number of orthonormal bases (ONB), which are chosen deterministically or randomly. In particular, we focus on two scenarios: Pauli basis measurements and measurements that are drawn randomly from the uniform measure over all ONBs. While the former setup is commonly employed in experiments [38], the latter is less restrictive, more amenable to theoretical analysis, and can serve as a benchmark for current experiments. We also consider covariant measurements in the context of the asymptotic theory of the least squares and respectively projected least squares estimators in sections 4.3 and 8.1, and refer to these for further details.

In the Pauli bases setup, one measures an arbitrary Pauli observable σ_x, σ_y, or σ_z on each of the n qubits simultaneously. Therefore, each measurement is labelled by a sequence $\boldsymbol{s} = (s_1, \ldots, s_n) \in \{x, y, z\}^n$, and there are 3^n possible measurement bases. Such a measurement produces a ± 1 outcome from each ion, and we let $\boldsymbol{o} \in \{+1, -1\}^n$ be the full record of outcomes from all the n qubits. The probability of obtaining a particular outcome $\boldsymbol{o}$ is given by $p_\rho(\boldsymbol{o}|\boldsymbol{s}) := \mathrm{Tr}(\rho P_{\boldsymbol{o}}^{\boldsymbol{s}})$, where $P_{\boldsymbol{o}}^{\boldsymbol{s}}$ is the one-dimensional projection onto the tensor product of eigenvectors of Pauli matrices

$$P_{\boldsymbol{o}}^{\boldsymbol{s}} = |e_{o_1}^{s_1}\rangle\langle e_{o_1}^{s_1}| \otimes \ldots \otimes |e_{o_n}^{s_n}\rangle\langle e_{o_n}^{s_n}|, \quad \sigma_s|e_o^s\rangle = o|e_o^s\rangle.$$

More generally, the measurement design is defined by a collection $\mathscr{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ of ONBs, which may be chosen deterministically or randomly. For each setting $\boldsymbol{s}$, independent measurements are performed on m identical copies of the state, and the counts $N(\boldsymbol{o}|\boldsymbol{s})$ are recorded, where $N(\boldsymbol{o}|\boldsymbol{s})$ is the number of times the outcome $\boldsymbol{o}$ is observed when measuring in setting $\boldsymbol{s}$. The total number of quantum samples is therefore $N = m \times k$. The resulting dataset $\mathcal{D} := \{N(\boldsymbol{o}|\boldsymbol{s}) : \boldsymbol{o} \in \{+1, -1\}^n, \boldsymbol{s} \in \mathscr{S}\}$ is a $2^n \times k$ table whose columns are independent and contain all the counts in a given setting. Its probability is given by the product of multinomials

$$p_\rho(\mathcal{D}|\mathcal{S}) = \prod_{\boldsymbol{s}} \frac{m!}{\prod_{\boldsymbol{o}} N(\boldsymbol{o}|\boldsymbol{s})!} \prod_{\boldsymbol{o}} p_\rho(\boldsymbol{o}|\boldsymbol{s})^{N(\boldsymbol{o}|\boldsymbol{s})}. \tag{1}$$

Our goal is to statistically reconstruct the density matrix ρ from the counts dataset $\mathcal{D}$. This can be seen as a statistical inverse problem of reversing the measurement map

$$\mathcal{M} : M(\mathbb{C}^d) \to \mathbb{R}^{d \cdot k}$$
$$\mathcal{M} : \rho \mapsto p_\rho(\cdot|\mathscr{S})$$

in the sense that given a sample $\mathcal{D}$ from $p_\rho(\cdot|\mathcal{S})$, one would like to produce an estimator $\hat{\rho}(\mathcal{D})$ which is close to ρ with respect to some meaningful distance measure. The next section lists several figures of merit considered here.

2.1. Error functions

Let us denote the *risk* or mean error of an estimator $\hat{\rho}$ as $\mathbb{E}[D(\hat{\rho}, \rho)]$, where $D(\hat{\rho}, \rho)$ represents a particular error function. In our theoretical analysis and simulations study we estimate the risk for several choices of the error function $D(\hat{\rho}, \rho)$, which are tabulated in table 1. Note that the Bures distance is defined only over the space of density matrices, and therefore applies only in the case where the estimators are density matrices. The classical analogue of the Bures distance is the Hellinger distance between two probability distributions. Here we consider the Hellinger distance $D_{\mathrm{H}}(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})^2$ between the eigenvalues $\{\lambda_1, \ldots, \lambda_d\}$ of the true state and those

Table 1. The different error functions used to measure the distance between the true state and the estimator. The Bures distance is defined only for states $\hat{\rho}, \rho \in \mathcal{S}_d$, and its classical analogue the Hellinger distance is defined between two probability distributions.

Error function	Definition
Frobenius norm squared	$\|\hat{\rho} - \rho\|_2^2 = \mathrm{Tr}\left[(\hat{\rho} - \rho)^2\right]$
Trace norm	$\|\hat{\rho} - \rho\|_1 = \mathrm{Tr}|\hat{\rho} - \rho|$
Operator norm	$\|\rho - \sigma\| = \lambda_{\max}(|\hat{\rho} - \rho|)$
Bures distance	$D_B(\hat{\rho}, \rho)^2 := 2\left(1 - \mathrm{Tr}\left[\sqrt{\sqrt{\rho}\hat{\rho}\sqrt{\rho}}\right]\right)$
Hellinger distance	$D_{\mathrm{H}}(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})^2 := 2\left(1 - \sum_i^d \sqrt{\hat{\lambda}_i \lambda_i}\right)$

of the estimator $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_d\}$, seen as probability distributions. Later on we will show that the behaviour of the Bures distance error is strongly correlated with that of the squared Hellinger distance error.

3. Fisher information, asymptotic normality and maximum likelihood

The maximum likelihood (ML) estimator is one of the most commonly used and well understood statistical tools with a universal range of applicability. Before considering its use in quantum tomography, we briefly review some of the key concepts and results related to the ML estimator and its asymptotic behaviour [51, 77]. Consider the scenario in which we are given N independent samples $X_1, \ldots, X_N$ drawn from the same discrete probability distribution $p_{\boldsymbol{\theta}}$ over a countable set Ω; the probability distribution is assumed to depend *smoothly* on an unknown parameter $\boldsymbol{\theta}$ which belongs to an *open* subset Θ of $\mathbb{R}^p$. The likelihood of the dataset is $p_{\boldsymbol{\theta}}(X_1, \ldots, X_N) = \prod_{j=1}^N p_{\boldsymbol{\theta}}(X_j)$. The ML estimator is the point in Θ where the likelihood of the observed data attains its maximum value

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} := \arg\max_{\boldsymbol{\theta}' \in \Theta} p_{\boldsymbol{\theta}'}(X_1, \ldots, X_N).$$

The likelihood can be expressed in terms of the *empirical distribution* of the data $\boldsymbol{f} = \sum_j \delta_{X_j}/N$, where δ_x denotes the Dirac distribution at x. The empirical distribution collects the frequencies of different outcomes in Ω. Indeed the log-likelihood can be written as (see (1))

$$\ell_{\boldsymbol{\theta}} := \log p_{\boldsymbol{\theta}}(X_1, \ldots, X_N) = N \sum_{i \in \Omega} \boldsymbol{f}(i) \log p_{\boldsymbol{\theta}}(i) + C(\boldsymbol{f})$$

$$= -NK(\boldsymbol{f}\|p_{\boldsymbol{\theta}}) + C'(\boldsymbol{f})$$

where $K(\cdot\|\cdot)$ relative entropy (Kullback–Leibler divergence), and C, C' do not depend on $\boldsymbol{\theta}$. The ML estimator is thus the closest point to $\boldsymbol{f}$ with respect to the relative entropy

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{\theta}' \in \Theta} K(\boldsymbol{f} \| p_{\boldsymbol{\theta}'})$$

Asymptotics. The appeal of the ML estimator lies in part in its asymptotic optimality properties. For large enough sample sizes N, a central limit behaviour emerges and the ML estimator becomes normally distributed around the true parameter, with a variance shrinking at the rate $1/N$

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, I^{-1}(\boldsymbol{\theta})). \tag{2}$$

The convergence above holds in distribution as $N \to \infty$ and the limit is a centred Gaussian distribution with covariance given by the inverse of the *Fisher information matrix*. The latter is a $p \times p$ positive matrix

$$I(\boldsymbol{\theta})_{i,j} = \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial \log p_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}_i} \frac{\partial \log p_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}_j} \right]$$

which encapsulates how informative a single sample X from $p_{\boldsymbol{\theta}}$ is about $\boldsymbol{\theta}$. The limiting covariance $I^{-1}(\boldsymbol{\theta})$ is equal to the lower bound appearing in the the Cramér–Rao inequality for unbiased estimators. Estimators which satisfy equation (2) are called *efficient* and statistical theory shows that one cannot improve on their asymptotic accuracy, except on a measure zero set of the parameter space [51]. In particular if $d(\cdot, \cdot)$ is a locally quadratic (positive) loss function on Θ,

$$d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T G(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2)$$

with $G(\boldsymbol{\theta})$ a positive weight matrix, then the risk of the MLE satisfies

$$N\mathbb{E}[d(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \boldsymbol{\theta})] \longrightarrow \mathrm{Tr}(G(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})).$$

It is important to note that the *asymptotic normality* property (2) relies on the smoothness of the model and the fact that $\boldsymbol{\theta}$ is an interior point of the parameter space Θ. For large N the ML estimator lies (with high probability) within in a small neighbourhood of $\boldsymbol{\theta}$ of size $1/\sqrt{N}$, and the parameter space looks like $\mathbb{R}^p$ for all practical purposes. However, when Θ has a boundary, the ML estimator will not be asymptotically normal for parameters $\boldsymbol{\theta}$ lying on the boundary, and one needs to analyse such models more carefully. This is the case in quantum tomography, when the unknown state is not full-rank and therefore lies on the boundary of $\mathcal{S}_d$, or is so close to the boundary that the asymptotic theory will kick in only for sample sizes that are much larger than those obtained in real experiments.

3.1. The maximum likelihood estimator in quantum tomography

We will now discuss in more detail the properties of the MLE in the specific case of quantum tomography. Given the measurement data encoded in the dataset $\mathcal{D}$, the MLE is defined as the maximum of $p_\tau(\mathcal{D}|\mathscr{S})$ over $\tau \in \mathcal{S}_d$. By passing to log-likelihood and discarding the constant factorial terms in (1), we obtain the following form of the estimator

$$\hat{\rho}_{\mathrm{ML}} := \arg\max_{\tau \in \mathcal{S}_d} \sum_{\boldsymbol{o,s}} \boldsymbol{f}(\boldsymbol{o}|\boldsymbol{s}) \, p_\tau(\boldsymbol{o}|\boldsymbol{s}) = \arg\min_{\tau \in \mathcal{S}_d} \sum_{\boldsymbol{s}} K(\boldsymbol{f}(\cdot|\boldsymbol{s}) \| p_\tau(\cdot|\boldsymbol{s})) \tag{3}$$

where $\boldsymbol{f}(\boldsymbol{o}|\boldsymbol{s}) = N(\boldsymbol{o}|\boldsymbol{s})/m$ are the empirical frequencies of the data. The MLE is commonly used in quantum tomography [10, 14, 32, 44], and several implementation methods have been proposed including Hradil's iterative algorithm [63]. Our specific implementation uses the CVX package for disciplined convex programming on MATLAB [2]. The general asymptotic theory guarantees that the ML estimator is asymptotically normal for *full rank states*, i.e. the interior of $\mathcal{S}_d$. To get more insight into the general asymptotic behaviour for a given state ρ, we will choose a *local* parametrisation defined in terms of the matrix elements of ρ with respect to its eigenbasis. Let $\lambda_1 \geqslant \lambda_2 \ldots, \geqslant \lambda_d > 0$ be the eigenvalues of a full-rank state ρ. Then any neighbouring state can be written as

$$\rho' = \rho_{\boldsymbol{\theta}} = \rho + \delta\rho_{\boldsymbol{\theta}}$$

where $\delta\rho_{\boldsymbol{\theta}}$ is trace zero matrix which is completely parametrised by

$$
\begin{aligned}
\boldsymbol{\theta} &:= \left(\theta^{(r)}; \theta^{(i)}; \theta^{(d)} \right) \\
&= (\text{Re}\delta\rho'_{1,2}, \dots, \text{Re}\delta\rho'_{d-1,d}\,;\, \text{Im}\delta\rho'_{1,2}, \dots, \text{Im}\delta\rho'_{d-1,d}\,;\, \delta\rho'_{2,2}, \dots, \delta\rho'_{d,d}) \in \mathbb{R}^{d^2-1}.
\end{aligned}
\tag{4}
$$

The ML estimator $\hat{\rho}_{\text{ML}} = \rho_{\hat{\boldsymbol{\theta}}_{\text{ML}}}$ has parameter $\hat{\boldsymbol{\theta}}_{\text{ML}}$ which is normally distributed around 0 with covariance $I^{-1}(\rho|\mathscr{S})/N$ where the Fisher information is the average of the individual informations for different sets $\mathbf{s} \in \mathscr{S}$

$$
I(\rho|\mathscr{S}) = \frac{1}{k} \sum_{\mathbf{s} \in \mathscr{S}} I(\rho|\mathbf{s}), \qquad I(\rho|\mathbf{s})_{a,b} := \sum_{\mathbf{o}:p(\mathbf{o}|\mathbf{s})>0} \frac{1}{p_\rho(\mathbf{o}|\mathbf{s})} \frac{\partial p_\rho(\mathbf{o}|\mathbf{s})}{\partial \theta_a} \cdot \frac{\partial p_\rho(\mathbf{o}|\mathbf{s})}{\partial \theta_b}.
\tag{5}
$$

In particular, for any locally quadratic loss function $d(\cdot, \cdot)$ (e.g. Frobenius distance, or Bures distance) with weight matrix $G(\boldsymbol{\theta})$, the associated risk has the asymptotic behaviour

$$
N\mathbb{E}\left[d(\rho_{\boldsymbol{\theta}}, \hat{\rho}_{\text{ML}}) \right] \longrightarrow \text{Tr}\left(I(\rho|\mathscr{S})^{-1} G(\boldsymbol{\theta}) \right).
\tag{6}
$$

Now, let us assume that the unknown state ρ is on the boundary of $\mathcal{S}_d$, and more precisely belongs to the space of rank r states $\mathcal{S}_d(r) \subset M(\mathbb{C}^d)$, for a *fixed and known* rank $r \leqslant d$. In its own eigenbasis, ρ is the diagonal matrix of eigenvalues $\text{Diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$, and any sufficiently close state is uniquely determined by its matrix elements in the first r rows (or columns). Intuitively this can be understood by noting that any rank-r state ρ' in the neighbourhood of ρ can be obtained by perturbing the eigenvalues and performing a small rotation of the eigenbasis; in the first order of approximations these transformation leave the $(d-r) \times (d-r)$ lower-right corner unchanged so

$$
\rho' = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} A & B \\ B^* & C \end{pmatrix}, \qquad D := \text{Diag}(\lambda_1, \dots, \lambda_r)
\tag{7}
$$

where $C = O(\|A\|^2, \|B\|^2)$. We therefore choose the (local) parametrisation $\rho' = \rho_{\boldsymbol{\theta}}$ where only the matrix elements of A and B enter the parameter $\boldsymbol{\theta}$. For this model, any rank-r state corresponds to an interior point of the parameter space $\mathcal{S}_d(r)$, and consequently the ML estimator obeys the asymptotic normality theory described above.

However, if the rank of ρ is not known in advance, then the diagonal block C of ρ' needs to be included in the model in order to describe neighbouring states of higher rank. As ρ' is a state, the block C is a positive matrix, and therefore its matrix elements are constrained. This complicates the analysis of the likelihood function, and invalidates the asymptotic normality of the ML estimator.

What should be the theory replacing asymptotic normality? At the moment there is not a complete answer to this question, but some important progress has been made in [65]. Following this work and [17], it is instructive to study an extended, 'non-physical' model in which the positivity requirement is dropped and (locally around ρ) the parameter space is taken to be that of selfadjoint matrices of trace-one $M_{\text{sa}}^1(\mathbb{C}^d)$. Therefore, the 'unphysical' parameter $\boldsymbol{\theta}$ consists of the matrix elements of the blocks A, C, B (except one diagonal element due to normalisation), with $\rho' = \rho_{\boldsymbol{\theta}}$ given by equation (7). We now make the 'mild' assumption that $p_\rho(\boldsymbol{o}|\boldsymbol{s}) > 0$ for all pairs $(\boldsymbol{o}, \boldsymbol{s})$; indeed this condition is satisfied for 'generic' states but fails for states whose support is orthogonal to some of the measurement basis vectors. Under this assumption, we find that locally around ρ, we can define a statistical model given by bona-fide probability distributions $p_{\boldsymbol{\theta}} := p_{\rho_{\boldsymbol{\theta}}}$. We can therefore define the *unconstrained* maximum likelihood (uML) estimator as in equation (3) where the optimisation is performed

over the unconstrained local neighbourhood $\mathcal{O} \subset M_{\mathrm{sa}}^1(\mathbb{C}^d)$ of ρ, rather than over the space of states $\mathcal{S}_d$. Since ρ is on the boundary of $\mathcal{S}_d$, the probability that $\hat{\rho}_{\mathrm{uML}}$ falls outside the state space is significant and does not vanish in the asymptotic limit. In this case, $\hat{\rho}_{\mathrm{uML}}$ does not coincide with the 'constrained' or regular ML estimator $\hat{\rho}_{\mathrm{ML}}$. In fact, each of them can be seen as the projection with respect to the relative entropy distance of the empirical distribution $\boldsymbol{f}$ onto the sets of probabilities $\mathcal{M}(\mathcal{O})$ and respectively $\mathcal{M}(\mathcal{S}_d)$, where $\mathcal{M} : M(\mathbb{C}^d) \to \mathbb{R}^{k \cdot d}$ is the measurement map.

Asymptotics. Since ρ corresponds to an interior point of the *unconstrained* parameter space, the general asymptotic normality theory applies again. In particular, the uML estimator is normally distributed around $\boldsymbol{\theta} = 0$ with variance $(NI(\rho|\mathscr{S}))^{-1}$, where $I(\rho|\mathscr{S})$ is the Fisher information (5) computed with respect to the unconstrained parametrisation. Moreover by Taylor expanding the log-likelihood function around the $\boldsymbol{\theta}_{\mathrm{uML}}$, and using the fact the the first derivative vanishes at this point, we obtain the second order approximation

$$\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_{\mathrm{uML}}) \approx -\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{uML}})^T I(\rho|\mathscr{S})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{uML}}). \tag{8}$$

This implies tht for large N, the ML estimator $\hat{\rho}_{\mathrm{ML}}$ is the projection of $\hat{\rho}_{\mathrm{uML}}$ onto $\mathcal{S}_d$ with respect to the quadratic distance determined by the Fisher information

$$d_I(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}'}) := (\boldsymbol{\theta} - \boldsymbol{\theta}')^T I(\rho|\mathscr{S})(\boldsymbol{\theta} - \boldsymbol{\theta}'). \tag{9}$$

Moreover, the probability distribution of the MLE is obtained by projecting the Gaussian distribution corresponding to $\hat{\rho}_{\mathrm{uML}}$, onto $\mathcal{S}_d$. Although the projection can be computed efficiently using convex optimisation, finding a general characterisation of the resulting distribution seems to be a challenging problem [70]. Nevertheless, [65] shows that the problem is tractable in those cases where the metric d_I is (approximately) isotropic, so that random matrix results such as the emergence of Wigner semicircle law can be used to treat the asymptotic theory. In section 8.1 we will use these ideas to study the asymptotic behaviour of the projected least squares estimator.

4. The least squares estimator

The least squares (LS) estimator [17, 62] is based on the observation that the outcome probabilities $p_\rho(\boldsymbol{o}|\boldsymbol{s})$ depend linearly on the state ρ. Let

$$\rho = \sum_{i=1}^{d^2} \beta_i \tau_i \tag{10}$$

be the decomposition with respect to an orthonormal basis of $M(\mathbb{C}^d)$ consisting of selfadjoint matrices $\{\tau_i : 1 \leqslant i \leqslant d^2\}$, and $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_{d^2})^T$ the corresponding vectorisation. Then the probabilities can be expressed as

$$p_\rho(\boldsymbol{o}|\boldsymbol{s}) = \mathrm{Tr}(\rho P_{\boldsymbol{o}}^{\boldsymbol{s}}) = \sum_i X_{(\boldsymbol{o}|\boldsymbol{s})}^i \beta_i, \qquad X_{(\boldsymbol{o}|\boldsymbol{s})}^i := \mathrm{Tr}(\tau_i P_{\boldsymbol{o}}^{\boldsymbol{s}})$$

where X is a $kd \times d^2$ fixed matrix depending on the measurement and the chosen vectorisation. In an experimental setup we do not have access to the true probability vector. Instead from the $d \times k$ dataset of counts, we can compute the empirical probabilities $f(\boldsymbol{o}|\boldsymbol{s}) := N(\boldsymbol{o}|\boldsymbol{s})/m$, whose expectations are $\mathbb{E}f(\boldsymbol{o}|\boldsymbol{s}) = p_\rho(\boldsymbol{o}|\boldsymbol{s})$. Replacing probabilities vector with empirical frequencies

$$\boldsymbol{f} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{11}$$

where $\epsilon \in \mathbb{R}^{dk}$ is a mean zero vector of statistical noise. The noise distribution is irrelevant for the definition of the LS estimator, but we will return to this when discussing its error, and the generalised least squares estimator.

The least squares solution to the above system of equations is defined as the minimiser of the following optimisation problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} := \arg\min_{v \in \mathbb{R}^{d^2}} \|Xv - \boldsymbol{f}\|^2 \tag{12}$$

and has the explicit form $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (X^T X)^{-1} \cdot X^T \cdot \boldsymbol{f}$. The final estimate $\hat{\rho}_{\mathrm{LS}}$ of the density matrix is then constructed from the estimated parameter vector $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ using equation (10). We note that the least squares estimator produces a state estimate $\hat{\rho}_{\mathrm{LS}} = \rho_{\boldsymbol{\beta}_{\mathrm{LS}}}$ that is not necessarily a density matrix.

Least squares as a projection. Let $\mathcal{M} : M(\mathbb{C}^d) \to \mathbb{C}^{k \cdot d}$ be the map associated to the measurement, and let $\hat{p}_{\mathrm{LS}} := \mathcal{M}(\hat{\rho}_{\mathrm{LS}}) = X \hat{\beta}_{\mathrm{LS}}$ be the 'probability distribution' corresponding to the LS estimator. This belongs to the hyperplane $\mathcal{L}_d = \mathcal{M}(M^1_{\mathrm{sa}}(\mathbb{C}^d))$ which contains the convex set of probabilities $\mathcal{P}_d = \mathcal{M}(\mathcal{S}_d)$. The LS estimator (12) can then be interpreted as the projection of the frequency vector $\boldsymbol{f}$ onto $\mathcal{L}_d$ with respect to the Euclidian distance in $\mathbb{R}^{k \cdot d}$.

4.1. Least squares as inversion of a measure-and-prepare channel

The LS estimator was introduced above by choosing a particular vectorisation of the density matrix. While this is useful for numerical implementations, the disadvantage is that one loses the physical interpretation of the quantum state and the measurement map. Conceptually, it is useful to define the LS estimator in a 'coordinate free' way which can be interpreted as the inversion of a certain measure-and-prepare map associated to the measurement [36]. Let $\mathcal{D} : M(\mathbb{C}^d) \to M(\mathbb{C}^d)$ be the quantum channel

$$\mathcal{D} : \rho \mapsto \frac{1}{|\mathscr{S}|} \sum_{o,s} \mathrm{Tr}(\rho P^s_o) \cdot P^s_o$$

which is the composition $\mathcal{P} \circ \mathcal{M}$ of the measurement with collection of bases $\mathscr{S}$, and the preparation map where the pure state P^s_o is prepared for each outcome $\boldsymbol{o}$ of the measurement in basis $\boldsymbol{s}$. If ρ is represented in its vectorised form, then the map $\mathcal{D}$ is given by the matrix $X^T X / |\mathscr{S}|$. On the other hand, the preparation map $\mathcal{P}$ has matrix X^T, so that the LS estimator can be expressed as

$$\hat{\rho}_{\mathrm{LS}} = \mathcal{D}^{-1} \circ \mathcal{P}(\boldsymbol{f}/|\mathscr{S}|). \tag{13}$$

From this expression we see immediately that $\mathrm{Tr}(\hat{\rho}_{\mathrm{LS}}) = 1$ as consequence of the fact that $\boldsymbol{f}/|\mathscr{S}|$ is a probability distribution and $\mathcal{D}$ is trace preserving. Additionally, we note the the accuracy of the LS estimator is linked to the properties of the channel $\mathcal{D}$. In particular, in the case of Pauli measurements the channel is given by a tensor product $\mathcal{D} = \mathcal{C}^{\otimes n}$ of qubit depolarising channels [17, 36]

$$\mathcal{C} : \rho \mapsto \frac{1}{3}\rho + \frac{2}{3}\frac{\boldsymbol{1}}{2}.$$

On the other hand, the measurements corresponding to the class of two-designs (which includes covariant measurements, mutually unbiased bases, stabiliser states, symmetric informationally complete measurements) have associated channel given by [36]

$$\mathcal{D} : \rho \mapsto \frac{1}{d+1}\rho + \frac{d}{d+1}\frac{\mathbf{1}}{d}.$$

4.2. Concentration bounds and asymptotic behaviour of the LS

For more insight into the structure of the LS estimator let us unpack equation (13) and note that $\hat{\rho}_{LS}$ can be written as an average of independent, identically distributed matrices $A^{(1)}, \ldots, A^{(m)}$ with $\mathbb{E}(A^{(i)}) = \rho$

$$\hat{\rho}_{LS} = \frac{1}{m}\sum_{i=1}^{m} A_i = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\mathscr{S}|}\sum_{\mathbf{s}\in\mathscr{S}}\mathcal{D}^{-1}(P_{\mathbf{o}_{i|\mathbf{s}}}^{\mathbf{s}}) \tag{14}$$

where $\mathbf{o}_{i|\mathbf{s}}$ are the outcomes of the measurements with respect to setting $\mathbf{s}$.

In this form, the LS estimator is amenable to non-asymptotic matrix concentration techniques, as well as asymptotic normality theory. The following result [36] was obtained by applying matrix Bernstein inequalities [75] and provides a non-asymptotic confidence bound for LS with respect to the operator norm distance, see also [17].

Theorem 4.1. *Let $\hat{\rho}_{LS}$ be the LS estimator of ρ for a dataset consisting of $N = m \times k$ samples. Then*

$$\Pr\left[\|\hat{\rho}_{LS} - \rho\| \geqslant \epsilon\right] \leqslant d e^{-\frac{3N\epsilon^2}{8g(d)}} \quad \epsilon \in [0,1]$$

where $g(d) = 2d$ for two-design measurements and $g(d) \simeq d^{1.6}$ for Pauli measurements.

The theorem provides upper bounds for risks with respect to commonly used loss functions such as the Frobenius distance (norm-two squared) and the trace-norm distance, see also [71] for related results. Indeed using the basic inequalities $\|A\|_2^2 \leqslant d\|A\|^2$ and $\|A\|_1 \leqslant d\|A\|$ we obtain the upper bounds

$$\mathbb{E}\|\rho - \hat{\rho}_{LS}\|_2^2 \leqslant c_2 \log(d)\frac{d^2}{N}, \qquad \mathbb{E}\|\rho - \hat{\rho}_{LS}\|_1 \leqslant c_1 \log(d)\frac{d\sqrt{d}}{\sqrt{N}} \tag{15}$$

for the two-design measurements and

$$\mathbb{E}\|\rho - \hat{\rho}_{LS}\|_2^2 \leqslant C_2 \log(d)\frac{d^{2.6}}{N}, \qquad \mathbb{E}\|\rho - \hat{\rho}_{LS}\|_1 \leqslant C_1 \log(d)\frac{d^{1.8}}{\sqrt{N}} \tag{16}$$

for the Pauli basis measurement, where c_1, c_2, C_1, C_2 are a numerical constants. Moreover, in [36] it was shown that the log factor in (15) can be removed when we deal with covariant measurements. On the other hand, in section 3 we have shown that for the maximally mixed state, the optimal convergence rate for the Frobenius distance is of the order $O(d^2/N)$; this means that the upper bound (15) cannot be improved, when seen as a uniform bound over *all* states. However, this leaves open the possibility that special classes of states can be estimated with higher accuracy that that provided by the LS estimator. Indeed, we will see that simple modifications of the LS estimator which take into account the positivity of the unknown state can significantly reduce the estimation errors for low rank states.

Asymptotics. Thanks to its simplicity, the LS estimator has a tractable asymptotic theory. As in the case of the concentration Theorem 4.1 the key point is that the error ρ_{LS} is a sum of independent, identically distributed matrices given by equation (14). In the limit of large m each

matrix element of ρ_{LS} has a Gaussian distribution; in terms of the vectorised form we have the Central Limit for $m \to \infty$

$$\sqrt{m}\left[\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}\right] \longrightarrow N(0, V_{\mathrm{LS}}), \quad V_{\mathrm{LS}} = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$$

where $\Omega := m \mathrm{Cov}\left[\boldsymbol{\epsilon}|\mathscr{S}\right]$ is the (renormalised) covariance of the noise $\boldsymbol{\epsilon}$. Due to the independent structure of the measurement settings, the latter has a non-trivial block diagonal form with $d \times d$ blocks corresponding to the different settings $\boldsymbol{s}$

$$[\Omega_{\boldsymbol{s}}]_{ij} = \mathrm{Cov}\left[\epsilon_{\boldsymbol{s}}\right]_{ij} = p_\rho(\boldsymbol{o}_i|\boldsymbol{s})\delta_{ij} - p_\rho(\boldsymbol{o}_i|\boldsymbol{s})\,p_\rho(\boldsymbol{o}_j|\boldsymbol{s}). \tag{17}$$

This allows us in principle to compute the asymptotic risk of an arbitrary quadratic loss function with weight matrix G (such as the Frobenius distance) as

$$m\mathbb{E}\left[(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta})^T G(\hat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta})\right] \longrightarrow \mathrm{Tr}(GV_{\mathrm{LS}}).$$

Since $\boldsymbol{f}$ is a vector of frequencies, it must satisfy k normalisation constraints, which is reflected in the fact each block $\Omega_{\boldsymbol{s}}$ is singular, with zero eigenvector $\boldsymbol{1} = (1, \ldots, 1)^T$. Consequently, the covariance matrix V_{LS} is singular with the zero eigenvector corresponding to the trace which is a fixed parameter. We will come back to this when discussing the generalised least squares estimator.

4.3. Asymptotic theory of LS for covariant measurements

While asymptotic normality tells us that the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ lies in an ellipsoid centred at $\boldsymbol{\beta}$, it treats the estimator as a vector rather that as a matrix. For instance, it does not immediately provide an asymptotic theory for the operator norm error $\|\hat{\rho}_{\mathrm{LS}} - \rho\|$. Random matrix theory provides us with other asymptotic results which take into account the matrix structure of the estimator. To explore this avenue we will make a simplifying assumption and place ourselves in the scenario of covariant measurements. The outcome of such a measurement is a one dimensional projection $P = |\psi\rangle\langle\psi|$ and the corresponding positive operator valued measure is

$$M(\mathrm{d}P) := d \cdot P \cdot \mathrm{d}P \tag{18}$$

where $\mathrm{d}P$ is the uniform measure over the space of one dimensional projections; the latter is the measure induced by the Haar measure over the unitaries U via the action $P \mapsto UPU^*$. An alternative way of obtaining a measurement sample is to choose a random basis and perform a measurement with respect to the chosen basis.

In this setup, the channel $\mathcal{D}$ is given by [36]

$$\mathcal{D} : \rho \mapsto \int \mathrm{Tr}(\rho M(\mathrm{d}P))P = \int \mathrm{d}P \mathrm{Tr}(\rho P)\mathrm{d}P = \frac{1}{d+1}(\rho + \boldsymbol{1})$$

and the LS estimator given by equation (14) can be written as

$$\hat{\rho}_{\mathrm{LS}} = \frac{d+1}{N} \sum_{i=1}^{N} \left(P_i - \frac{\boldsymbol{1}}{d+1}\right)$$

where P_i are the independent outcomes of measurements.

We will be particularly interested in the behaviour of the LS estimator for low rank states, as well as states close to the maximally mixed one. Due to covariance it suffices to choose states which are diagonal in the standard basis $\{|i\rangle : i = 1, \ldots, d\}$, and for simplicity we will

restrict ourselves to rank-r states with equal eigenvalues $\rho_r = \sum_{i=1}^{r} |i\rangle\langle i|/r$. As in section 3.1 we write the LS estimator as

$$\hat{\rho}_{LS} = \frac{1}{r}\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\sqrt{N}}\begin{pmatrix} A & B \\ B^* & C \end{pmatrix}, \tag{19}$$

where A, B, C are block matrices of mean zero, and I_r is the $r \times r$ identity block. We will deal with each block separately.

Block C. By covariance, the distribution of the block C is invariant under unitary transformations in $\mathbb{C}^{d-r}$. As the LS estimator is unbiased, the matrix elements of C are centred. The off-diagonal elements have real and imaginary parts with variances equal to

$$\text{Var}(\text{Re}C_{ij}) = \text{Var}(\text{Im}C_{ij}) = \frac{1}{2}\mathbb{E}|C_{ij}|^2 = \frac{v_c}{2}$$

with

$$v_c = d(d+1)^2 \int \langle Uk|\rho_r|Uk\rangle \cdot |\langle i|U|k\rangle|^2 \cdot |\langle k|U^*|j\rangle|^2 \, dU$$

$$= d(d+1)^2 \int U_{1k}U_{ik}U_{jk}U_{k1}^*U_{ki}^*U_{kj}^* \, dU$$

$$= \frac{d(d+1)^2}{d(d+1)(d+2)} = \frac{d+1}{d+2}$$

where we have written $P = U|k\rangle\langle k|U^*$ for $r < k \neq i,j$ and replaced the integration over projections with that over unitaries; the integral was then evaluated using Weingarten formulas [19]. Similarly, the variance and covariance of the diagonal elements are

$$\text{Var}(C_{ii}) = \frac{d}{d+2} \qquad \text{Cov}(C_{ii}, C_{jj}) = -\frac{1}{d+2}$$

and all off-diagonal elements (of all blocks A, B, C) have zero covariance with other matrix elements. By the Central Limit theorem, in the limit of large N, the off-diagonal elements of $\hat{\rho}_{LS}$ become normally distributed and independent of all other elements. On the other hand, as the covariance of diagonal elements is non-zero, they converge to correlated Gaussian variables. However, if we also take the limit of large d we find that the covariance of the diagonal elements vanishes and the matrix C is distributed approximately as the Gaussian unitary ensemble (GUE). Universality results for random matrices [64] imply that the empirical distribution of the eigenvalues of $C/\sqrt{d-r}$ converges weakly to Wigner's semicircle law on the interval $[-2,2]$, whose probability density is

$$w(x) = \frac{1}{2\pi}\sqrt{4-x^2}. \tag{20}$$

Indeed, panel (a) of figure 2 shows a good match between the histogram of the eigenvalues of the error block $C/\sqrt{N}$ for a rank-one state of $n = 7$ atoms and $N = 10^6$ samples, and the corresponding Wigner distribution.

Block A. Let us consider now block A of equation (19). For the same symmetry reasons, its off-diagonal elements (when $r > 1$) have real and imaginary parts with variance

$$\text{Var}(\text{Re}A_{ij}) = \text{Var}(\text{Im}A_{ij}) = \frac{1}{2}\mathbb{E}|A_{ij}|^2 = \frac{v_a}{2}$$

with

$$
v_a = d(d+1)^2 \int \langle Uk | \rho_r | Uk \rangle \cdot |\langle i | U | k \rangle|^2 \cdot |\langle k | U^* | j \rangle|^2 \, \mathrm{d}U
$$

$$
= \frac{2d(d+1)^2}{r} \int U_{ik} U_{ik} U_{jk} U_{ki}^* U_{ki}^* U_{kj}^* \, \mathrm{d}U + \frac{(r-2)d(d+1)^2}{r} \int U_{lk} U_{ik} U_{jk} U_{ki}^* U_{kl}^* U_{kj}^* \, \mathrm{d}U
$$

$$
= \frac{2d(d+1)^2}{r} \cdot \frac{2}{d(d+1)(d+2)} + \frac{(r-2)d(d+1)^2}{r} \cdot \frac{1}{d(d+1)(d+2)} = \frac{r+2}{r} \cdot \frac{d+1}{d+2}
$$

where all indices i, j, k, l are different. The diagonal elements have variance

$$
\mathrm{Var}(A_{ii}) = 2 \frac{d+1}{d+2} \frac{r+2}{r} - \left(1 + \frac{1}{r} \right)^2
$$

while the covariance of the diagonal elements is

$$
\mathrm{Cov}(A_{ii}, A_{jj}) = \frac{d+1}{d+2} \frac{r+2}{r} - \left(1 + \frac{1}{r} \right)^2 .
$$

We note that for small r the diagonal elements do not become independent when N and d are large. However, if we also take r to be large, the covariance vanishes, as in the case of the C block. By the same argument, the distribution of the eigenvalues of $A/\sqrt{r}$ converges to the Wigner semicircle law (20). This is illustrated in the right panel of figure 2, for the case of the totally mixed state of $n = 7$ atoms.

Block B. The elements of block B have variances equal to

$$
\mathrm{Var}(\mathrm{Re} B_{ij}) = \mathrm{Var}(\mathrm{Im} B_{ij}) = \frac{1}{2} \mathbb{E} |B_{ij}|^2 = \frac{v_b}{2}
$$

with $i \leqslant r < j$ and

$$
v_b = d(d+1)^2 \int \langle Uk | \rho_r | Uk \rangle \cdot |\langle i | U | k \rangle|^2 \cdot |\langle k | U^* | j \rangle|^2 \, \mathrm{d}U
$$

$$
= \frac{d(d+1)^2}{r} \int U_{ik} U_{ik} U_{jk} U_{ki}^* U_{ki}^* U_{kj}^* \, \mathrm{d}U + \frac{(r-1)d(d+1)^2}{r} \int U_{lk} U_{ik} U_{jk} U_{k1}^* U_{kl}^* U_{kj}^* \, \mathrm{d}U
$$

$$
= \frac{d(d+1)^2}{r} \cdot \frac{2}{d(d+1)(d+2)} + \frac{(r-1)d(d+1)^2}{r} \cdot \frac{1}{d(d+1)(d+2)} = \frac{r+1}{r} \cdot \frac{d+1}{d+2} .
$$

Eigenvalues distribution of the LS estimator. To better understand the LS estimator, it is instructive to study the distribution of its eigenvalues. This will also be used in the study of the projected least squares estimator furher on. Assuming N is large, we note that $\hat{\rho}_{\mathrm{LS}}$ can be written as

$$
\hat{\rho}_{\mathrm{LS}} = U \left(\frac{I_r}{r} + \frac{D}{\sqrt{N}} \right) U^* + o(N^{-1/2}) \tag{21}
$$

where

$$
U = \exp \left(\frac{\mathrm{i} r \tilde{B}}{\sqrt{N}} \right), \quad \tilde{B} := \begin{pmatrix} 0 & \mathrm{i} B \\ -\mathrm{i} B^* & 0 \end{pmatrix}, \quad D := \begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix} .
$$

This means that in the leading order in $N^{-1/2}$ the set of eigenvalues of $\hat{\rho}_{\mathrm{LS}}$ is the union of those of $I_r/r + A/\sqrt{N}$ and $C/\sqrt{N}$. In the case of large N, d and small rank r, the first group of eigenvalues are small fluctuations of size $r/\sqrt{N}$ around $1/r$; in particular these eigenvalues

are positive with high probability; the second group of eigenvalues are distributed according to the Wigner law, and have maximum absolute value approximately equal to $2\sqrt{(d-r)/N}$, and roughly half of them will be negative.

We can now evaluate the asymptotic risk of the LS estimator with respect to the chosen loss functions.

Frobenius risk. For the Frobenius distance, the asymptotic mean square error is

$$
\begin{aligned}
N\mathbb{E}\|\hat{\rho}_{\mathrm{LS}} - \rho_r\|_2^2 &= \mathbb{E}\|A\|_2^2 + 2\mathbb{E}\|B\|_2^2 + \mathbb{E}\|C\|_2^2 \\
&= \sum_{i=1}^{r} \mathrm{Var}(A_{ii}) + 2\sum_{i<j\leqslant r} \mathbb{E}|A_{ij}|^2 + 2\sum_{i\leqslant r<j} \mathbb{E}|B_{ij}|^2 \\
&\quad + \sum_{r<i<j} 2\mathbb{E}|C_{ij}|^2 + \sum_{i=r+1}^{d} \mathrm{Var}(C_{ii}) \\
&= (r+1)(r+2)\frac{d+1}{d+2} - \frac{(r+1)^2}{r} + 2\frac{(r+1)(d-r)(d+1)}{d+2} \\
&\quad + \frac{(d-r)(d-r-1)(d+1)}{d+2} + \frac{(d-r)d}{d+2}.
\end{aligned}
$$

In particular, the leading contribution to the Frobenius risk is d^2/N, and the dependence with r is weak. As illustrated in panels (a) and (b) of figure 3 the theoretical prediction match the simulation results for 5 and 6 atom states.

Operator norm risk. The operator-norm error of the LS estimator is

$$
\|\hat{\rho}_{\mathrm{LS}} - \rho_r\| = \frac{1}{\sqrt{N}}\left\|\begin{pmatrix} A & B \\ B^* & C \end{pmatrix}\right\|.
$$

Above we found that the entries of A, B, C become independent in the limit of large N, except for correlations between the diagonal elements which vanish if we additionally take the limit of large d. Moreover, the variances of the elements in the B block and the off-diagonal elements of A differ from those of the off-diagonal elements of C by factors $(r+2)/r$ and respectively $(r+1)/r$. Therefore, the error block matrix as a whole is not distributed according to GUE ensemble. However, the universality of the Wigner semicircle law, the limit holds not only for highly symmetric ensembles like GUE but also for 'small' perturbations of this ensemble, e.g. random matrices with independent entries, whose variances do not deviate too much from a fixed value [23]. In particular for low rank $r \ll d$, the total size of the blocks A,B,B^* is of the order rd which is much smaller than than the size $(d-r)^2$ of C. Therefore, the asymptotic behaviour of the error matrix is determined by that of C and its spectrum converges to the Wigner law in the limit of large sample size and large dimension. In particular, the leading contribution to the norm error $\|\hat{\rho}_{\mathrm{LS}} - \rho_r\|$ for low rank states is $2\sqrt{d/N}$. A similar situation occurs for high rank states $r \approx d$ where the dominant variances are provided by the block A.

Indeed simulations for a pure and the fully mixed state of $n=7$ atoms with $N=10^6$ samples gave norm errors 0.0225 and respectively 0.0221, while the above theoretical prediction is 0.0226. Further simulation results for $n=5$ and $n=6$ atoms states of different ranks show a good match with the above estimate.

Trace norm risk. As in the case of the operator norm, the trace norm error can be estimated thanks to the fact that the eigenvalues of error matrix follow an approximate Wigner semicircle distribution. Using

$$\frac{1}{2\pi} \int_{-2}^{2} |x| \sqrt{4 - x^2} \mathrm{d}x = \frac{8}{3\pi}$$

we find that for low rank, or close to full rank states, the leading contribution to $\mathbb{E}(\|\hat{\rho}_{\mathrm{LS}} - \rho\|_1)$ is $8d^{3/2}/(3\pi\sqrt{N})$. This rate agrees with the upper bound in equation (15), but provides an exact asymptotic constant for this rate. Indeed simulations with a pure and maximally mixed states of $n = 7$ atoms and $N = 10^6$ samples gave norm-one errors of 1.228 and respectively 1.229 while the theoretical prediction is 1.229.

5. Generalised least squares estimator

Consider the generic linear regression problem

$$\mathbf{y} = A\mathbf{x} + \mathbf{n} \tag{22}$$

with $\mathbf{x} \in \mathbb{R}^a$ and unknown vector, $\mathbf{y} \in \mathbb{R}^b$ a vector of observations and $\mathbf{n}$ the 'noise' term with *fixed and known* covariance matrix C, which is assumed to be strictly positive. While the LS estimator $\hat{\mathbf{x}}_{\mathbf{LS}}$ is optimal in the sense of minimising the *prediction error* $\|A\hat{\mathbf{x}} - \mathbf{y}\|$, this is not the case when considering an *estimation error* e.g. the mean square error $\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$, unless the covariance matrix is proportional to the identity. To recover the 'equal noise' situation we multiply equation (22) from the left by $C^{-1/2}$, to obtain

$$\mathbf{y}' = A'\mathbf{x} + \mathbf{n}', \qquad \mathrm{Cov}(\mathbf{n}') = \mathrm{Id}.$$

The LS estimator of the last regression equation has the smallest covariance matrix among linear unbiased estimators

$$\hat{\mathbf{x}}_{\mathrm{GLS}} = \arg\min_{\mathbf{x}'} \|\mathbf{y}' - A'\mathbf{x}'\|^2 = (A^T C^{-1} A)^{-1} A^T C^{-1} \mathbf{y}.$$

This is called the generalised least squares (GLS) estimator when the noise distribution is Gaussian, the GLS estimator coincides with the MLE. In the i.i.d. setting where m samples are available, the GLS estimator is asymptotically normal and efficient

$$\sqrt{m}(\hat{\mathbf{x}}_{\mathrm{GLS}} - \mathbf{x}) \to N\left(\mathbf{0}, (A^T C^{-1} A)^{-1}\right). \tag{23}$$

GLS as a projection. Similarly to the LS estimator we consider the image $\hat{\mathbf{y}}_{\mathrm{GLS}} = A\hat{\mathbf{x}}_{\mathrm{GLS}}$ of the GLS estimator $\mathbf{x}_{\mathrm{GLS}}$. Then $\hat{\mathbf{y}}_{\mathrm{GLS}}$ is the projection of the data $\mathbf{y}$ onto the subspace $\mathrm{Ran}(A) \subset \mathbb{R}^b$ with respect to the covariance-dependent metric

$$d_C(\mathbf{y}, \mathbf{z}) = (\mathbf{y} - \mathbf{z})^T C^{-1}(\mathbf{y} - \mathbf{z}).$$

GLS for tomography. Let us return now to the tomography regression problem of the form given by equation (11). While the LS estimator is optimal in the sense of equation (12), in general it is not optimal in the estimation sense for any locally quadratic distance function. To remedy this we would like to construct a corresponding generalised least square estimator to take into account the nontrivial form of the noise covariance matrix Ω given by equation (17). However, there are two issues which prevent us from directly applying the GLS methodology to the tomography data. The first is that Ω is unknown since it depends on the true probabilities, and therefore on the unknown state ρ. We therefore propose to use an estimate of the covariance matrix instead. We describe the computation of this estimate in appendix A.2. The second difficulty is that Ω is a singular matrix due the constraint $\sum_o f(o|s) = 1$ for each setting $\mathbf{s}$. This means that for each setting $\mathbf{s}$ the vector $|\mathbf{1_s}\rangle \in \mathbb{C}^d$ is a zero eigenvector for $\Omega_\mathbf{s}$ and one should work within the orthogonal complement of $|\mathbf{1_s}\rangle$. This can be achieved by choosing

a block-diagonal isometry $V : \mathbb{C}^{k(d-1)} \to \mathbb{C}^{k \cdot d}$ such that each block $V_{\mathbf{s}}$ satisfies $\langle \mathbf{1}_s | V_{\mathbf{s}} = 0$, and defining the new 'frequencies' vector $\tilde{\boldsymbol{f}}$ with settings components $\tilde{\boldsymbol{f}}_{\mathbf{s}} = V_{\mathbf{s}}^* \boldsymbol{f}_{\mathbf{s}}$. Similarly, we can also remove the trace-one constraint for states by choosing the vectorisation (10) such that the last basis element is $\tau_{d^2} = \mathbb{1}/\sqrt{d}$. In this case the state is uniquely determined by the free parameters $\{\beta_i : 1 \leqslant i \leqslant d^2 - 1\}$. If we denote by $J : \mathbb{C}^{d^2-1} \to \mathbb{C}^{d^2}$ the isometric embedding with respect to the standard basis, then $\tilde{\boldsymbol{\beta}} := J^* \boldsymbol{\beta}$ is the truncated vector of free parameters for ρ.

The regression problem can now be written in terms of the 'tilde' vectors and matrices

$$\tilde{\boldsymbol{f}} = \tilde{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}}$$

where $\tilde{X} = V^* X J$ and $\tilde{\boldsymbol{\epsilon}}$ has covariance matrix $\tilde{\Omega} = V^* \Omega V$ which is *non-singular*. We can now apply the GLS methodology and define the estimator as

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\tilde{X}^* (\hat{\tilde{\Omega}})^{-1} \tilde{X})^{-1} \tilde{X}^* (\hat{\tilde{\Omega}})^{-1} \tilde{\boldsymbol{f}},$$

where $\hat{\tilde{\Omega}}$ is the (non-singular) estimated covariance matrix. We denote $\hat{\rho}_{\mathrm{GLS}}$ as estimate of the density matrix constructed from its vectorised form $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$. For future use, let us denote $\tilde{\mathcal{M}} : M^1_{\mathrm{sa}}(\mathbb{C}^d) \to \mathbb{R}^{k(d-1)}$ the map $\rho \mapsto \tilde{X}\tilde{\boldsymbol{\beta}}$, so that $\hat{\tilde{p}}_{\mathrm{GLS}} = \tilde{\mathcal{M}}(\hat{\rho}_{\mathrm{GLS}})$.

5.1. Asymptotic theory of GLS

What are the asymptotic properties of $\hat{\rho}_{\mathrm{GLS}}$? For large m, the noise distribution becomes Gaussian, and the covariance estimator converges to the actual covariance. The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ becomes asymptotically normal as $m \to \infty$

$$\sqrt{m}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta}) \to N\left(\mathbf{0}, (\tilde{X}^* \tilde{\Omega}^{-1} \tilde{X})^{-1}\right) \tag{24}$$

which means that the corresponding asymptotic average Fisher information per sample is $I_{\mathrm{GLS}} = \tilde{X}^* \tilde{\Omega}^{-1} \tilde{X}/k$. In appendix A.1 we show that I_{GLS} coincides with the Fisher information $I(\rho|\mathscr{S})$ defined as in equation (5) for the parametrisation $\tilde{\boldsymbol{\beta}}$. This equality implies that the error rates of the GLS estimator have the same asymptotic behaviour as those of the uML estimator, and both estimators satisfy asymptotic normality. In fact one can make the stronger statement that the two estimators are asymptotically close to each other.

Equivalence of GLS and uML. As stated above, the GLS can be interpreted as the projection of the data $\tilde{\boldsymbol{f}}$ onto the image of $\tilde{X}$ in $\mathbb{R}^{k(d-1)}$ with respect to the metric

$$d(\boldsymbol{g}, \boldsymbol{h}) = (\boldsymbol{g} - \boldsymbol{h})^T \hat{\tilde{\Omega}}^{-1} (\boldsymbol{g} - \boldsymbol{h})^T.$$

On the other hand, in section 3.1 we showed that for large N, we can define the uML estimator as the projection of $\boldsymbol{f}$ onto the hyperplane $\mathcal{L}_d := \mathcal{M}(M^1_{\mathrm{sa}}(\mathbb{C}^d))$, with respect to the relative entropy distance. Since in the first order of approximation the relative entropy is given by the quadratic form $\tilde{\Omega}^{-1}$, the two projections become identical in the asymptotic limit.

6. Thresholded least squares estimator

The LS estimator (as well as GLS) suffers from the disadvantage that it does not necessarily produce a density matrix, i.e, a positive semi-definite estimate of trace one. While the significant eigenvalues can be estimated reasonably well with enough data, the LS estimator will

typically have negative eigenvalues corresponding to small or zero eigenvalues of the true state. The thresholded least squares estimator (TLS) proposed in [17], improves the LS estimator by selecting the state which is the closest in Frobenius norm to $\hat{\rho}_{LS}$ and whose non-zero eigenvalues are above a certain threshold $\nu \geqslant 0$, i.e.

$$\hat{\rho}_{TLS} := \underset{\tau \in \mathcal{S}_d^{(\nu)}}{\arg\max} \|\hat{\rho}_{LS} - \tau\|_2^2$$

with $\mathcal{S}_d^{(\nu)}$ the set of density matrices with spectrum in $\{0\} \cup [\nu, 1]\}$. The choice of the statistical noise threshold is informed by the accuracy of the LS estimate, and a theoretical and a 'data-driven' choices for this threshold are detailed in [17], see also [22, 68]. In practice it is found that estimator's performance improves if the threshold is allowed to be 'data-driven', by using cross-validation to choose the optimal value of the threshold, see appendix A.2.

It turns out that computing $\hat{\rho}_{TLS}$ is very efficient and can be easily implemented. Let

$$\hat{\rho}_{LS} = \sum_{i=1}^{d} \hat{\lambda}_i |\hat{g}_i\rangle\langle\hat{g}_i|$$

be the spectral decomposition of $\hat{\rho}_{LS}$ where we assume that the eigenvalues are sorted in descending order $\hat{\lambda}_1 \geqslant \ldots \geqslant \hat{\lambda}_d$. The thresholded estimator

$$\hat{\rho}_{TLS} = \sum_{i=1}^{d} \hat{\lambda}_i(\nu) |\hat{g}_i\rangle\langle\hat{g}_i|$$

has the same eigenvectors as $\hat{\rho}_{TLS}$, and its eigenvalues can be computed in terms of $\hat{\lambda}_i$ as summarised in algorithm 1. In words, the algorithm checks if the smallest eigenvalue of $\hat{\rho}_{LS}$ is above the noise threshold, and if it is then $\hat{\rho}_{TLS} = \hat{\rho}_{LS}$ (note that by construction $\text{Tr}(\hat{\rho}_{LS}) = 1$, so there is no need to renormalise the LS estimator). On the other hand if the smallest eigenvalue is below the threshold, it is set to zero and the remaining eigenvalues are suitably shifted accordingly so that their sum is equal to one. The final estimate $\hat{\rho}_{TLS}$ is constructed by replacing the eigenvalues of $\hat{\rho}_{LS}$ with these thresholded eigenvalues $\hat{\lambda}_i(\nu)$. The theoretical properties of the TLS estimator are presented in [17] and are similar to those of the projected least squares estimator which is discussed in more detail below.

Algorithm 1. Algorithm to threshold the eigenvalues of the LS estimate.

Input: eigenvalues of LS estimator $\hat{\lambda}_1 \geqslant \ldots \geqslant \hat{\lambda}_d$, and noise threshold ν
Output: eigenvalues $\hat{\lambda}_1(\nu) \geqslant \ldots \geqslant \hat{\lambda}_d(\nu)$ of thresholded estimate $\hat{\rho}_{TLS}$
for $p = 1, \ldots, d$ **do**
    if $\hat{\lambda}_{d-p+1} \geqslant \nu$ **then**
        STOP;
    else
        $\hat{\lambda}_{d-p+1} \leftarrow 0$;
        for $j = 1, \ldots, d - p$ **do**
            $\hat{\lambda}_j \leftarrow \hat{\lambda}_j + \frac{1}{d-p}\left(1 - \sum_{m=1}^{d-p} \hat{\lambda}_m\right)$
        end
    end
end

6.1. Thresholded generalised least squares estimator (TGLS)

This estimator is obtained by using the GLS estimate $\hat{\rho}_{\text{GLS}}$ instead of the LS estimate as a starting point for the thresholding procedure. The constant for thresholding is chosen in the same way by using cross-validation. The advantage of TGLS is that it starts from a more accurate estimator than LS, which leads to smaller errors compared to TLS. The price to pay is that it is somewhat more involved computationally, although still faster than ML.

7. Positive least squares estimator

The positive least squares (posLS) estimator is obtained by restricting the minimisation in (12) to parameters that correspond to density matrices $\tau \in \mathcal{S}_d$. Let $\mathcal{M} : M(\mathbb{C}^d) \mapsto \mathbb{C}^{kd}$ be the measurement map defined as $[\mathcal{M}(\tau)]_{o,s} = \text{Tr}\left[\tau P_o^s\right]/|\mathcal{S}|$, Then, similarly to section 4.1 we can express the posLS estimator as

$$\hat{\rho}_{\text{posLS}} := \arg\min_{\tau \in \mathcal{S}_d} \|\mathcal{M}(\tau) - \boldsymbol{f}\|^2. \tag{25}$$

By comparing with (12) we see that $\hat{\rho}_{\text{posLS}}$ is the projection of $\hat{\rho}_{\text{LS}}$ with respect to the distance on matrices induced by the euclidian distance on measurement probabilities $d_{\mathcal{M}}(\rho, \tau) = \|\mathcal{M}(\rho) - \mathcal{M}(\tau)\|$. To our knowledge, with the exception of the results in [46], its theoretical properties have not been studied in detail. While its statistical performance greatly improves on that of the LS, the posLS estimator has the drawback that it cannot be expressed in a closed form, and its computational complexity is comparable to that of the ML estimator.

7.1. Positive least squares estimator for covariant measurements

Let us consider the special case of the covariant measurement defined in equation (18). This measurement maps states into probability distributions in an (almost) isometric way [36]

$$\|\mathcal{M}(\rho) - \mathcal{M}(\tau)\|^2 = \frac{d}{d+1}\|\rho - \tau\|_2^2.$$

Indeed using Weingarten formulas [19] we get for any traceless X

$$\|\mathcal{M}(X)\|^2 = d^2 \int \mathrm{d}P \text{Tr}(PX)^2 = d^2 \int \mathrm{d}U |\langle U1|X|U1\rangle|^2$$

$$= d^2 \sum_{ijkl} X_{ij}\bar{X}_{kl} \int U_{j1}U_{k1}U_{1i}^*U_{1l}^* \,\mathrm{d}U = \frac{d}{d+1}\|X\|_2^2.$$

As a corollary, we find that the posLS estimator is the projection of the LS estimator with respect to the Frobenius distance. Therefore, for this specific measurement, the posLS estimator coincides with the projected least squares estimator which will be discussed in section 8.

7.2. Positive generalised least squares estimator

This estimator is defined in much the same way as the posLS estimator, by restricting the minimisation in (5) to parameters that correspond to density matrices. In keeping with the discussion in section 5, we consider the truncated frequency vector $\tilde{\boldsymbol{f}}$. By analogy with the GLS estimator, the positive generalised least squares (posGLS) estimator is defined as

$$\hat{\rho}_{\mathrm{posGLS}} := \underset{\tau \in \mathcal{S}_d}{\arg\min} \left\| \hat{\tilde{\Omega}}^{-1/2} \left(\tilde{\mathcal{M}}(\tau) - \tilde{f} \right) \right\|^2. \tag{26}$$

We will show that $\hat{\rho}_{\mathrm{posGLS}}$ is asymptotically equivalent to the ML estimator $\hat{\rho}_{\mathrm{ML}}$.

Asymptotic equivalence of posGLS and ML. As in the case of GLS and uML, it is easier to work with the images in $\mathbb{C}^{k(d-1)}$ of the different estimators through the (injective) map $\tilde{\mathcal{M}}$; we denote by $\hat{\tilde{p}}_{\mathrm{posGLS}} = \tilde{\mathcal{M}}(\hat{\rho}_{\mathrm{posGLS}})$ the probability vector corresponding to the posGLS estimator, and similarly for other estimators. Let us equip the the space of 'frequencies' $\mathbb{C}^{k(d-1)}$ with the distance

$$d_\Omega(\tilde{p}, \tilde{q}) = \left\| \tilde{\Omega}^{-1/2} (\tilde{p} - \tilde{q}) \right\|^2 = (\tilde{p} - \tilde{q})^T \tilde{\Omega}^{-1} (\tilde{p} - \tilde{q}).$$

Therefore, $\hat{\tilde{p}}_{\mathrm{posGLS}}$ is the projection with respect to d_Ω of $\tilde{f}$ onto the convex set

$$\mathcal{P}_d = \tilde{\mathcal{M}}(\mathcal{S}_d) \subset \tilde{\mathcal{M}}(M_{\mathrm{sa}}^1(\mathbb{C}^d)) \subset \mathbb{C}^{k(d-1)}.$$

On the other hand, the GLS estimator is the projection of $\tilde{f}$ onto the hyperplane $\tilde{\mathcal{M}}(M_{\mathrm{sa}}^1(\mathbb{C}^d))$ which contains $\mathcal{P}_d$. Therefore, by properties of projections on convex subsets we find that $\hat{\tilde{p}}_{\mathrm{posGLS}}$ is the projection of $\hat{\tilde{p}}_{\mathrm{GLS}}$ onto $\mathcal{P}_d$.

In section 5.1 we showed that the GLS estimator is asymptotically equivalent to the uML, as a consequence of the fact asymptotically with m both projections are determined by the Fisher information metric. Since posGLS and ML are obtained by applying the same projections onto the smaller space $\mathcal{P}_d$, they are also asymptotically equivalent. Unfortunately, this equivalence does not provide us with general estimation method which is more efficient than ML; the projection involved in posGLS does not seem to have closed form expression and requires a similar optimisation process as ML.

8. Projected least squares

Recently, it has been shown that the theoretical properties $\hat{\rho}_{\mathrm{TLS}}$ are preserved even if the threshold ν is chosen to be zero [36]. The *projected least squares* (PLS) estimator is defined as

$$\hat{\rho}_{\mathrm{PLS}} := \underset{\tau \in \mathcal{S}_d}{\arg\inf} \| \hat{\rho}_{\mathrm{LS}} - \tau \|_2$$

where the optimisation is performed over all states τ. As in the case of the TLS estimator, this optimisation can be performed efficiently, and it only involves computing the spectral decomposition of $\hat{\rho}_{\mathrm{LS}}$ and applying algorithm 1 with $\nu = 0$. The PLS estimator is therefore faster than the data-driven TLS and turns out to have quite similar behaviour to the latter. Note that in general PLS is different from posLS, as both can be seen as projections of the LS estimator with respect to different metrics. However, as noted before, the estimators coincide in the case of covariant measurements. In the next section we will study this scenario in more detail.

Using the LS concentration bound of theorem 4.1, the following *rank dependent* norm-one bound for the PLS estimator was derived in [36], where the measurement is either a two-design or the Pauli bases measurement.

Theorem 8.1. *Let $\hat{\rho}_{\mathrm{PLS}}$ be the PLS estimator of ρ for a dataset consisting of $N = m \times k$ samples. Then*

$$\Pr\left[\|\hat{\rho}_{\mathrm{PLS}} - \rho\|_1 \geqslant \epsilon \right] \leqslant \mathrm{d}e^{-\frac{N\epsilon^2}{43g(d)r^2}} \quad \tau \in [0,1]$$

where $g(d) = 2d$ for two-design measurements and $g(d) \simeq d^{1.6}$ for Pauli measurements.

This gives a upper bound of $O(r^2 \cdot d \log d / \sqrt{N})$ on the convergence rate of norm-one for two-design measurements and $O(\log r^2 \cdot d^{1.6} \log d / \sqrt{N})$ for Pauli measurements. In the first case the $\log d$ can be removed for covariant measurements and the resulting rate is optimal in the sense that it achives general lower bounds from [37].

8.1. The asymptotic behaviour of PLS for covariant measurements

In this section we look in more detail at the PLS estimator in the context of covariant measurements defined in equation (18). We have already seen that theorem 8.1 provides a concentration bound on the norm-one risk of the PLS estimator. While such results are very valuable thanks to their non-asymptotic nature, it is instructive and useful to also understand the asymptotic behaviour for large sample size N and dimension d. Indeed, in section 4.3 we showed how central limit and random matrix theory can be used to obtain tighter bounds on estimation risks of the LS estimator.

We will be particularly interested in the behaviour of PLS for *low rank states*. Due to covariance it suffices to choose states which are diagonal in the standard basis $\{|i\rangle : i = 1, \ldots, d\}$, and for simplicity we will restrict ourselves to rank-r states with equal eigenvalues $\rho_r = \sum_{i=1}^{r} |i\rangle\langle i|/r$. As in section 4.3 we write the PLS estimator as

$$\hat{\rho}_{\text{PLS}} = \rho_r + \frac{1}{\sqrt{N}} \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{B}^* & \tilde{C} \end{pmatrix}, \tag{27}$$

for some error blocks $\tilde{A}, \tilde{B}, \tilde{C}$ whose dependence on the A, B, C blocks of the LS estimator needs to be determined.

Our analysis draws on the asymptotic theory of the LS described in section 4.3, and the arguments developed in [65] for the analysis of the ML estimator. We know that $\hat{\rho}_{\text{PLS}}$ has the same eigenbasis as $\hat{\rho}_{\text{LS}}$, and its eigenvalues are obtained from those of $\hat{\rho}_{\text{LS}}$ by the truncation procedure which involves repeatedly setting to zero negative eigenvalues and shifting the remaining ones so that they add up to one. The following 3 step procedure aims to identify the leading contributions to PLS for low rank states, and $N \gg d \gg 1$.

1. *Diagonalisation.* Recall that for large N the LS estimator can be block diagonalised by means of a 'small' unitary rotation U, see equation (21)

$$U^* \hat{\rho}_{\text{LS}} U = \begin{pmatrix} I_r/r + A/\sqrt{N} & 0 \\ 0 & C/\sqrt{N} \end{pmatrix} + o(N^{-1/2}).$$

   Therefore the eigenvalues of $\hat{\rho}_{\text{LS}}$ can be grouped in two sets. The first is the set of eigenvalues of the block $C/\sqrt{N}$, which for large d are distributed according to the Wigner semicircle law and lie between $\pm 2\sqrt{(d-r)/N}$. The second set consists of the eigenvalues of the block $I_r/r + A/\sqrt{N}$, which are small fluctuations of order $r/\sqrt{N}$ around $1/r$.

2. *Truncation.* As long as $N \gg d \gg 1$ and r is small, the second set of eigenvalues is well separated from the first and it is very unlikely that any of these eigenvalues will be set to zero in the truncation process. Therefore, the cut-off point for the eigenvalues of $C/\sqrt{N}$ depends only on the sum of the larger eigenvalues

$$\text{Tr}(I_r/r + A/\sqrt{N}) = 1 + \text{Tr}(A)/\sqrt{N} =: 1 + a/\sqrt{N}.$$

Moreover, since $(d - r) \gg 1$, the eigenvalues of $C/\sqrt{N}$ are (approximately) distributed according to the Wigner distribution (20) on the interval $[-2\sqrt{(d-r)/N}, 2\sqrt{(d-r)/N}]$. Therefore, we can write the cut-off point q as the solution of the following equation [65]

$$1 = 1 + \frac{a}{\sqrt{N}} - rq + (d - r) \int_q^{2\sqrt{\frac{d-r}{N}}} (x - q) \frac{N}{2\pi(d-r)} \sqrt{\frac{4(d-r)}{N} - x^2}\, dx$$

where the integral corresponds to the sum of the eigenvalues of the block $C/\sqrt{N}$ above q, after being shifted by q. Writing $q = 2\sqrt{d - r/N}\epsilon$ with $\epsilon = \epsilon(r, d, N, a) < 1$, we get the following equation for ϵ

$$r\epsilon - \frac{a}{2\sqrt{d}} = \frac{2(d-r)}{\pi} \int_\epsilon^1 (y - \epsilon)\sqrt{1 - y^2}\,dy. \tag{28}$$

As the left side is smaller than r, the integral needs to be smaller than $r\pi/2(d - r)$, which means that ϵ is close to 1 for $r \ll d$. This agrees with the intuition that a large part of the eigenvalues of the lower block will be set to zero by projecting the LS onto states. Further details on finding an (approximate) solution to (28) can be found in [65]. In particular, we will approximate ϵ by the *deterministic* solution of equation (28) in which $a = \mathrm{Tr}(A)$ is set to zero; indeed, for large d this will have a negligible effect on ϵ but will allow us to compute ϵ it in terms of r and d deterministically.

3. *Rotation to original basis.* Once the cut-off point q has been computed, the projection of the rotated LS estimator $U^* \hat{\rho}_{\mathrm{LS}} U$ onto states can be written as

$$\begin{pmatrix} \alpha I_r + \frac{A}{\sqrt{N}} & 0 \\ 0 & \frac{C'}{\sqrt{N}} \end{pmatrix} + o\left(\frac{1}{\sqrt{N}}\right),$$

where

$$\alpha := \frac{1}{r} - 2\sqrt{\frac{d-r}{N}}\epsilon, \quad C' = \left[C - 2\sqrt{d-r}\epsilon I_{d-r}\right]_+$$

and *[X]*$_+$ denotes the positive part of *X*. The PLS estimator is now obtained by performing the inverse rotation

$$\begin{aligned} \hat{\rho}_{\mathrm{PLS}} &= U \begin{pmatrix} \alpha I_r + \frac{A}{\sqrt{N}} & 0 \\ 0 & \frac{C'}{\sqrt{N}} \end{pmatrix} U^* + o\left(\frac{1}{\sqrt{N}}\right) \\ &= \rho_r + \frac{1}{\sqrt{N}} \begin{pmatrix} A - 2\sqrt{d-r}\epsilon I_r & \alpha r B \\ \alpha r B^* & C' \end{pmatrix} + o\left(\frac{1}{\sqrt{N}}\right) \\ &= \rho_r + \frac{1}{\sqrt{N}} \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{B}^* & \tilde{C} \end{pmatrix}. \end{aligned} \tag{29}$$

We can now estimate the asymptotic risk of the PLS estimator with respect to the chosen loss functions.

Frobenius risk. The mean square error scales as N^{-1} and its rescaled version is

$$N\mathbb{E}\|\hat{\rho}_{\mathrm{PLS}} - \rho_r\|_2^2 = \mathbb{E}\|\tilde{A}\|_2^2 + 2\mathbb{E}\|\tilde{B}\|_2^2 + \mathbb{E}\|\tilde{C}\|_2^2.$$

The contribution from $\tilde{B}$ is

$$2\mathbb{E}\|\tilde{B}\|_2^2 = 2(\alpha r)^2 \mathbb{E}\|B\|_2^2 = \frac{2(d+1)(d-r)(r+1)}{d+2}\left(1 - 2r\sqrt{\frac{d-r}{N}}\epsilon\right)^2$$

where the variance of *B* has been computed as in section 4.3; note that the term $2r\sqrt{(d-r)/N}\epsilon$ vanishes for $N \gg d$. This error is of the order $2rd$ and can be seen as stemming from the uncertainty in estimating the eigenbasis of ρ_r. For $\tilde{A}$ we write

$$\mathbb{E}\|\tilde{A}\|_2^2 = \mathbb{E}\|A\|_2^2 + 4r(d-r)\epsilon^2 = (r+1)(r+2)\frac{d+1}{d+2} - \frac{(r+1)^2}{r} + 4r(d-r)\epsilon^2.$$

Finally, the term $\mathbb{E}\|\tilde{C}\|_2^2$ is given by the sum of the squares of the remaining eigenvalues which can be approximated using the limit Wigner law as

$$\mathbb{E}\|\tilde{C}\|_2^2 = N(d-r)\int_q^{2\sqrt{\frac{d-r}{N}}}(x-q)^2\frac{N}{2\pi(d-r)}\sqrt{\frac{4(d-r)}{N} - x^2}\,dx$$
$$= \frac{8(d-r)^2}{\pi}\int_\epsilon^1 (y-\epsilon)^2\sqrt{1-y^2}\,dy.$$

Although the last term appears to be of order $(d-r)^2$, a careful analysis of the integral [65] shows that it is of lower order than *rd* due to the fact that $1 - \epsilon$ leading term in a Taylor expansion is proportional to $(r/(d-r))^{2/5}$. Therefore, by adding the three terms, we find that the Frobenius risk scales as $6rd/N$. This agrees with the non-asymptotic results of [36], and provides the exact asymptotic constant of the Frobenius rate. By comparing with the lower bound to the asymptotic minimax rate of $2r(d-r)/N$ derived in [17] we find that PLS is optimal for such states, within a constant which is at most 3.

For the $n = 7$ atoms state with $r = 10$ the Frobenius error is 0.041, compared to the theoretical one 0.043. For rank $r = 1$ state of 8 atoms with $N = 10^5$ samples, the Frobenius error was 0.017, while the theoretical prediction is 0.016. The simulations results for all ranks of $n = 5$ and $n = 6$ atoms are presented in panels (a) and respectively (b) of figure 3.

Operator norm risk. Unlike the case of the LS estimator, the error matrix of PLS does not approach a Wigner distribution. For this reason, obtaining the asymptotic operator-norm risk seems difficult. However, the following lower bound follows from equation (29)

$$\sqrt{N}\|\hat{\rho}_{\text{PLS}} - \rho_r\| \geqslant \max(\|\tilde{A}\|, \|\tilde{C}\|).$$

Due to the truncation, the matrix $\tilde{C}$ is positive and the largest eigenvalue is approximtely $2\sqrt{d-r}(1-\epsilon)$. On the other hand, $\tilde{A}$ is dominated by the term $2\sqrt{d-r}\epsilon I_r$ whose norm is $2\sqrt{d-r}\epsilon$. Since $\epsilon \approx 1$ for large *d*, the lower bound is $2\sqrt{d-r}\epsilon$. This complements the non-asymptotic upper bound of [36] which has rate $O(\sqrt{d})$. Moreover, the lower bound seems to be a good approximation to the actual risk. A simulation with a rank $r = 10$ state of $n = 7$ atoms and $N = 10^6$ samples gave an operator-norm error of 0.02 while the lower bound is 0.01. For rank $r = 1$ state of 8 atoms with $N = 10^5$ samples, the operator-norm error was 0.12, while the lower bound is 0.08.

Trace-norm risk. As in the case of the norm-error, we could not derive the asymptotic expression of the trace-norm risk but we can formulate a lower bound based on the pinching inequality

$$\sqrt{N}\|\hat{\rho}_{\text{PLS}} - \rho_r\|_1 \geqslant \|\tilde{A}\|_1 + \|\tilde{C}\|_1.$$

Note that $\tilde{A} = A - 2\sqrt{d - r}\epsilon I_r$, and the variance of the elements of A is of the order 1. Therefore, for $r \ll d$ the shift $2\sqrt{d-r}$ dominates the eigenvalues of A and $\tilde{A}$ is a negative matrix. In first approximation its norm-one is then $2r\sqrt{d-r}\epsilon$. On the other hand, $\tilde{C}$ is positive and its trace can be approximated as

$$
\begin{aligned}
\mathbb{E}\|\tilde{C}\|_1 &= \sqrt{N}(d-r)\int_q^{2\sqrt{\frac{d-r}{N}}}(x-q)\frac{N}{2\pi(d-r)}\sqrt{\frac{4(d-r)}{N} - x^2}\,\mathrm{d}x \\
&= \frac{4(d-r)^{3/2}}{\pi}\int_\epsilon^1 (y-\epsilon)\sqrt{1-y^2}\,\mathrm{d}y \\
&= 2r\sqrt{d-r}\epsilon
\end{aligned}
$$

where in last step we used equation (28) with $a = 0$. Therefore the lower bound to the trace-norm error is $4r\sqrt{d-r}\epsilon$. For the $n = 7$ atoms state with $r = 10$ the trace-norm error was 0.33 while the lower bound is 0.21. For rank $r = 1$ state of 8 atoms with $N = 10^5$ samples, the operator-norm error was 0.24, while the lower bound is 0.16.

Bures risk. The Bures distance error can be expressed in terms of the blocks $\tilde{A}, \tilde{B}, \tilde{C}$ as follows

$$
\begin{aligned}
d_B(\hat{\rho}_{\text{PLS}}, \rho_r) &= 2\left(1 - \text{Tr}\left(\sqrt{\sqrt{\rho_r}\hat{\rho}_{\text{PLS}}\sqrt{\rho_r}}\right)\right) = 2\left(1 - \frac{1}{\sqrt{r}}\text{Tr}\left(\sqrt{I_r\hat{\rho}_{\text{PLS}}I_r}\right)\right) \\
&= 2\left(1 - \frac{1}{\sqrt{r}}\text{Tr}\left(\sqrt{I_r/r + \tilde{A}/\sqrt{N}}\right)\right) = 2\left(1 - \frac{1}{r}\text{Tr}\left(\sqrt{I_r + r\tilde{A}/\sqrt{N}}\right)\right) \quad (30) \\
&= -\frac{1}{\sqrt{N}}\text{Tr}(\tilde{A}) + \frac{r}{4N}\text{Tr}(\tilde{A}^2) + o(N^{-1})
\end{aligned}
$$

where we used the Taylor expansion in the last step. The leading term of the Bures risk is then

$$
-\frac{1}{\sqrt{N}}\mathbb{E}\text{Tr}(\tilde{A}) = \frac{2r\sqrt{d-r}}{\sqrt{N}}\epsilon(r, d)
$$

where we used the fact that the LS block A has centred distribution. The second order term is

$$
\begin{aligned}
\frac{r}{4N}\mathbb{E}\text{Tr}(\tilde{A}^2) &= \frac{r}{4N}(\mathbb{E}\|A\|^2 + 4r(d-r)\epsilon^2) \\
&= \frac{r}{4N}\left((r+1)(r+2)\frac{d+1}{d+2} - \frac{(r+1)^2}{r} + 4r(d-r)\epsilon^2\right).
\end{aligned}
$$

The simulation results for all ranks states of $n = 5$ and $n = 6$ atoms are presented panels (c) and respectively (d) of figure 3.

9. Comparative numerical simulations

In this section we detail the methodology and results of a general simulation study which compares the performance of the estimators presented in the previous sections for a set of states and against several estimation criteria. The states we consider are rank-r states with a fixed spectrum of r equal eigenvalues of magnitude $1/r$ each, and have randomly generated eigenvectors. This choice is motivated by the fact that such states are arguably harder to estimate among rank-r states (in analogy to the fact that an unbiased coin is harder to estimate than a biased one. Additionally, having such a spectrum allows for a more consistent comparison of the estimators across several ranks.
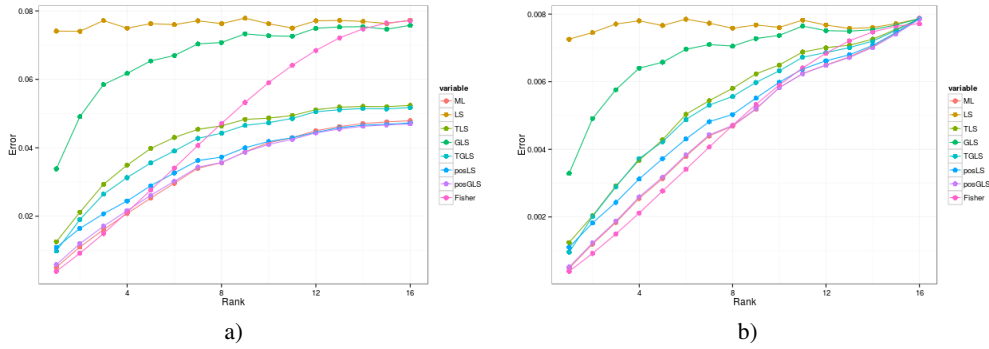
Figure 4. The mean squared Frobenius error $\mathbb{E}\left[\|\hat{\rho} - \rho\|_2^2\right]$ of the estimators for random 4 qubit rank-r states of equal eigenvalues, with Pauli measurements. (a) $n = 4$, $k = 81$ and $m = 100$. (b) $n = 4$, $k = 81$ and $m = 1000$.

We generate the above mentioned states for 3 or 4 qubits, and for a particular 'true state' we simulate a dataset $\mathcal{D}$ of counts from which the state is to be reconstructed. The outcome statistics depend on a few variables that we may vary, namely the type of measurement design (random basis versus Pauli), the number of repetitions per settings m, and in the case of the random basis measurements the total number k of measured settings. This allows us to study the performance of the estimators across several different combinations of variables: types of states, ranks, measurement design, number of repetitions per setting m, the total number of settings k and the number of qubits n.

We will present plots of the estimated mean errors $\mathbb{E}\left[D(\hat{\rho}, \rho)\right]$ of the estimators for the equal eigenvalues states. For each given rank r and number of qubits n, we generate a state with equal eigenvalues $\left(\frac{1}{r}, \ldots, \frac{1}{r}, 0, \ldots, 0\right)$ and random eigenbasis. Then for each choice of measurement design and values of k and m, the several estimates of the true state are evaluated. The error of each resulting estimate is computed using all the error functions listed in table 1, and the corresponding mean errors are estimated from 100 different runs of the experiment.

In order to make the results of the simulation study more accessible, we have made all plots for 3 and 4 atoms simulation available online via an interactive Rshiny application (https://rudhacharya.shinyapps.io/plots/), while a selection is presented in the paper.

9.1. Squared Frobenius norm

Figures 4 and 5 show the (estimated) Frobenius risk (mean square error) for states 4 qubits measured with the Pauli and the random basis design, respectively. The states are chosen randomly from the family of rank-r states with equal non-zero eigenvalues. In both cases we note that the Frobenius risk of the LS estimator has no significant dependence on the rank of the true state, and its performance is poor for small rank states. In contrast, the remaining estimators all show a scaling of the Frobenius risk with the rank of the true state. We also note that the performance of several of the estimators matches well with the Fisher-predicted risk. This is remarkable as the latter (6) was defined for a rank-r parameterisation of states, while none of the estimators have any prior knowledge of the rank.

We further note that for relatively small values of $N = m \times k$ the TLS, TGLS, posLS, posGLS, ML estimators significantly better that LS and GLS at higher ranks, while the errors approach each other for larger N, see figures 4(a) and 5(a) versus 4(b) and 5(b). This reflects
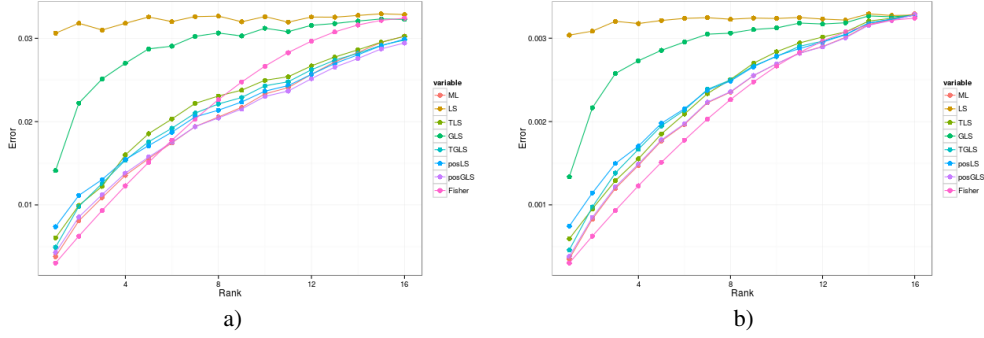
a)   b)

Figure 5. The mean squared Frobenius error $\mathbb{E}\left[\|\hat{\rho} - \rho\|_2^2\right]$ of the estimators for 4 qubits rank-*r* states of equal eigenvalues with $k = 100$ random bases and different repetition numbers. (a) $n = 4$, $k = 100$ and $m = 100$. (b) $n = 4$, $k = 100$ and $m = 1000$.

the fact that in the low N case the asymptotic regime has not been reached and constrained estimators have an advantage even for full rank states which lie in the interior of the parameter space. Indeed for an eigenvalue λ of order $1/d$ and relatively small values of N, the unconstrained estimates $\hat{\lambda}$ will have a standard deviation of order $\sqrt{d/N}$ which may be comparable or larger than the magnitude of the eigenvalues themselves. Therefore without the constraint of positivity such estimators may produce estimates with $\hat{\lambda} < 0$. In contrast, for the constrained estimators the positivity constraint provides additional information when the eigenvalues are small. For large values of N however, the uncertainty in the eigenvalues is very small and the Fisher risk acts as a lower bound for all of the estimators.

Across both the Pauli and the random measurement designs we note that the performance of the posGLS and the ML estimators is very similar, and for large m almost identical. This confirms our asymptotic analysis which shows that the posGLS and the ML estimators are equivalent in the limit of large m, see section 7.

9.2. Bures and Hellinger distance

As the Bures distance $D_B(\hat{\rho}, \rho)^2$ is well defined only over density matrices, we plot the mean Bures errors only for the ML, TLS, TGLS, posLS , posGLS estimators. Figures 6 and 7 show the mean Bures errors for different sample sizes in the Pauli, and respectively random bases measurement design. For comparison, in figures 8 and 9 we also plot the corresponding average Hellinger errors $D_H(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})^2$, see table 1.

We note that the behaviour of the Hellinger errors is very similar to that of the Bures errors, with a better match for larger values of N. To give some intuition about this, we look at what happens in the case of qubits. The Bures distance between neighbouring qubit states can be approximated by the sum of Hellinger distance and a quadratic form in the parameters of the unitary connecting the eigenbases of the two states [4]

$$D_B(\rho, \hat{\rho})^2 \approx D_H(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})^2 + \frac{1}{4}\frac{(1 - 2\lambda)(1 - 2\hat{\lambda})}{\sqrt{(1 - \lambda)(1 - \hat{\lambda})} + \sqrt{\lambda\hat{\lambda}}}\Phi^2,$$

(31)

where Φ is the angle between the Bloch vectors of ρ and $\hat{\rho}$. In a non-adaptive measurement scenario such as those considered here, the Bloch vector parameters can be estimated at rate $1/\sqrt{N}$ which means that the second term on the right side of (31) is always of the order $1/N$.
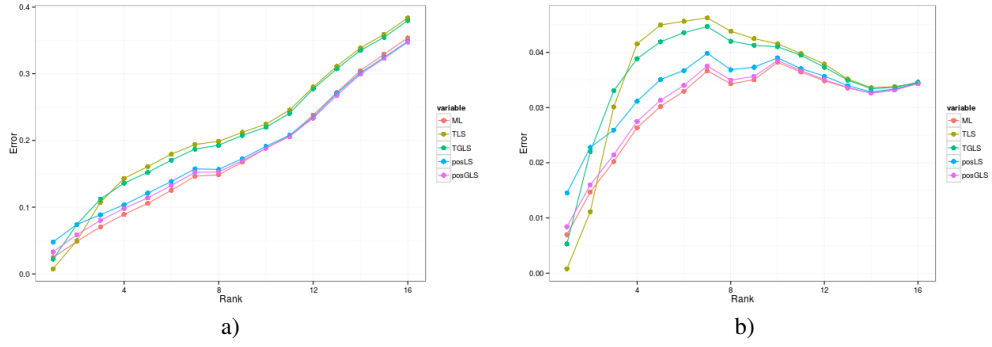
a)

b)

Figure 6. The mean Bures error $\mathbb{E}\left[D_B(\hat{\rho}, \rho)^2\right]$ of the estimators for random 4 qubits rank-*r* states of equal eigenvalues, with Pauli measurement design. (a) $n = 4$, $k = 81$ and $m = 100$. (b) $n = 4$, $k = 81$ and $m = 1000$.
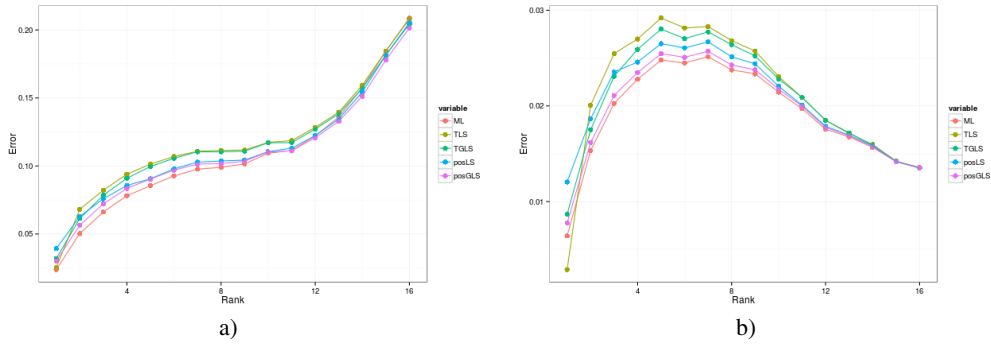

a)

b)

Figure 7. The mean Bures error $\mathbb{E}\left[D_B(\hat{\rho}, \rho)^2\right]$ of the estimators for random rank-*r* states of equal eigenvalues, with random bases measurement design. (a) $n = 4$, $k = 100$ and $m = 100$. (b) $n = 4$, $k = 100$ and $m = 1000$.
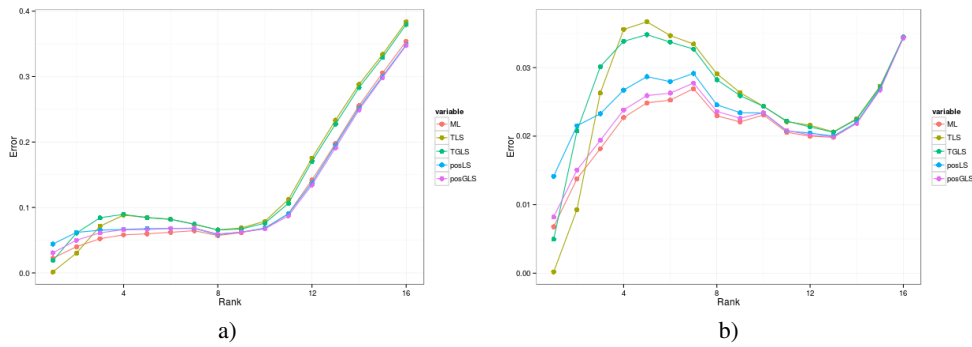

a)

b)

Figure 8. The mean Hellinger error $\mathbb{E}\left[D_H(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})^2\right]$ of the estimated eigenvalues $\hat{\boldsymbol{\lambda}}$ for random 4 qubits rank-*r* states of equal eigenvalues, for Pauli measurement design. (a) $n = 4$, $k = 81$ and $m = 100$. (b) $n = 4$, $k = 81$ and $m = 1000$.
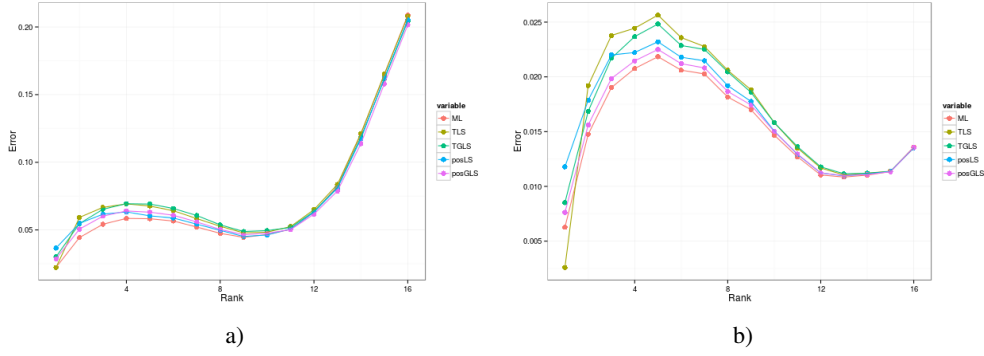
Figure 9. The mean Hellinger error $\mathbb{E}\left[D_{\mathrm{H}}(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})^2\right]$ of the estimated eigenvalues $\hat{\boldsymbol{\lambda}}$ for random 4 qubits rank-r states of equal eigenvalues, for random bases measurement design. (a) $n = 4$, $k = 100$ and $m = 100$. (b) $n = 4$, $k = 100$ and $m = 1000$.

However, for states which are very pure ($\lambda \approx 0$) the Hellinger component has the dominant contribution to the Bures distance, and is responsible for the 'non-standard' scaling of $1/\sqrt{N}$ in the minimax risk [4]. The simulation results indicate that a similar phenomenon may occur in higher dimensional systems. For full rank states, both the Bures and the Hellinger distance have a quadratic expansion, and in this sense a relation similar to (31) can be derived by splitting the Bures distance into quadratic contributions coming from changes in the eigenvalues and small basis rotations respectively, see also [60]. Alternatively, the block-matrix techniques used for analysing the Bures risk in section 8.1 can be extended to rank deficient states of arbitrary spectrum to show that the leading $1/\sqrt{N}$ contribution comes from the Hellinger $D_{\mathrm{H}}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})^2$.

Another noticeable feature across both the Pauli and the random bases measurement design is that for large N the mean errors are seen to be larger for states of middling ranks than for the full rank states, see figures 6(b) and 7(b). This is however not true for smaller values of N, as shown in figures 6(a) and 7(a). More precisely, for large N we see a steep increase from pure states to low rank states, followed by an almost linear decrease down to the full rank state. A similar behaviour has been uncovered in the analysis of the PLS estimator for covariant measurements in section 8.1. There, we found that for large d and low rank r, the eigenvalues distribution of the error block C of the LS estimator converges to the Wigner distribution, which allows us to compute the leading orders of the Bures risk of the PLS estimator. Since the covariant measurement arises in the large m limit of the random basis measurement [3], it is therefore expected that the risks behave similarly in the two cases. Our simulations indicate that the mechanism governing the asymptotics of the Bures risk seems to be robust with respect to the details of the measurement; although the Pauli basis measurement is not expected to produce an LS estimator whose error matrix is Wigner distributed, the Bures and Hellinger risks have similar characteristics as those of the covariant measurement. These findings are in line with those of [65], and are worthy of further theoretical investigation.

While the Bures risk for rank deficient states scales as $1/\sqrt{N}$, for full rank states the Bures distance is locally quadratic, and standard asymptotic results imply that the convergence rate is in this case $1/N$, see section 3.1. In general,

$$D_B(\rho_{\boldsymbol{\theta}}, \rho_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) = (\boldsymbol{\delta\theta})^T G_B(\boldsymbol{\theta})(\boldsymbol{\delta\theta}) + O(\|\boldsymbol{\delta\theta}\|^2),$$

a)   b)
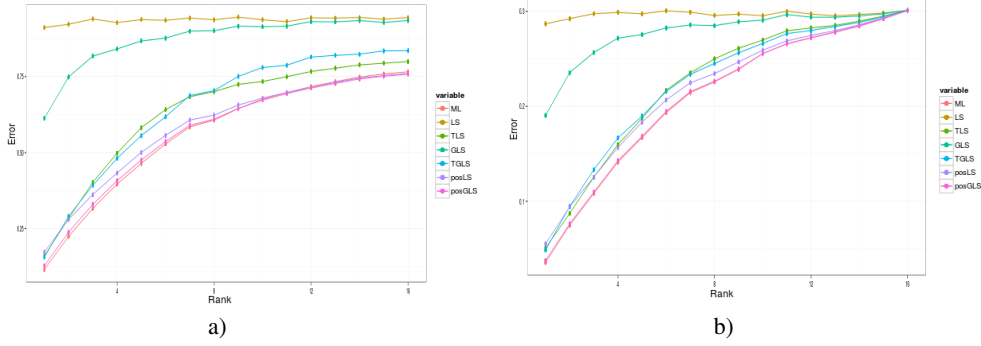
Figure 10. The mean trace-norm error $\mathbb{E}\left[\|\hat{\rho} - \rho\|_1\right]$ for Pauli bases measurements. Panels (a) and (b) show the risks for 4 qubits rank-*r* states of equal eigenvalues with different repetition numbers. (a) $n = 4$, $m = 100$. (b) $n = 4$, $m = 1000$.
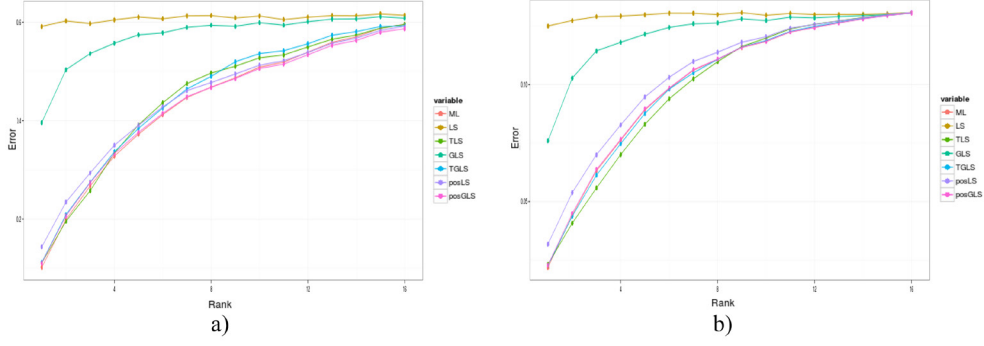

a)   b)

Figure 11. The mean trace-norm error $\mathbb{E}\left[\|\hat{\rho} - \rho\|_1\right]$ for random bases measurements. Panels a) and b) show the risks for 4 qubits rank-*r* states of equal eigenvalues with different number of bases and repetition numbers. (a) $n = 4$, $k = 100$ and $m = 100$. (b) $n = 4$, $k = 200$ and $m = 1000$.

with weight matrix $G_B(\boldsymbol{\theta}) = F(\rho_{\boldsymbol{\theta}})/4$, where $F(\rho_{\boldsymbol{\theta}})$ is the quantum Fisher information. For the maximally mixed state $\rho_{r=d}$, and the parametrisation (4), the latter has two uncorrelated blocks [3] corresponding to off-diagonal parameters

$$F_{a,b} = 2d\delta_{a,b}, \qquad 1 \leqslant a, b \leqslant d^2 - d$$

and respectively diagonal parameters

$$F_{a,b} = d(1 + \delta_{a,b}), \qquad d^2 - d \leqslant a, b \leqslant d^2 - 1.$$

On the other hand, the classical average Fisher information for random basis measurements is given by $I = F/(d + 1)$ [17]. Since the ML estimator is asymptotically normal with variance $I(\boldsymbol{\theta})^{-1}/N$, we have

$$N\mathbb{E}\left[D_B(\hat{\rho}_{\text{ML}}, \rho)^2\right] \to \text{Tr}(G_B I^{-1}) = \frac{(d^2 - 1)(d + 1)}{4}.$$

For 3 qubits with $N = 100 \cdot 1000$ samples, and 4 qubits with $N = 100 \cdot 1000$, the asymptotic predictions for Bures errors are 0.0014 and 0.01 which match closely the simulations results [1].

9.3. Trace-norm distance

The risks for the trace-norm distance exhibit a similar behaviour to those of the Frobenius distance, as illustrated in figures 10 and 11. A noticeable feature is that all constrained estimators have close risks for large sample sizes in both the Pauli and the random bases setting.

10. Conclusions

In this paper we studied theoretical an practical aspects of quantum tomography methods. The unifying theme is that each estimator can be seen as a projection of the data onto a parameter space with respect to an appropriate metric. We considered estimators without positivity constraints (unconstrained maximum likelihood, least squares, generalised least squares) and with positivity constraints (maximum likelihood, positive least squares, thresholded least squares and projected least squares), and investigated the relationships between different estimators. While no estimator makes use of the state's rank (which is assumed to be unknown) the constrained estimators have significantly lower errors than the unconstrained estimators, for low rank states. To better understand this behaviour we derived new asymptotic error rates for the least squares estimator and for the projected least squares estimators, for a class of given rank states and covariant measurements. These results capture the exact rate dependence on rank and dimension and complement non-asymptotic concentration bounds of [17, 36, 71], showing that PLS has strong optimality properties; for instance the leading contribution to the Frobenius risk is 6rd which is 'almost optimal', in that it is only 3 times larger than the minimax lower bound established in [17], which assumes that the rank is known. Our analysis has uncovered an interesting behaviour of the Bures error, which increases with respect to rank for small ranks and then decreases towards full rank states. The extensive simulations study shows that this behaviour (as well as the monotonic increase of other errors) is robust with respect to the measurement design.

Computationally, maximum likelihood and positive least squares involve an optimisation over states and are significantly slower than projected least squares which only requires the diagonalisation of the least squares matrix followed by a simple truncation procedure. Our results confirm and strengthen those of [36] and show that projected least squares is an attractive alternative to maximum likelihood, which is routinely used in practice. Additional improvements can be achieved by using generalised least squares as starting point, in a two steps procedure.

An interesting and practically relevant open question is whether any of the 'fast' estimators analysed here is statistically 'optimal' for more realistic measurements such as the Pauli bases. More generally, one can ask whether these methods can be adapted to non-informationally complete measurement scenarios, and other physically motivated lower dimensional statistical models.

Acknowledgments

Appendix

A.1. Fisher information for GLS

Here we prove the equality $I(\rho|\mathscr{S}) = I_{\text{GLS}}$ stated in section 5.1. Note that by definition

$$I(\rho|\mathscr{S}) = \frac{1}{k}\tilde{X}^*\text{Diag}(p_\rho^{-1})\tilde{X} = \frac{1}{k}J^*X^*\text{Diag}(p_\rho^{-1})XJ$$

and

$$I_{\text{GLS}} = \frac{1}{k}\tilde{X}^T\tilde{\Omega}^{-1}\tilde{X} = \frac{1}{k}J^*X^*V(V^*\Omega V)^{-1}V^*XJ. \tag{A.1}$$

We will show that

$$V(V^*\Omega V)^{-1}V^* = \text{Diag}(p_\rho^{-1}) + c|\mathbf{1}\rangle\langle\mathbf{1}| + d(|\mathbf{1}\rangle\langle p_\rho^{-1}| + |p_\rho^{-1}\rangle\langle\mathbf{1}|) \tag{A.2}$$

where c, d are some constants and $|p_\rho^{-1}\rangle \in \mathbb{C}^{k \cdot d}$ is the vector whose entries are the inverses of measurement probabilities. Indeed, since $J^*X^*|\mathbf{1}\rangle = 0$, equation (A.2) implies (A.1). Now $V(V^*\Omega V)^{-1}V^*$ is the pseudo-inverse of Ω and since $V^*|\mathbf{1}\rangle = 0$, it satisfies

$$V(V^*\Omega V)^{-1}V^* = (\Omega - |\mathbf{1}\rangle\langle\mathbf{1}|)^{-1} + |\mathbf{1}\rangle\langle\mathbf{1}|.$$

To compute the inverse on the right side we use the definition (17) of Ω and apply the Sherman–Morrison formula [67]

$$(A - |x\rangle\langle x|)^{-1} = A^{-1} + \frac{A^{-1}|x\rangle\langle x|A^{-1}}{1 - \langle x|A^{-1}|x\rangle}.$$

This gives

$$(\Omega - |\mathbf{1}\rangle\langle\mathbf{1}|)^{-1} = (\text{Diag}(p_\rho) - |p_\rho\rangle\langle p_\rho| - |\mathbf{1}\rangle\langle\mathbf{1}|)^{-1} = B^{-1} + \frac{B^{-1}|p_\rho\rangle\langle p_\rho|B^{-1}}{1 - \langle p_\rho|B^{-1}|p_\rho\rangle} \tag{A.3}$$

where $B = \text{Diag}(p_\rho) - |\mathbf{1}\rangle\langle\mathbf{1}|$. By applying the Sherman–Morrison formula again we get

$$B^{-1} = \text{Diag}(p_\rho^{-1}) + \frac{|p_\rho^{-1}\rangle\langle p_\rho^{-1}|}{1 - \langle\mathbf{1}|\text{Diag}(p_\rho^{-1})|\mathbf{1}\rangle} = \text{Diag}(p_\rho^{-1}) + \frac{|p_\rho^{-1}\rangle\langle p_\rho^{-1}|}{1 - q}$$

where $q = \sum_{\mathbf{o},\mathbf{s}} p_\rho^{-1}(\mathbf{o}|\mathbf{s})$. Moreover,

$$B^{-1}|p_\rho\rangle = |\mathbf{1}\rangle + \frac{kd}{1 - q}|p_\rho^{-1}\rangle, \qquad \langle p_\rho|B^{-1}|p_\rho\rangle = 1 + \frac{(kd)^2}{1 - q}.$$

By plugging the last three expresssions into (A.3) we get

$$(\Omega - |\mathbf{1}\rangle\langle\mathbf{1}|)^{-1} = \text{Diag}(p_\rho^{-1}) + \frac{|p_\rho^{-1}\rangle\langle p_\rho^{-1}|}{1 - q}$$

$$- \frac{1 - q}{(kd)^2}\left(|\mathbf{1}\rangle\langle\mathbf{1}| + \frac{kd}{1 - q}(|\mathbf{1}\rangle\langle p_\rho^{-1}| + |p_\rho^{-1}\rangle\langle\mathbf{1}|) + \frac{(kd)^2}{(1 - q)^2}|p_\rho^{-1}\rangle\langle p_\rho^{-1}|\right)$$

$$= \text{Diag}(p_\rho^{-1}) + c|\mathbf{1}\rangle\langle\mathbf{1}| + d|\mathbf{1}\rangle\langle p_\rho^{-1}| + \bar{d}|p_\rho^{-1}\rangle\langle\mathbf{1}|$$

which concludes the proof of (A.2) and of $I(\rho|\mathscr{S}) = I_{\text{GLS}}$.

A.2. Implementations of estimators

Here we list certain practical details about the implementation of the estimators described in the paper. In particular we discuss how we compute the covariance matrix estimator $\hat{\hat{\Omega}}$ used in the GLS, TGLS, posGLS estimators, and we also describe the cross-validation procedure we use to select a constant C for the thresholded in the TLS and TGLS estimators.

1. The covariance matrix estimator $\hat{\hat{\Omega}}$ for the generalised estimators (GLS, TGLS and posGLS) is computed as follows. Given a dataset $\mathcal{D}$, we first obtain the LS estimate $\hat{\rho}_{\mathrm{LS}}$ and then construct the TLS estimate (see algorithm 1) with threshold $\nu = 0$. From this we obtain an estimate of the probabilities $\hat{p}(\boldsymbol{o}|\boldsymbol{s}) = \mathrm{Tr}\,[\hat{\rho}_{\mathrm{TLS}} P^{\boldsymbol{s}}_{\boldsymbol{o}}]$. The matrix $\hat{\hat{\Omega}}$ is then constructed from these estimated probabilities via (17). The generalised estimates (GLS/TGLS/posGLS) are then evaluated using $\hat{\hat{\Omega}}$ and the same dataset $\mathcal{D}$.

2. As mentioned briefly in section 6, the threshold for the TLS and TGLS estimators is selected using cross-validation. We describe this cross-validation method below [17].

   - For a particular number of repetitions per setting m, we simulate data in 5 independent batches $m/5$ repetitions per setting in each batch and we denote the corresponding datasets as $\mathcal{D}_1, \ldots, \mathcal{D}_5$. The total dataset of counts is the sum $\mathcal{D} = \sum_{i=1}^{5} \mathcal{D}_i$.
   - We choose a vector of constants C forming a mesh over the interval $[0, 1]$. For each value of C, and for each $j \in \{1, \ldots, 5\}$ we compute the following estimators. We hold out the dataset $\mathcal{D}_j$, and compute the TLS/TGLS estimate $\hat{\rho}^{-j}_{\mathrm{T(G)LS}}(C)$ for the dataset $\mathcal{D}_{-j} = \sum_{i \neq j} \mathcal{D}_i$, with threshold $\nu = C\sqrt{\frac{4}{m} \log 2^{n+1}}$, cf. [17]. For each $\mathcal{D}_j$ the LS estimate $\hat{\rho}^{j}_{\mathrm{LS}}$ is also evaluated.
   - For all values of C, the empirical discrepancy is evaluated for a choice of error function $D(\hat{\rho}, \rho)$

$$CV_D(C) = \frac{1}{5} \sum_{i=1}^{5} D\left(\hat{\rho}^{-j}_{\mathrm{T(G)LS}}(C), \hat{\rho}^{j}_{\mathrm{LS}} \right).$$

   - This function $CV(C)$ is then minimised over all values C

$$\hat{C}_D = \arg\min_{C} CV_D(C)$$

   this gives an estimate for the holding constant, which is then used to evaluate the TLS or the TGLS estimators with threshold $\nu = \hat{C}_D \sqrt{\frac{4}{m} \log 2^{n+1}}$.

Notice that the cross-validation procedure picks different constants for different choices of the error function. An important caveat here is that the Bures distance is not defined for the LS estimates $\hat{\rho}^{j}_{\mathrm{LS}}$, and therefore the procedure above cannot apply. Instead in the simulations we estimate the thresholding constant $\hat{C}_{D_B}$ using the ML estimate as

$$\hat{C}_{D_B} = \arg\min_{C} D_B(\hat{\rho}_{\mathrm{T(G)LS}}, \hat{\rho}_{\mathrm{ML}})^2.$$

ORCID iDs

Mădălin Guţă ● https://orcid.org/0000-0002-9017-2222

References

[1] Complete simulation results (https://rudhacharya.shinyapps.io/plots/)
[2] Cvx 2017 Matlab software for disciplined convex programming (http://cvxr.com/cvx/)
[3] Acharya A and Guta M 2017 Statistical analysis of compressive low rank tomography with random measurements *J. Phys. A: Math. Theor.* **50** 195301
[4] Acharya A and Guta M 2018 Minimax estimation of qubit states with Bures risk *J. Phys. A: Math. Theor.* **51** 175307
[5] Acharya A, Kypraios T and Guta M 2016 Statistically efficient tomography of low rank states with incomplete measurements *New J. Phys.* **18** 043018
[6] Artiles L, Gill R D and Guta M 2005 An invitation to quantum tomography *J. R. Stat. Soc.* B **67** 109–34
[7] Audenaert K M R and Scheel S 2009 Quantum tomographic reconstruction with error bars: a kalman filter approach *New J. Phys.* **11** 023028
[8] Bagan E, Baig M and Munoz-Tapia R 2002 Optimal scheme for estimating a pure qubit state via local measurements *Phys. Rev. Lett.* **89** 277904
[9] Bagan E, Ballester M A, Gill R D, Monras A and Munoz-Tapia R 2006 Optimal full estimation of qubit mixed states *Phys. Rev.* A **73** 032301
[10] Banaszek K, D'Ariano G M, Paris M G A and Sacchi M F 1999 Maximum-likelihood estimation of the density matrix *Phys. Rev.* A **61** 010304
[11] Barndorff-Nielsen O E, Gill D R and Jupp E P 2003 On quantum statistical inference (with discussion) *J. R. Stat. Soc.* B **65** 775–816
[12] Baumgratz T, Nüßeler A, Cramer M and Plenio M B 2013 A scalable maximum likelihood method for quantum state tomography *New J. Phys.* **15** 125004
[13] Belavkin V P 1976 Generalized uncertainty relations and efficient measurements in quantum systems *Theor. Math. Phys.* **26** 213–22
[14] Blume-Kohout R 2010 Hedged maximum likelihood quantum state estimation *Phys. Rev. Lett.* **105** 200504
[15] Blume-Kohout R 2010 Optimal, reliable estimation of quantum states *New J. Phys.* **12** 043034
[16] Braunstein S L and Caves C M 1994 Statistical distance and the geometry of quantum states *Phys. Rev. Lett.* **72** 3439–43
[17] Butucea C, Guta M and Kypraios T 2015 Spectral thresholding quantum tomography for low rank states *New J. Phys.* **17** 113050
[18] Christandl M and Renner R 2012 Reliable quantum state tomography *Phys. Rev. Lett.* **109** 120403
[19] Collins B and Śniady P 2006 Integration with respect to the haar measure on unitary, orthogonal and symplectic group *Commun. Math. Phys.* **264** 773–95
[20] Cramer M, Plenio M B, Flammia S T, Somma R, Gross D, Bartlett S D, Landon-Cardinal O, Poulin D and Liu Y-K 2010 Efficient quantum state tomography *Nat. Commun.* **1** 149
[21] D'Ariano G M, Macchiavello C and Paris G A M 1994 Detection of the density matrix through optical homodyne tomography without filtered back projection *Phys. Rev.* A **50** 4298–302
[22] Dryden I L, Le H, Preston S P and Wood A T A 2014 Mean shapes, projections and intrinsic limiting distributions *J. Stat. Plan. Inference* **145** 25–32
[23] Erdös L 2011 Universality of wigner random matrices: a survey of recent results *Russ. Math. Surv.* **66** 507
[24] Faist P and Renner R 2016 Practical and reliable error bars in quantum tomography *Phys. Rev. Lett.* **117** 010404
[25] Ferrie C and Blume-Kohout R 2016 Minimax quantum tomography: estimators and relative entropy bounds *Phys. Rev. Lett.* **116** 090407
[26] Ferrie C 2014 Self-guided quantum tomography *Phys. Rev. Lett.* **113** 190404
[27] Flammia S T, Gross D, Liu Y-K and Eisert J 2012 Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators *New J. Phys.* **14** 095022
[28] Fujiwara A and Nagaoka H 1995 Quantum fisher metric and estimation for pure state models *Phys. Lett.* A **201** 119–24
[29] Gao W-B, Lu C-Y, Yao X-C, Xu P, Guhne O, Goebel A, Chen Y-A, Peng C-Z, Chen Z-B and Pan J-W 2010 Experimental demonstration of a hyper-entangled ten-qubit schrodinger cat state *Nat. Phys.* **6** 331–5

J. Phys. A: Math. Theor. **52** (2019) 234001

A Acharya *et al*

[30] Gill R D and Guta M 2013 On asymptotic quantum statistical inference From Probability to Statistics and Back: High-Dimensional Models and Processes (A Festschrift in Honor of Jon A Wellner) (Beachwood, OH: Institute of Mathematical Statistics) pp 105–27

[31] Gill R D and Massar S 2000 State estimation for large ensembles *Phys. Rev.* A **61** 042312

[32] Goncalves D S, Gomes-Ruggiero M A, Lavor C, Farias O J and Ribeiro P H S 2012 Local solutions of maximum likelihood estimation in quantum state tomography *Quant. Info. Comput.* **12** 775–90

[33] Granade C, Combes J and Corrie D G 2016 Practical bayesian tomography *New J. Phys.* **18** 033024

[34] Granade C, Ferrie C and Flammia S T 2017 Practical adaptive quantum tomography *New J. Phys.* **19** 113017

[35] Gross D, Liu Y-K, Flammia S T, Becker S and Eisert J 2010 Quantum state tomography via compressed sensing *Phys. Rev. Lett.* **105** 150401

[36] Guta M, Kahn J, Kueng R and Tropp J A 2018 Fast state tomography with optimal error bounds (arXiv:1809.11162)

[37] Haah J, Harrow A W, Ji Z, Wu X and Yu N 2017 Sample-optimal tomography of quantum states *IEEE Trans. Inf. Theory* **63** 5628–41

[38] Haffner H *et al* 2005 Scalable multiparticle entanglement of trapped ions *Nature* **438** 643–6

[39] Hannemann Th, Reiss D, Balzer Ch, Neuhauser W, Toschek P E and Wunderlich Ch 2002 Self-learning estimation of quantum states *Phys. Rev.* A **65** 050303

[40] Hayashi M 2005 *Asymptotic Theory of Quantum Statistical Inference: Selected Papers* (Singapore: World Scientific)

[41] Hayashi M and Matsumoto K 2008 Asymptotic performance of optimal state estimation in qubit system *J. Math. Phys.* **49** 102101–33

[42] Helstrom C W 1969 Quantum detection and estimation theory *J. Stat. Phys.* **1** 231–52

[43] Holevo A S 2011 *Probabilistic and Statistical Aspects of Quantum Theory* vol 1, 1st edn (Pisa: Edizioni della Normale) (https://doi.org/10.1007/978-88-7642-378-9)

[44] James D F V, Kwiat P G, Munro W J and White A G 2001 Measurement of qubits *Phys. Rev.* A **64** 052312

[45] Kahn J and Guta M 2009 Local asymptotic normality for finite dimensional quantum systems *Commun. Math. Phys.* **289** 597–652

[46] Baldwin C H, Deutsch I H and Kalev A 2016 Strictly-complete measurements for bounded-rank quantumstate tomography *Phys. Rev.* A **93** 052105

[47] Keyl M and Werner R F 2001 Estimating the spectrum of a density operator *Phys. Rev.* A **64** 052311

[48] Koltchinskii V and Xia D 2015 Optimal estimation of low rank density matrices *J. Mach. Learn. Res.* **16** 1757–92

[49] Kueng R, Rauhut H and Terstiege U 2017 Low rank matrix recovery from rank one measurements *Appl. Comput. Harmon. Anal.* **42** 88–116

[50] Lanyon B P *et al* 2017 Efficient tomography of a quantum many-body system *Nat. Phys.* **13** 1158–62

[51] Lehmann E L and Casella G 1998 *Theory of Point Estimation* (New York: Springer)

[52] Leonhardt U 2010 *Essential Quantum Optics From Quantum Measurements to Black Holes* 1st edn (Cambridge: Cambridge University Press)

[53] Li X, Shang J, Ng H K and Englert B-G 2016 Optimal error intervals for properties of the quantum state *Phys. Rev.* A **94** 062112

[54] Liu Y-K 2011 Universal low-rank matrix recovery from pauli measurements *Advances in Neural Information Processing Systems 24: 25th Annual Conf. on Neural Information Processing Systems 2011. Proc. of a meeting held (Granada, Spain, 12–14 December 2011)* 1638–46

[55] Mahler D H, Rozema L A, Darabi A, Ferrie C, Blume-Kohout R and Steinberg A M 2013 Adaptive quantum state tomography improves accuracy quadratically *Phys. Rev. Lett.* **111** 183601

[56] Massar S and Popescu S 1995 Optimal extraction of information from finite quantum ensembles *Phys. Rev. Lett.* **74** 1259–63

[57] Matsumoto K 2002 A new approach to the cramer-rao type bound of the pure state model *J. Phys. A: Math. Gen.* **35** 3111–23

[58] Monz T, Schindler P, Barreiro J T, Chwalla M, Nigg D, Coish W A, Harlander M, Hänsel W, Hennrich M and Blatt R 2011 14-qubit entanglement: creation and coherence *Phys. Rev. Lett.* **106** 130506

[59] Paris M G A and Rehacek J (ed) 2004 *Quantum State Estimation* (*Lecture Notes in Physics* vol 649) (Berlin: Springer)

[60] Pereira L, Zambrano L, Cortés-Vega J, Niklitschek S and Delgado A 2018 Adaptive quantum tomography in high dimensions *Phys. Rev.* A **98** 012339

[61] Petz D and Jencova A 2006 Sufficiency in quantum statistical inference *Commun. Math. Phys.* **263** 259–76

[62] Qi B, Hou Z, Li L, Dong D, Xiang G and Guo G 2013 Quantum state tomography via linear regression estimation *Sci. Rep.* **3** 3496

[63] Rehacek J, Hradil Z, Knill E and Lvovsky A I 2007 Diluted maximum-likelihood algorithm for quantum tomography *Phys. Rev.* A **75** 042108

[64] Schenker J H and Schulz-Baldes H 2005 Semicircle law and freeness for random matrices with symmetries or correlations *Math. Res. Lett.* **12** 531–42

[65] Scholten T L and Blume-Kohout R 2018 Behavior of the maximum likelihood in quantum state tomography *New J. Phys.* **023050** 20

[66] Schwemmer C, Tóth G, Niggebaum A, Moroder T, Gross D, Gühne O and Weinfurter H 2014 Experimental comparison of efficient tomography schemes for a six-qubit state *Phys. Rev. Lett.* **113** 040503

[67] Sherman J and Morrison W J 1950 Adjustment of an inverse matrix corresponding to a change in one element of a given matrix *Ann. Math. Stat.* **21** 124–7

[68] Smolin J A, Gambetta J M and Smith G 2012 Efficient method for computing the maximum-likelihood quantum state from measurements with additive gaussian noise *Phys. Rev. Lett.* **108** 070502

[69] Steffens A, Riofrío C A, McCutcheon W, Roth I, Bell B A, McMillan A, Tame M S, Rarity J G and Eisert J 2017 Experimentally exploring compressed sensing quantum tomography *Quantum Sci. Technol.* **2** 025005

[70] Suess D, Rudnicki L, Maciel T O and Gross D 2017 Error regions in quantum state tomography: computational complexity caused by geometry of quantum states *New J. Phys.* **19** 093013

[71] Sugiyama T, Turner P S and Murao M 2013 Precision-guaranteed quantum tomography *Phys. Rev. Lett.* **111** 160406

[72] Moroder T, Hyllus P, Tóth G, Schwemmer C, Niggebaum A, Gaile S, Gühne O and Weinfurter H 2012 Permutationally invariant state reconstruction *New J. Phys.* **14** 105001

[73] Teo Y S, Englert B-G, Řeháček J, Hradil Z and Mogilevtsev D 2012 Verification of state and entanglement with incomplete tomography *New J. Phys.* **14** 105020

[74] Teo Y S, Zhu H, Englert B-G, Řeháček J and Hradil Z 2011 Quantum-state reconstruction by maximizing likelihood and entropy *Phys. Rev. Lett.* **107** 020404

[75] Tropp J A 2012 User-friendly tail bounds for sums of random matrices *Found. Comput. Math.* **12** 389–434

[76] Vidal G, Latorre J I, Pascual P and Tarrach R 1999 Optimal minimal measurements of mixed states *Phys. Rev.* A **60** 126

[77] Young G A and Smith R L 2005 *Essentials of Statistical Inference* (Cambridge: Cambridge University Press)

[78] Yuen H P and Lax M 1973 Multiple-parameter quantum estimation and measurement of non-selfadjoint observables *IEEE Trans. Inf. Theory* **19** 740