Smoothing parameter selection using the L-curve

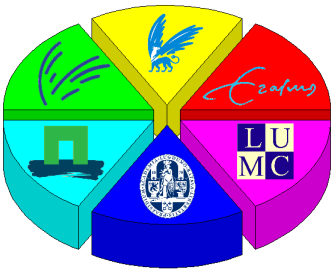
Master Thesis

Leiden University
Mathematics
Specialization Statistical Science

Defended on March 1, 2012

Gianluca Frasso

Thesis Advisor:
Prof. Dr. Paul H. C. Eilers



Contents

1 Introduction

Penalized regression has a prominent place in modern smoothing. It combines a rich set of basis functions with a roughness penalty, to tune smoothness of the estimated curve. B-splines are a popular choice for the basis functions, but others prefer truncated power functions. The penalty can be derived from classical roughness measures, like the integrated squared second derivative, or it can be discrete, working directly on the regression coefficients. An extensive discussion is presented by Eilers and Marx [6].

P-splines (Eilers and Marx [5]) combine a B-spline matrix with a penalty on (higher order) differences of their coefficients. If we have data on equally spaced positions and go to the limit, we will have a basis function for each observation and the regression basis will be the identity matrix. This brings us back to the Whittaker's smoother [20] and also [4], which recently became popular in econometric literature as the Hodrick-Prescott filter [11]. It is an attractive smoother, because effectively the basis functions disappear and with just one smoothing parameter one can move all the way from a straight line fit to essentially reproducing the data themselves.

It is desirable to have an automatic procedure for selecting a value for the smoothing parameter. In principle many choices are available. The most straightforward ones are leave-one-out cross-validation (LOO-CV) and AIC (Akaike's Information Criterion) or BIC (Bayesian Information Criterion). It is also possible to exploit the similarity between penalized regression and mixed models and then the smoothing parameter becomes a ratio of variances.

The established method for selection of the smoothing parameter have two things in common: 1) they require the computation of the effective model dimension, and 2) they are sensitive to serial correlation in the noise around the trend. The effective dimension is equal to the trace of the smoother matrix, and so inversion of a large matrix is required; for long data series this is prohibitive. Serial correlation generally leads to under-smoothing. At first sight this is surprising, but it is not hard to see why it happens. Indeed cross validation methods assume data with independent noise. If, otherwise, the error component shows serial correlation, cross validation tends to suggest small smoothing parameter, in order to obtain uncorrelated residuals.

In this paper we present and alternative approach, based on the L-curve method for ridge regression [9] and [7]. The curve is a plot of the logarithm of the magnitude of the penalty against the log of the sums of squares of the residuals, parameterized by the regularization parameter λ. In the case of ridge regression a very pronounced L-shape is obtained and a good value of the regularization parameter is found in its corner (see for example figure 2). As far as we know, the L-curve has not been used for smoothing, but it turns out to be very valuable there. The shape is less pronounced that of an L, but a corner is present and it can easily be located numerically by following the path parameterized by λ.

There is no need to compute the effective model dimension, so using the L-curve makes smoothing of long data series practical. And, very surprisingly, it is not affected by correlated noise. This is illustrated in figure 1, showing historical data of the price of orange juice [17], with very strong serial correlation. For the upper panel the amount of smoothing was chosen automatically by LOO-CV, while in the lower panel the L-curve was used. It is quite clear that no meaningful trend can be obtained with LOO-CV; AIC and the mixed model approach lead to very similar results (not shown). On the other hand the trend in the lower panel of the figure appears to summarize the data well.

We have no compelling explanations of why the L-curve works so well. Of course, the

corner of an L-shaped curve is a special point, but it is not clear why it marks a good choice of smoothing parameter. The relative changes of both the penalty and the size of the residuals are small there, and approximately equal, and apparently that matters. The insensitivity to serial correlation in the noise is also hard to explain. On the other hand this method shows excellent performances in practice.

This work is organized as follows: in Section 2 we introduce the smoothing procedures that will be used in our discussion, together with some standard smoothing selection procedures, Section 3 describes in more details the L-curve method, and finally, Section 4 compares the L-curve with a LOO-CV approach using real data.

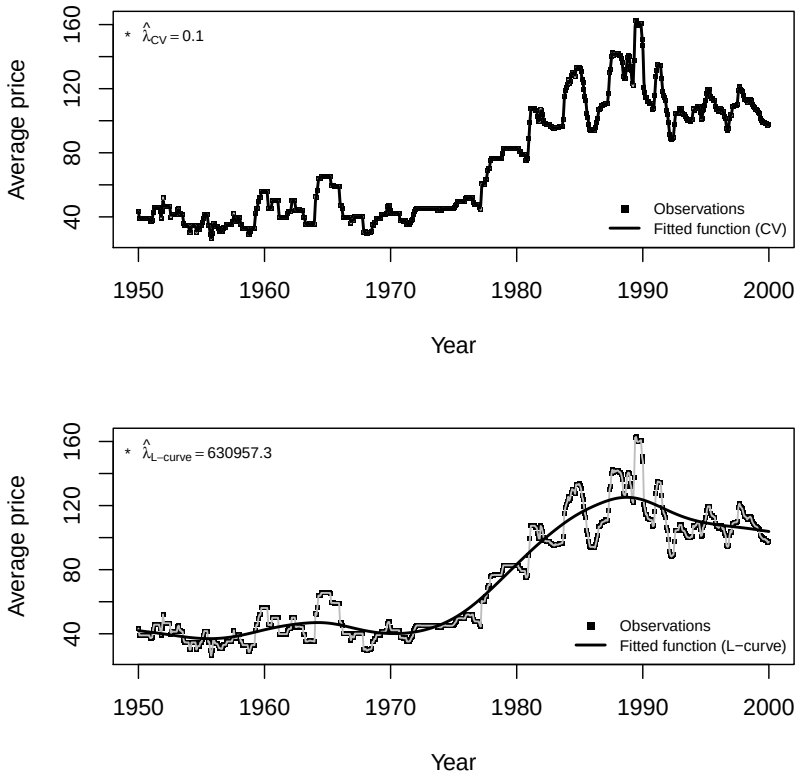


Figure 1: Hodrick-Prescott filter applied to the orange juice dataset (see Section 4). The smoothing parameter have been selected by cross validation (upper panel) and by L-curve procedure (lower panel). In both cases the smoothing parameter was considered in the range $\log(\lambda) \in [-4, 6]$.

2 P-splines, Whittaker smoother and Hodrick-Prescott filter

P-splines were proposed by Eilers and Marx [5]. Suppose that a set of data $\{x, y\}$, where x represents the independent (explanatory) variable and y the dependent variable, has been observed. We want to find a smooth function that describes y appropriately. Let $B_j(x; q)$ denote the value of the jth B-spline of degree q defined on a set of equidistant knots (taking

equidistant knots does not affect our further results but, in our opinion, it is generally a good idea, see [6]). A curve that describes the data $\{x, y\}$ is given by $\widehat{y}(x) = \sum_{j=1}^{n} \widehat{z}_j B_j(x; q)$ where $\widehat{z}_j$ (with $j = 1, ..., n$) are the estimated B-splines coefficients. Unfortunately this curve, that is obtained minimizing $\|y - Bz\|^2$ w.r.t. z, usually shows more variation than is justified by the data. To avoid this over-fitting effect we can introduce a roughness penalty so that the estimation problem can be reformulated as follows:

$$\widehat{z} = \underset{z}{\operatorname{argmin}} \|y - Bz\|^2 + \lambda \|Dz\|^2 \tag{1}$$

where D is a mth order difference penalty matrix and λ is the smoothing parameter that controls the trade-off between smoothness and goodness of fit. Solving (1) for the spline coefficients we get:

$$\widehat{z} = (B^T B + \lambda D^T D)^{-1} B^T y \tag{2}$$

The Whittaker smoother [20] can be viewed as a special case of the P-spline smoother. It arises when $B = I$ and the observations are located on an equispaced grid and a knot is placed at each abscissa point. In the econometric literature this smoother is also known as Hodrick-Prescott filter [11]. It was proposed by Hodrick and Prescott as a tool to separate the cyclical component of a time series from raw data in order to obtain a smoothed version of the series. The smoothed time series has the advantage to be less influenced by short term fluctuations than by long term ones. In their paper Hodrick and Prescott suggest $\lambda = 1600$ as a good choice. Others, such as Kauermann et al. [12], have suggested to select the λ parameter using automatic procedures. Well known methods for the selection of the smoothing parameter are: Akaike Information Criterion, Cross Validation and Generalized Cross Validation.

AIC measures the relative goodness of fit correcting the log-likelihood of the fitted model by its effective dimension (number of parameters): $AIC = 2ED - 2\ell$ (with ED we indicate the effective dimension of the model and with ℓ the log-likelihood). Following Hastie and Tibshirani [10] we can compute the effective dimension as $ED = \text{tr}[(B^T B + \lambda D^T D)^{-1} B^T B]$ for the P-spline smoother while $ED = \text{tr}[(I + \lambda D^T D)^{-1}]$ is the effective dimension of the Whittaker smoother. Instead of the log-likelihood it is convenient to use the deviance $dev(y, z) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 / \widehat{\sigma}_0^2$, where $\widehat{\sigma}_0^2$ is a constant equal to the variance of the residuals obtained considering $\lambda = 0$ as shown in [5]. The AIC function for our smoothing problem can be written as:

$$AIC(\lambda) = dev(y, z) + 2ED + 2n \ln \widehat{\sigma}_0^2 - n \ln 2\pi \tag{3}$$

The optimal parameter is the one that minimizes the value of $AIC(\lambda)$.

An alternative selection method is the cross validation. This criterion suggests to choose the parameter that minimizes:

$$CV(\lambda) = \sum_{i=1}^{n} \left[\frac{y_i - \widehat{y}_i}{1 - h_{ii}} \right]^2 \tag{4}$$

where h_{ii} is the ith diagonal entry of $H = (B(B^T B + \lambda D^T D)^{-1} B^T)$ for a P-spline smoother or $H = (I + \lambda D^T D)^{-1}$ in the case of Whittaker smoother.

Analogous to CV is the generalized cross validation measure [18]:

$$GCV(\lambda) = \sum_{i=1}^{n} \left[\frac{y_i - \widehat{y}_i}{n - ED} \right]^2 \tag{5}$$

where $ED = \text{tr}(H)$. In analogy with cross validation procedure we select the smoothing parameter that minimizes $GCV(\lambda)$.

3 L-curve selection method

In Section 1 we mentioned that the L-curve is a parameterized curve describing the trade-off between the two goals of every regularization/smoothing procedure: the goodness of fit and the smoothness of the final estimate. This approach was originally proposed by Hansen [8] to tune the strength of the penalty in a ridge regression framework.

3.1 The L-curve in ridge regression

Ridge regression is a common regularization tool in statistics. The idea behind this approach is to add a penalty term to the standard regression problem in order to obtain a shrinkage of the final estimates. The ridge penalty takes into account the magnitude of the coefficient vector:

$$\begin{aligned} &\underset{\beta}{\text{argmin}} \, \|y - X\beta\|^2 + \lambda\|\beta\|^2 \\ &\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T y \end{aligned} \tag{6}$$

The strength of the shrinkage depends on λ. Methods analogous to those briefly introduced in Section 2 can be used to select this parameter but, the L-curve represents a valid alternative selection procedure.

Let us define

$$\{\omega(\lambda); \theta(\lambda)\} = \{\|y - X\beta\|^2; \|\beta\|^2\}$$

then the L-curve is given by:

$$L = \{\psi(\lambda); \phi(\lambda)\} = \{\log(\omega); \log(\theta)\} \tag{7}$$

Figure 2 shows a L-curve obtained for a toy ridge regression analysis. The data where simulated considering two explanatory variables (say x_1 and x_2) extracted from a bivariate normal distribution with mean vector $\mu = [0, 0]$ and a high correlation. The dependent variable was obtained as a linear combination of the two independent variables plus a random noise $y_i = 2x_{i,1} + x_{i,2} + N(0, 1)$ with $i = 1, ..., 200$.

The shape of the curve explains its name. We notice that it shows a corner in an area characterized by intermediate values of ψ, ϕ and λ. Hansen suggested to select the regularization parameter located in the corner of the L-curve. The corner corresponds to the point of maximum curvature. The curvature can be computed using:

$$k(\lambda) = \frac{\psi' \phi'' - \psi'' \phi'}{[(\psi')^2 + (\phi')^2]^{3/2}} \tag{8}$$

The maximization of $k(\lambda)$ requires the computation of the first and second derivatives but the computations can be simplified in some cases. We will show this simplified procedure in the next sections using some simulated and real examples. Furthermore in the Appendix A we will also give some mathematical results that help to understand this result.

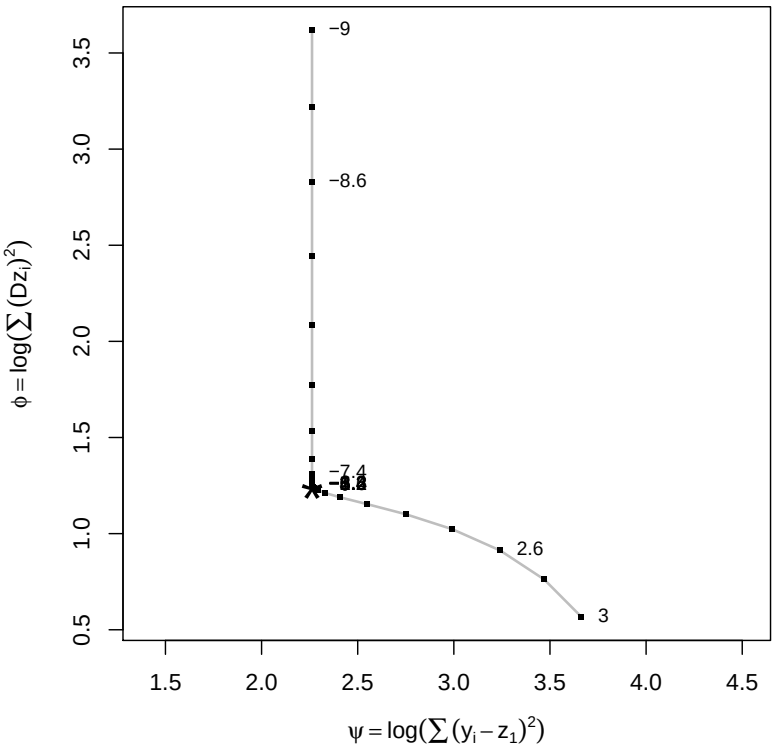


Figure 2: A typical L-curve obtained for a simulated ridge regression example. The star point represents the point of maximum curvature while the other points represent the points used to draw the curve. For some points the associated logarithmic value of the parameter is shown ($\log(\lambda) \in [-9, 3]$).

3.2 The L-curve in penalized smoothing

In Section 2 we briefly introduced the P-spline and the Whittaker smoothers. Let consider a P-spline smoother. In this case the following quantities can be defined:

$$\{\omega(\lambda);\theta(\lambda)\} = \{\|y - Bz\|^2; \|Dz\|^2\}$$

and the L-curve is given by:

$$L = \{\psi(\lambda);\phi(\lambda)\} = \{\log(\omega);\log(\theta)\} \tag{9}$$

The L-curve for a simple smoothing problem is depicted in figure 3. These results were obtained using 200 observations simulated as follows: $y = 5\sin(x) + N(0, 0.5)$ where $x \in [0, 2\pi]$.

The lower panel of figure 3 shows the pointwise curvature function and the Euclidean distance between adjacent points of the L-curve of the upper panel. We notice that the smoothing parameters selected maximizing the curvature and minimizing the Euclidean distance between adjacent points are the same. To understand why it is the case we can look at the L-curve closely (see figure 2 or the second panel of figure 3). The density of the points defining the curve tends to increase moving from the tail to the corner of the L. We can exploit this characteristic to simplify the selection procedure.

The curvature function in the case of a L-curve built for a smoothing problem is given by:

$$k(\lambda) = \frac{\psi'\phi'' - \psi''\phi'}{((\psi')^2 + (\phi')^2)^{3/2}}$$

The rate of change of the arc length distance between each point of the curve w.r.t. the λ parameter describes how the density of the points changes. This quantity is given by:

$$\frac{ds}{d\lambda} = \sqrt{\left(\frac{d\psi}{d\lambda}\right)^2 + \left(\frac{d\phi}{d\lambda}\right)^2} \tag{10}$$

The derivative $\frac{ds}{d\lambda}$ is a monotone function of the denominator of $k(\lambda)$. Minimizing it we obtain a good approximation of $\max\{k(\lambda)\}$ when the L-curve shows a clear convex area (i.e. when there is a corner). On the other hand, this also means that we can further simplify the selection procedure. Indeed the corner, if it exists, coincides (at least approximately) with the point satisfying:

$$\min\left\{\sqrt{(\Delta\psi)^2 + (\Delta\phi)^2}\right\} \tag{11}$$

The criterion in (11) suggests that the best smoothing parameter can be selected minimizing the Euclidean distance between adjacent points on the L-curve as shown in the last panel of figure 3.

However if the L-curve does not show a distinguishable convex area, i.e. if its curvature function is negative everywhere, the selection criterion cannot be reduced to (11). In these cases the curvature based selection criterion has to be preferred because it tends to select a λ parameter as close as possible to the optimal one.

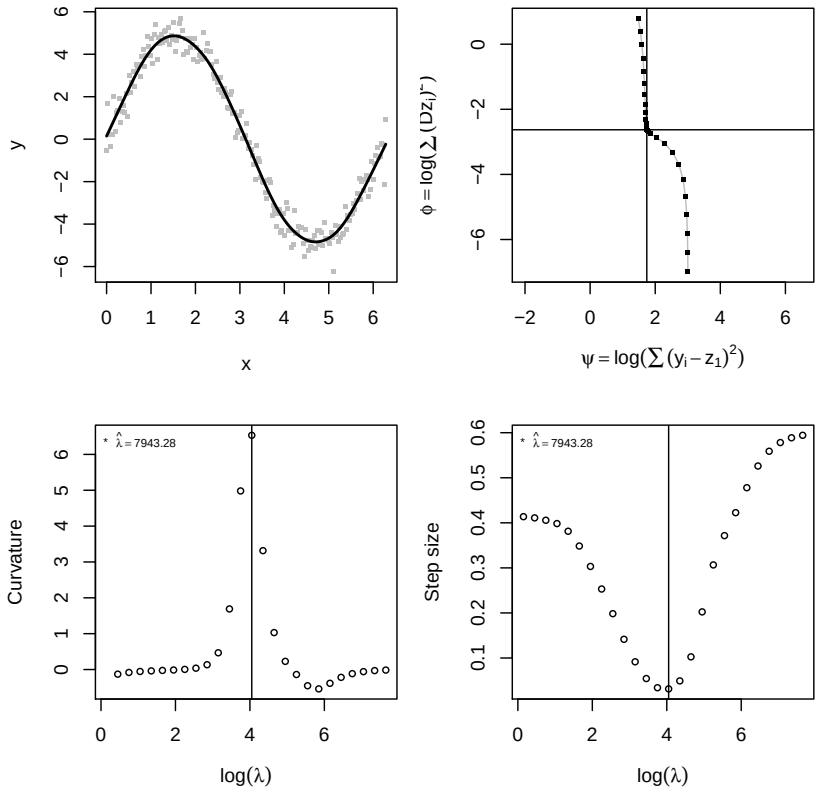


Figure 3: P-spline smoothing of simulated data using a L-curve approach. The first panel shows the obtained smoothing function (line) and the second the associated L-curve. The lower panels show the curvature function and Euclidean distance between adjacent points of L-curve. The values of these functions are plotted against different values of $\log(\lambda)$.

3.3 The shape of the L-curve

The shape of the L-curve is influenced by the data characteristics and determines the reliability of the selection procedure. To evaluate the impact of the data on the shape of the L-curve in this section we show some simulated examples. A formal discussion about this topic can be found in Appendix A.

Let us start by considering figure 5. The first panel shows the L-curve for an H-P filter applied to data simulated using the following scheme: $y = 10^{c_j}\sin(x_i) + N(0,1)$ with $x_i = 1,...,2\pi$ for $i = 1,...,200$ and $c_j = 2.5,...,-2$ for $j = 1,...,7$. It is clear that the convex region tends to disappear when the white noise component tends to be dominant on the trend component. Furthermore the four panels in the lower part of the figure show how the obtained smoothing functions change in accordance with the characteristics of the data. The smoothers seem to reproduce effectively the behavior of the data but we have to spend some words on the example in the first panel obtained for approximately white noise data. The estimated λ parameter was selected using a globally concave L-curve (see upper plot). The selection procedure suggests a high smoothing parameter (as we expected). In this and similar circumstances we suggest to select the smoothing parameter on the L-curve maximizing the curvature function. Indeed, even if the results cannot be considered reliable because there is not a clear convex region on the L-curve, the curvature selection criterion suggests a regularization parameter as large as possible corresponding to an area of the curve with a curvature as close as possible to be positive.

Also the characteristics of the error component influence the shape of the L-curve. First of all the the variability of the white noise component plays a role. Figure 4 shows that higher the variability is, less sharp the L-curve appears. These results where obtained using 200 observation simulated as follows: $y = \sin(x) + N(0,\sigma_j)$ with $x = 1,...,2\pi$ and $\sigma_j \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$. All the smoothing functions shown in the lower panels seem to efficiently catch the behavior of the data. Figure 6 reproduces the distributions of the residual mean square errors computed for the fitted values and the underlying trend component considering data with random noise showing different standard deviations (we considered now 1000 simulated observations). The mean deviations for each variability level is close to zero even if those with a less variable noise are smaller.

The L-curve is particularly useful if we want to smooth data with autocorrelated noise. However the shape of the curve depends on the strength of the serial correlation of the error component. Figure 8 shows some results obtained using the Whittaker smoother on a set of data simulated as follows: $y = 3\sin(x_i) + AR(1,\rho_j,\sigma = 1)$ with $x_i = 1,...,2\pi$ for $i = 1,...,200$ and $\rho_j = 0,...,0.9$ for $j = 1,...,7$ where ρ indicates the autocorrelation coefficient. Lightly autocorrelated noise produces really sharp L-curves while higher degrees of serial correlation reduce the sharpness. In any case the curves show clear convex areas. The lower part of figure 8 shows smoothing functions associated to some of the L-curves plotted in the upper part. It is possible to appreciate how well they reproduce the underlying data behavior in each scenario. Furthermore figure 7 summarizes the performances obtained using a larger set of data (1000 observations) simulated as before. It shows the distributions of the scaled squared deviations of the estimated smoothing functions from the underlying trend component for each simulation setting. The mean deviations are all close to zero even if the variability of the distributions seem to be influenced by the autocorrelation of the noise.

In addition to these considerations we found also that the L-curve seems to be less sharp in smoothing spline regression than in the applications proposed in the literature (typical

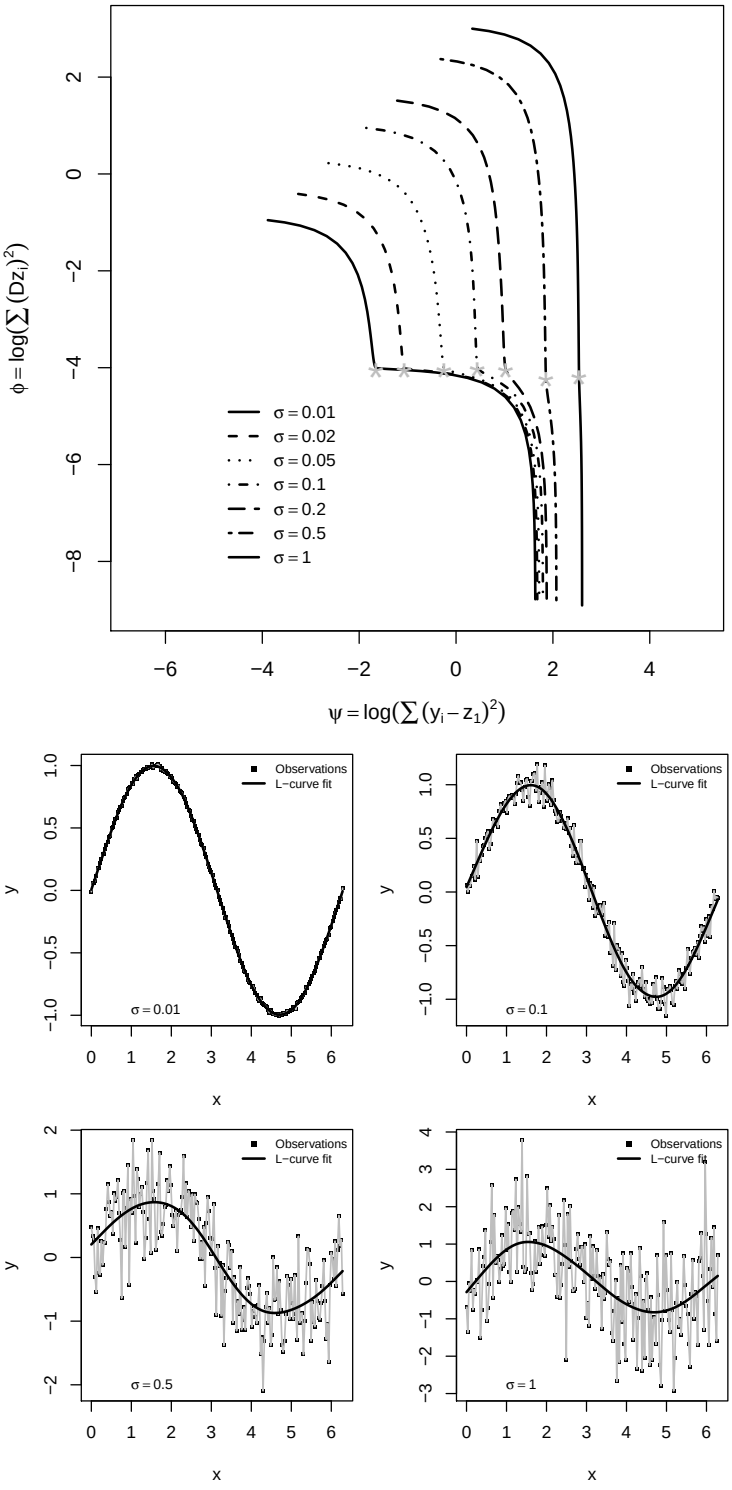


Figure 4: The first panel shows seven L-curves obtained for a Whittaker smoother estimated on data, with different variability for the white noise component. The lower panels show four smoothers obtained considering data simulated using noise components with increasing standard deviations (indicated in the lower legend of each plot). These smoothing functions were obtained considering $\log(\lambda) \in [-2, 9]$. The trend component was obtained considering $x \in [0, 2\pi]$ and $y_0 = \sin(x)$ for $i = 1, ..., 200$.

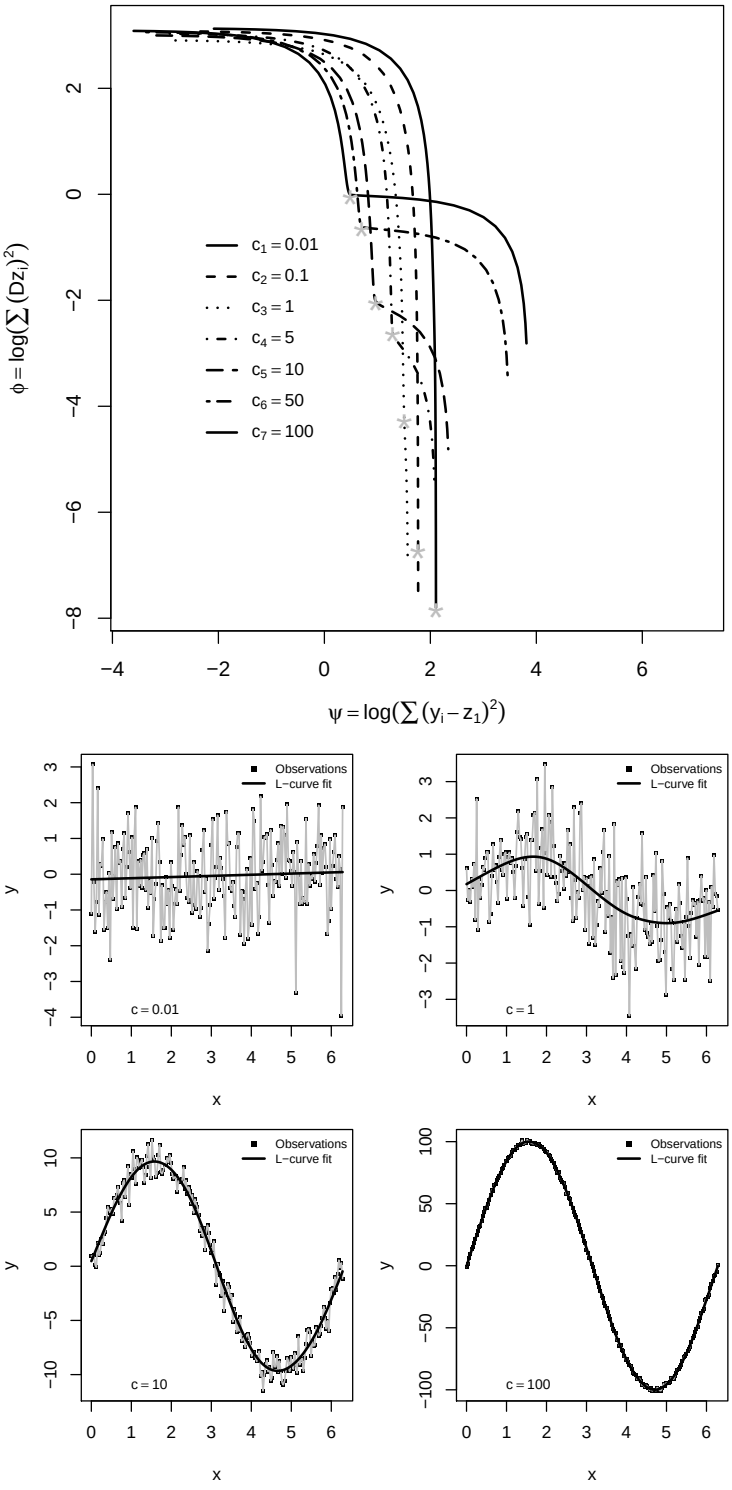


Figure 5: The first panel shows seven L-curves obtained for a Whittaker smoother estimated on data with different weights for the signal component while, the lower panels show four smoothers obtained considering data simulated using different weights for the this component (indicated in the lower legend of each plot). All these results were obtained considering $\log(\lambda) \in [0, 9]$.

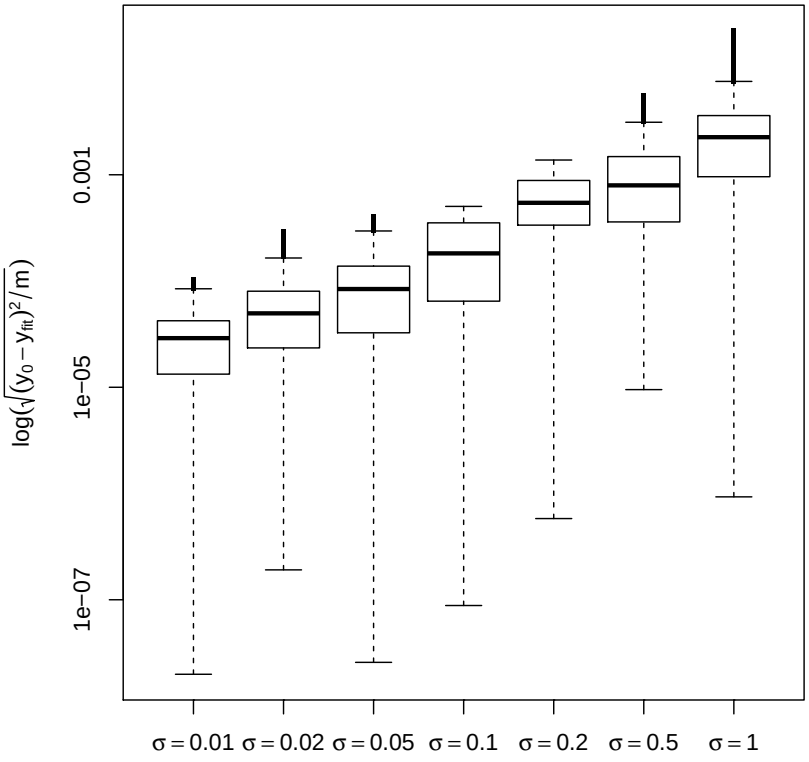


Figure 6: Distributions of RMSE in log scale computed on the estimated values and the underlying trend considering data with noise components characterized by different degrees of variability. These results were obtained considering $\log(\lambda) \in [-2, 9]$ and simulating 1000 observations.

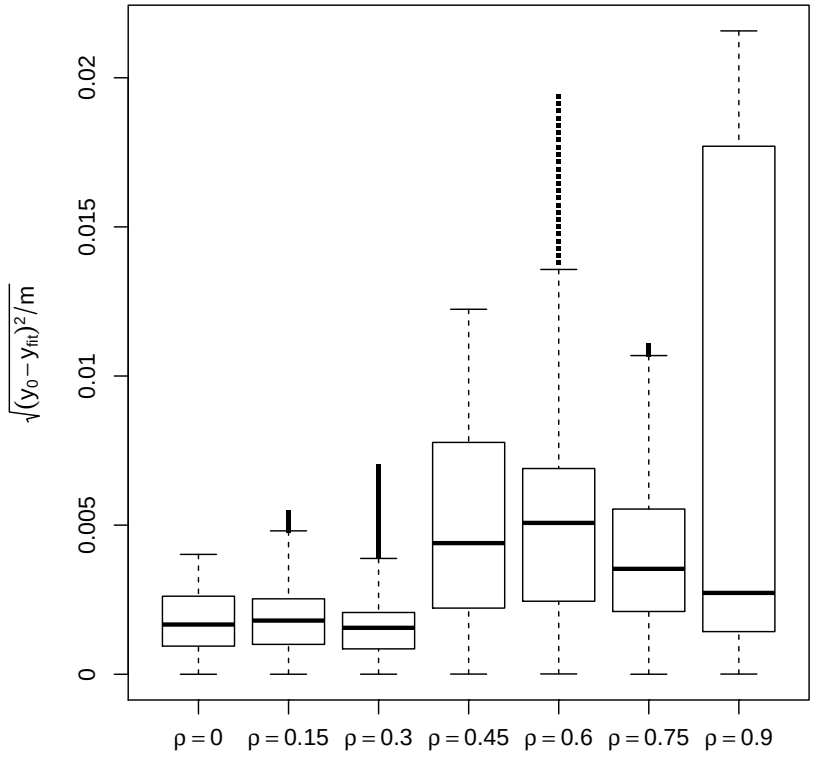


Figure 7: Distributions of the RMSE between estimated values and underlying function for different autocorrelation coefficients for the noise component of the data. These results were obtained considering $\log(\lambda) \in [-2, 9]$ and simulating 1000 observations.

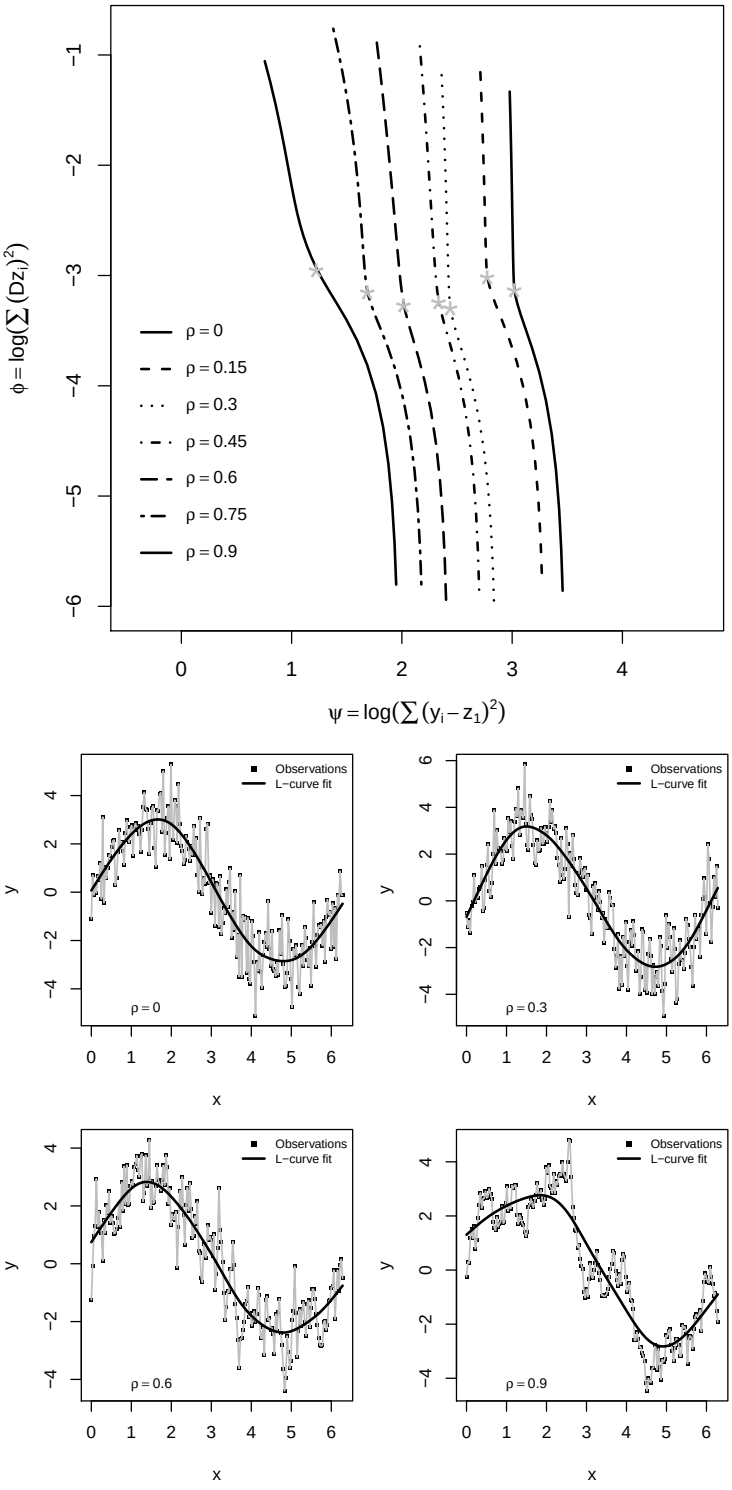


Figure 8: The first panel shows seven L-curves for a Whittaker smoother estimated on data with an increasing autocorrelation coefficient for the noise component while, the lower panels show the smoothing functions associated to four L-curves depicted in the upper panel. All these results were obtained considering $\log(\lambda) \in [-2, 9]$.

example is the ridge regression analysis). In our opinion it does not invalidate the applicability of the methodology. Indeed we believe that, as long as a convex area is well distinguishable, the procedure can be considered reliable.

4 Some examples

In this section we test the performances of the L-curve criterion using real datasets coming from different scientific fields. The first example that we would like to discuss is the orange juice price data [17] already introduced in Section 1. The original dataset contains three time series: the average producer price for frozen orange juice, the producer price index for finished goods and the number of freezing degree days at the Orlando airport. The orange juice price series was divided by the overall Producer Price Index for finished goods to adjust for general price inflation. As we did before, we will concentrate only on the monthly series of the prices.

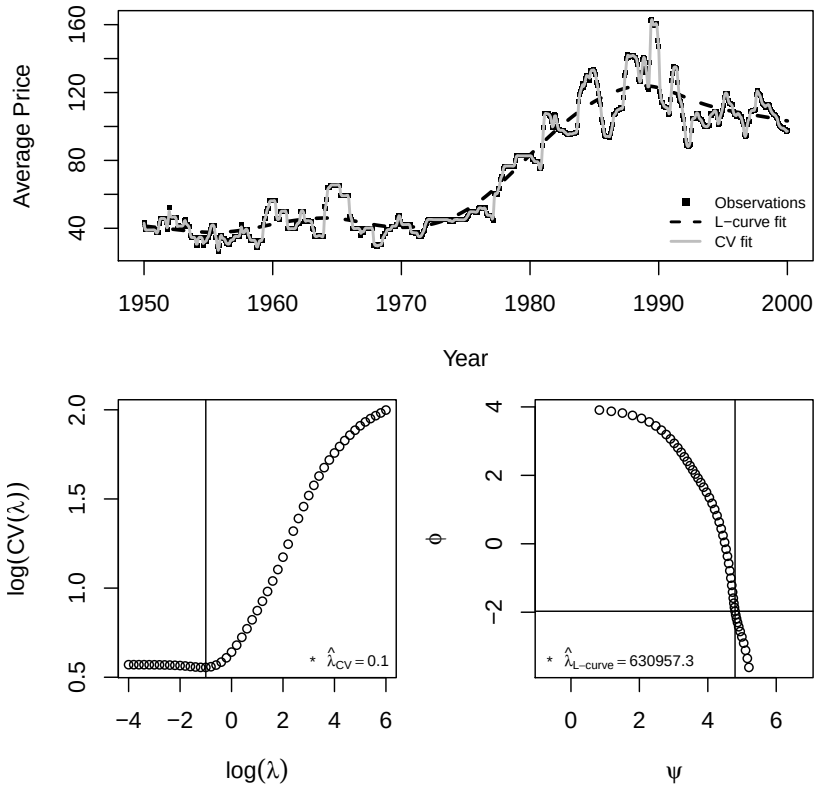


Figure 9: The upper panel compare the results obtained selecting the λ parameter of a H-P filter using the cross validation an the L-curve methods. The lower panels show the cross validation and L-curve profiles and indicate the selected smoothing parameters. For both selection methods we considered $\log(\lambda) \in [-4, 6]$.

In figure 9 the results obtained smoothing this time series with a Hodrick-Prescott filter and the associated cross validation and L-curve profiles are shown. The smoothing parameters selected by these two procedures lead to very different smoothing functions as we already

noticed in Section 1. Cross validation suggests a small λ and the result is a rough fitting function. On the other hand the L-curve suggests a larger parameter and the estimated H-P filter is able to reproduce the trend behind the data.

Another interesting application that we would like to show concerns the smoothing of the wood data. Pandit and Wu [14] present a dataset describing 320 measurements of a block of wood that was subject to grinding. In figure 10 the profile height at different distances is drawn. The profile variation follows a curve determined by the radius of the grinding stone. We use the Whittaker smoother to analyze these data and compare the performances of the L-curve and the cross validation for the smoothing parameter selection. The fitted curves and the related selection criteria are shown in figure 11. Also in this case the smoothing procedure built using the L-curve efficiently reproduces the trend in the data while the filter based on cross validation does not.

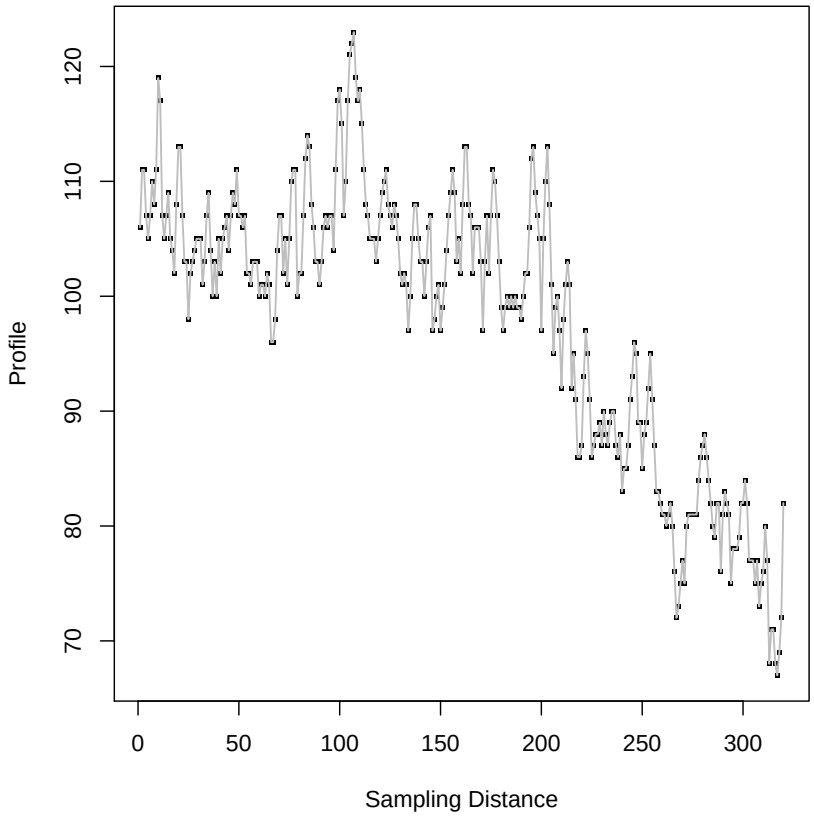


Figure 10: Profile of a block of wood subject to grinding.

As last example we analyze the time series of the annual mean sea level registered by the Dutch station of Delfzijl. This time series considers a period between 1865 and 2010 without missing values. The annual mean values are in millimeters. The data can be downloaded from `http://www.psmsl.org/data/obtaining/rlr.annual.data/24.rlrdata` and are summarized in figure 12. As in the previous cases the cross validation procedure gives a rough smoothing function while the H-P filter obtained tuning the smoothing parameter with

the L-curve catches the trend in the data. The behavior of the Euclidean distance between adjacent points (third panel in the lower part of the figure) shows some local minima. Indeed this dataset gives us the opportunity to briefly discuss another important issue related to the smoothers of the P-spline class. Welham and Thompson [19] showed the possibility of bimodality in the the smoothing parameter log-likelihood profile. Figure 13 shows that the bimodality of the cross validation profile is reproduced in the Euclidean distance profile. However the criterion computed on the L-curve shows a large difference between the two minima while it is not true for the cross validation criterion.

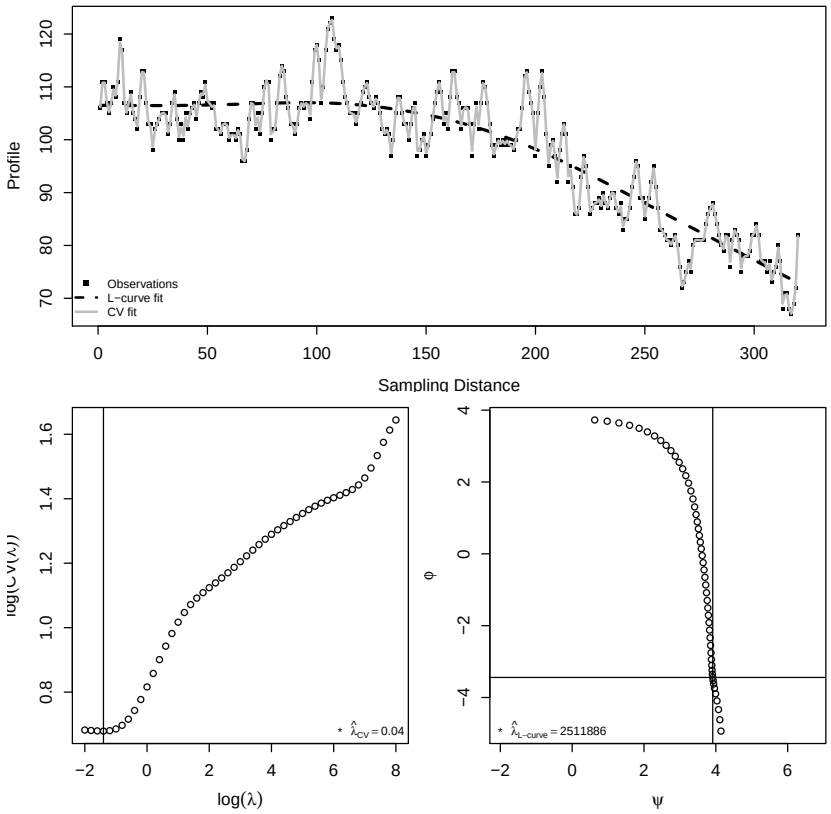


Figure 11: Whittaker smoothing of wood data. The upper panel shows the results obtained selecting the parameter by the L-curve and the cross validation. The lower panels represent the L-curve and the cross validation functions and the selected parameters. These results were obtained considering $\log(\lambda) \in [-2, 8]$.

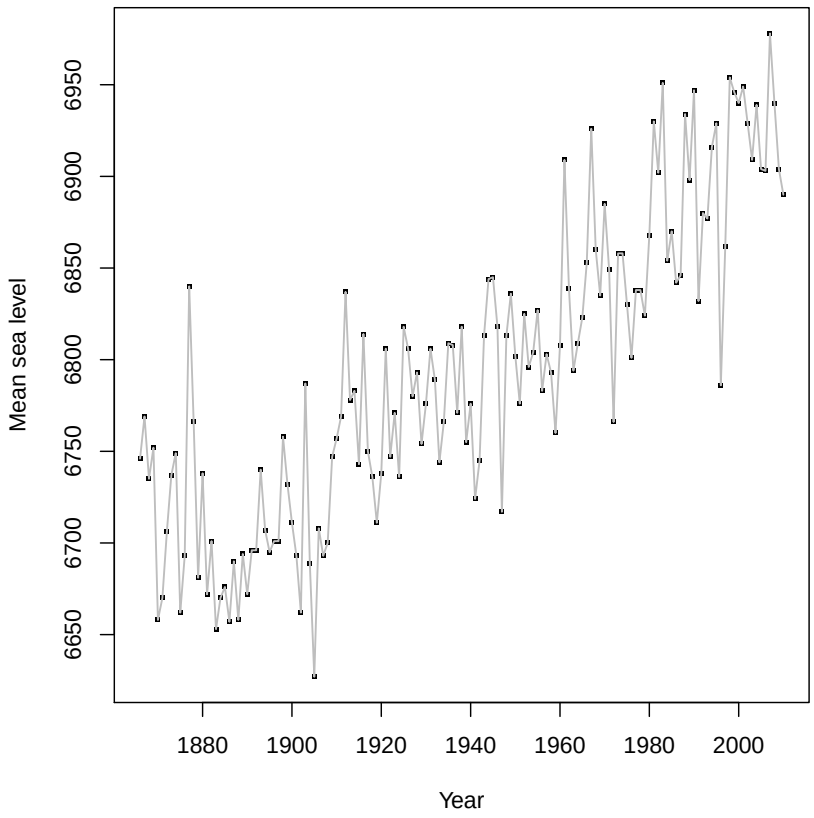


Figure 12: Annual mean sea level registered in Delfzijl for the period 1865-2010.

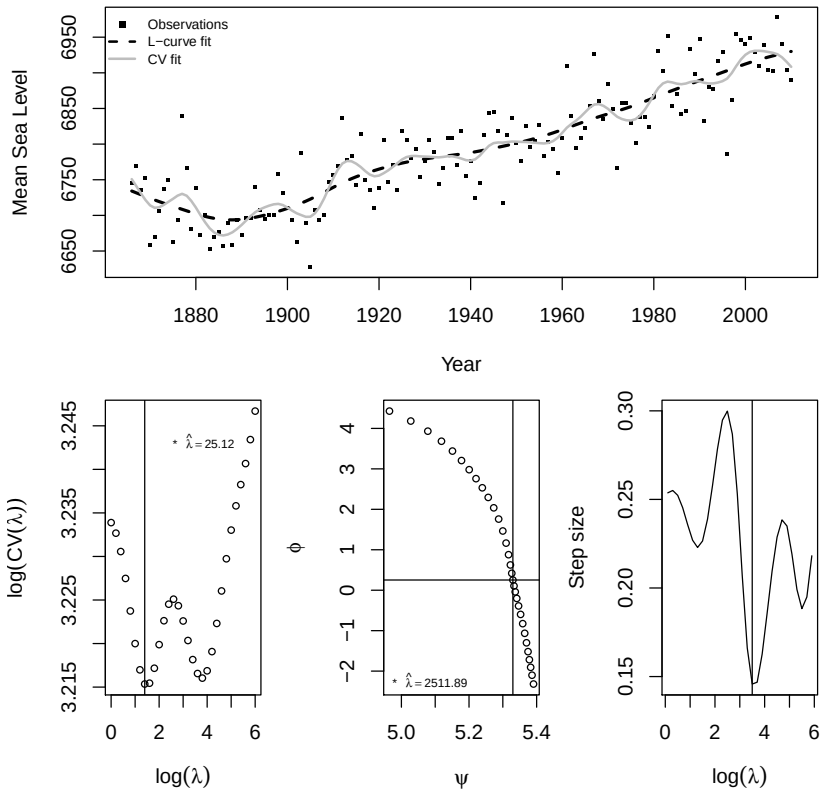


Figure 13: Whittaker smoothing of mean sea level data. The upper panel show the results obtained selecting the parameter using the L-curve and the cross validation ($\log(\lambda) \in [0, 9]$). The lower panels reproduce the L-curve, the cross validation curves and the selected parameters.

5 Discussion

In this work a L-curve procedure for the selection of the smoothing parameter in a P-splines framework is presented. This approach was originally introduced by Hansen et al. to select the optimal regularization parameter in ridge regression analyzes [8] and [9].

The L-curve selects the optimal smoothing parameter comparing the goodness of fit and the roughness of the final estimates. The compromise between this two issues of every smoothing analysis is summarized by a L-shaped curve for a range of λ parameters. The optimal smoothing parameter is selected maximizing the local curvature on the 'L', i.e. locating its corner. However it is possible to show that, under regularity conditions, the procedure for the location of the optimal smoothing parameter on the L-curve can be simplified. Indeed it is approximatively possible to locate the corner of the 'L' minimizing the Euclidean distance between adjacent points on the curve.

The proposed approach shows some relevant advantages. First of all it guarantees good performances in those cases in which the methods inspired by a cross validation framework lead to inappropriate results. We refer, for example, to those cases in which the data show a serial correlated noise component. The L-curve procedure is also advantageous in terms of computational efficiency. Indeed it does not require the computation of the effective dimension of the smoother at each step.

Other robust selection methods have been proposed in the literature. For example Currie and Durban [1] suggested to use a REML approach in order to take into account the correlation structure of the data. In their work the authors successfully analyzed the wood data [14] modeling the correlation structure through an $AR(2)$ model. They obtained results really close to ours (see Section 4). However we believe that our approach, besides being computationally more convenient, guarantees more flexibility because it is easily applicable to extreme cases such as the orange juice example.

On the other hand the L-curve criterion could lead to no reliable results in some cases. It usually happens smoothing data approaching to a pure white noise or when the signal component underlying the data tends to disappear. In these cases the L-curve procedure based on the maximization of the curvature function has to be preferred to the criterion based on the minimization of the Euclidean distance between adjacent points on the curve. This can be evaluated in figure 14 where a white noise is smoothed selecting the optimal λ parameter on the L-curve through both procedures.

The L-curve procedure offers many opportunities for further research. Indeed, in our opinion, it is not still clear why the corner contains information about the optimal smoothing parameter and why it is a robust selection method in the case of data with correlated noise. Furthermore we believe that it is also possible to generalize this methodology. Our future research will concentrate on a L-curve criterion for multivariate smoothing analyzes and on a L-curve based procedure suitable for spatially adaptive smoothing problems. We will also study the applicability of this procedure to a generalized linear model setting for smoothing of counts and binary data.

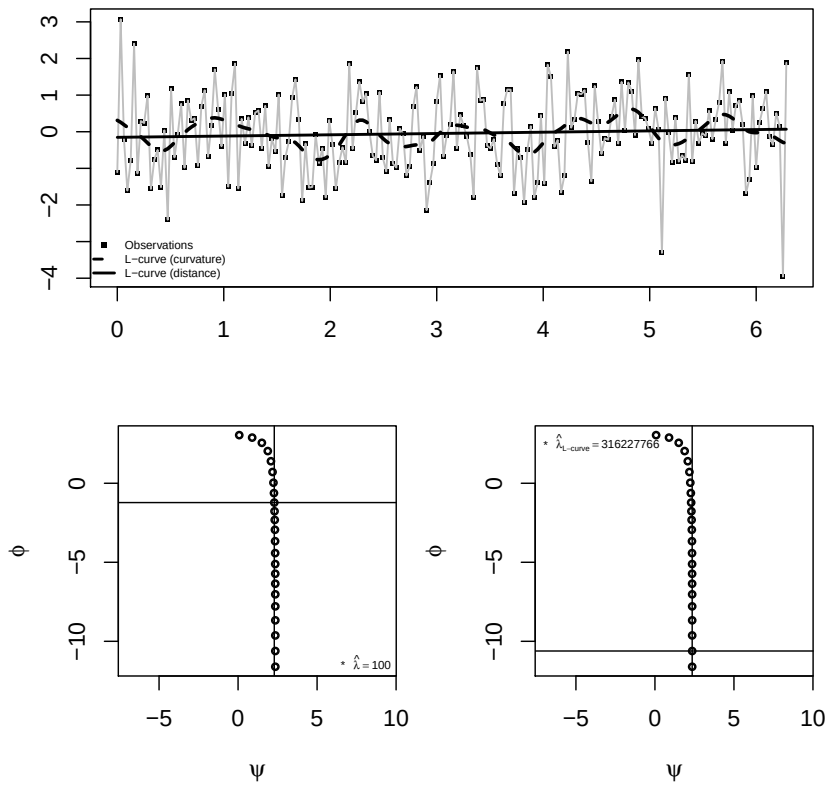


Figure 14: Whittaker smoothing of a white noise. The upper panel shows the results obtained selecting the parameter using the L-curve and the cross validation ($\log(\lambda) \in [0, 9]$). The lower panels reproduce the L-curve selections based on the curvature and on the Euclidean distance criteria.

References

[1] I. D. Currie and M. Durban. Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 4:pp. 333–349, 2002.

[2] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, 1978.

[3] Paul Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, 1995.

[4] Paul H. Eilers. A perfect smoother. *Analytical chemistry*, 75(14):3631–3636, July 2003.

[5] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):pp. 115–121, 1996.

[6] Paul H. C. Eilers and Brian D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.

[7] P. C. Hansen. The L-curve and its use in the numerical treatment of inverse problems. In *Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

[8] Per C. Hansen. Analysis of Discrete Ill-Posed Problems by Means of the L-curve. *SIAM Review*, 34(4):pp. 561–580, 1992.

[9] Per C. Hansen and Dianne P. O'Leary. The use of the L-Curve in the regularization of discrete ill-posed problems. *SIAM J. SCI. COMPUT.*, 14(6):pp. 1487–1503, 1993.

[10] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

[11] Robert J. Hodrick and Edward C. Prescott. Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking*, 29(1):pp. 1–16, 1997.

[12] Goeran Kauermann, Tatyana Krivobokova, and Willi Semmler. Filtering time series with penalized splines. *Studies in Nonlinear Dynamics & Econometrics*, 15(2):2, 2011.

[13] Tatyana Krivobokova and Goran Kauermann. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102:1328–1337, December 2007.

[14] S. M. Pandit and S. M. Wu. *Time series and system analysis with applications*. Krieger, 1993.

[15] A Pressley. *Elementary Differential Geometry*. Springer-Verlag, 2001.

[16] T. Reginska. A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.*, 17:740–749, May 1996.

[17] James H. Stock and Mark W. Watson. *Introduction to econometrics*. The Addison-Wesley series in economics. Addison-Wesley, Boston, Mass. [u.a.], internat. ed edition, 2003.

[18] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[19] S. J. Welham and R. Thompson. A note on bimodality in the log-likelihood function for penalized spline mixed models. *Comput. Stat. Data Anal.*, 53:920–931, February 2009.

[20] E. T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.

A Formal discussion about the shape of the L-curve

It is possible to show some analytical results about the shape of the L-curve. These results give some intuitions about the characteristics of the curve described above. To keep the mathematics as simple as possible let consider the Whittaker smoother even if our further reasoning can be generalized to other smoothing procedures. Let $\widehat{z}_\lambda = (I + \lambda D^T D)^{-1} y$ be the vector of regularized fitted values (we will omit the hat and the subscript symbols in order to simplify the notation). Plugging z_λ in the expressions for $\omega(\lambda)$ and $\theta(\lambda)$ we get:

$$\begin{aligned} \omega(\lambda) &= \|y - (I + \lambda P)^{-1} y\|^2 \\ \theta(\lambda) &= \|D(I + \lambda P)^{-1} y\|^2 \end{aligned}$$

where $P = D^T D$.

For our discussion it is necessary do determine the first derivative of z with respect to λ. We can compute this derivative using implicit differentiation as follows:

$$\begin{aligned} y &= (I + \lambda P) z \\ 0 &= (I + \lambda P) \frac{dz}{d\lambda} + Pz \\ \frac{dz}{d\lambda} &= -(I + \lambda P)^{-1} Pz \end{aligned}$$

Let $H = (I + \lambda P)^{-1}$ and differentiate ω and θ with respect to λ:

$$\begin{aligned} \frac{d\omega}{d\lambda} &= \left(\frac{d\omega}{dz}\right)^T \frac{dz}{d\lambda} \\ &= -2(y - z)^T \frac{dz}{d\lambda} \\ &= 2(z^T P z - z^T H P z) \\ &= 2[z^T (I - H) P z] \\ &= 2\lambda z^T P H P z \end{aligned} \tag{A.1}$$

$$\begin{aligned} \frac{d\theta}{d\lambda} &= \left(\frac{d\theta}{dz}\right)^T \frac{dz}{d\lambda} \\ &= (2Pz)^T \frac{dz}{d\lambda} \\ &= -2z^T P^T H P z \end{aligned} \tag{A.2}$$

Given that $P^T = P$, it is easy to notice that:

$$\frac{d\omega}{d\lambda} = -\lambda \frac{d\theta}{d\lambda} \tag{A.3}$$

From (A.3) we deduce that θ is a decreasing function of ω. Considering relation (A.3) we can analyze the convexity of the curve:

$$
\begin{aligned}
\frac{d\omega}{d\lambda}\frac{d\lambda}{d\theta} &= -\lambda \\
\frac{d\theta}{d\omega} &= -\frac{1}{\lambda} \\
\frac{d^2\theta}{d\omega^2} &= \frac{d}{d\omega}\left(-\frac{1}{\lambda}\right) \\
&= \left(\frac{1}{\lambda^2}\right)\frac{d\lambda}{d\omega} \\
&= \left(\lambda^2\frac{d\omega}{d\lambda}\right)^{-1} \\
&= \left(2\lambda^3 z^T P^T H P z\right)^{-1} \qquad \text{(A.4)}
\end{aligned}
$$

Equation (A.4) shows that the curve in normal scale is a convex function. As we said before the L-curve is defined in log-log scale. Let define the following quantities:

$$\psi = \log(\omega);\ \phi = \log(\theta)$$

Using logarithmic derivatives we obtain:

$$\frac{d\psi}{d\lambda} = \frac{d\omega}{d\lambda}\omega^{-1};\ \frac{d\phi}{d\lambda} = \frac{d\theta}{d\lambda}\theta^{-1}$$

Following the notation $\psi' = \frac{d\psi}{d\lambda}$ and $\phi' = \frac{d\phi}{d\lambda}$ we can write down the curvature function as:

$$k(\lambda) = \frac{\psi'\phi'' - \psi''\phi'}{((\psi')^2 + (\phi')^2)^{3/2}} \qquad \text{(A.5)}$$

The curvature function tells us something about the convexity of the curve. A convex part of the curve is characterized by a positive curvature. We are now considering the curve defined by $\{\psi(\lambda);\ \phi(\lambda)\}$. We can reparameterize this curve using $\{\psi;\ \phi(\psi)\}$ which has curvature function equal to:

$$k^*(\lambda) = \frac{\frac{d^2\phi}{d\psi^2}}{\left[1 + \left(\frac{d\phi}{d\psi}\right)^2\right]^{3/2}} \qquad \text{(A.6)}$$

This reparameterized curve has the same shape of the parametric one. Hence it has a positive curvature when the numerator of (A.6) is positive. Let start considering the denominator. The right part of the denominator of (A.6) can be written as follows:

$$
\begin{aligned}
\frac{d\theta}{d\omega} &= -\frac{1}{\lambda} \\
\frac{d\phi}{d\psi} &= -\frac{1}{\lambda}\frac{\omega}{\theta} \\
&= -\frac{\|y - z\|^2}{\|Dz\|^2}\lambda^{-1} = -\frac{S}{\lambda R} \qquad \text{(A.7)}
\end{aligned}
$$

Equation (A.5) clarifies that the curvature function can also be negative under some conditions. These can be evaluated considering the second derivative of ϕ w.r.t. ψ (i.e. using the curvature definition in (A.6)):

$$\frac{d^2\phi}{d\psi^2} = -\frac{d\gamma}{d\psi} = -\omega\frac{d\gamma}{d\omega} \tag{A.8}$$

where $\gamma = \frac{\omega}{\lambda\theta}$. We can now compute $\frac{d\gamma}{d\omega}$ taking into account the equation for γ:

$$\frac{d\gamma}{d\omega} = \frac{1}{\lambda\theta} - \frac{\omega}{\lambda^2\theta^2}\left[\frac{d\lambda}{d\omega}\theta + \frac{d\theta}{d\lambda}\lambda\right] \tag{A.9}$$

but we know that $\frac{d\theta}{d\lambda} = -\frac{1}{\lambda}$ and $\frac{d\lambda}{d\phi} = -\theta\frac{d\lambda}{d\omega}\lambda$. So we can write equation (A.9) as follows:

$$\frac{d\gamma}{d\omega} = \frac{1}{\lambda\theta} - \frac{\omega}{\lambda^2\theta^2}\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] \tag{A.10}$$

The numerator of the curvature function (A.6) is then equal to:

$$\frac{d^2\phi}{d\psi^2} = \frac{d\phi}{d\psi} + \left(\frac{d\phi}{d\psi}\right)^2\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] \tag{A.11}$$

The L-curve procedure suggests to locate the optimal smoothing parameter in the point of maximum curvature. For this reason it is convenient to find a condition for which $\frac{d^2\phi}{d\psi^2}$ assumes positive values. Given that $\frac{d\phi}{d\psi} < 0$, in order to have a positive curvature the following inequality has to hold:

$$\begin{aligned}
\left(\frac{d\phi}{d\psi}\right)^2\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] &> -\frac{d\phi}{d\psi} \\
\frac{d\phi}{d\psi}\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] &< -1 \\
-\frac{1}{\lambda}\frac{d\lambda}{d\phi} &> -\frac{d\psi}{d\phi} + 1 \\
-\frac{1}{\lambda}\frac{d\lambda}{d\phi} &> \frac{-d\psi + d\phi}{d\phi} \\
\frac{1}{\lambda}d\lambda &> d\psi - d\phi
\end{aligned} \tag{A.12}$$

Differentiating both sides of (A.12) w.r.t. λ and considering that ψ and ϕ are in log scale and that $\frac{d\lambda}{\lambda} = d\log(\lambda)$ we get the final relation:

$$1 > \frac{d\psi}{d\lambda} - \frac{d\phi}{d\lambda} \tag{A.13}$$

This condition can be verified numerically. Let consider, as examples, two cases showed above. In particular we take a case in which a clear corner is present and another case the convex area of the L-curve is not pronounced. The first case is given by the example in figure 3. Figure 15 shows the curvature function related with this example. The smaller segments

under this curve show the positive curvature points founded using (A.13). The second panel of figure 15 plots the numerator of the curvature function and the vertical lines indicate the points of positive curvature founded applying the criterion in (A.13). On the other hand figure 16 shows the results obtained considering the mean sea level example (in this example there was not a clear corner in the L-curve).

In both cases the criterion in (A.13) selects correctly all the points of positive curvature (and positive numerator of the curvature function) even if the L-curve does not show a really clear convex area (as in the second example).

Relation (A.13) tells us also something else. Indeed it says that in the area of positive curvature the rate of change of the numerator of k w.r.t. λ is slower than the rate of change for the denominator. To understand this we refer now to the first definition of the curvature function in equation (A.5). Given that the curvature is independent from the parameterisation the numerator of (A.5) as to be positive when the numerator of (A.6) is positive.

The denominator of (A.5) is equal to $((\psi')^2 + (\phi')^2)^{3/2}$. Remembering that $\frac{d\phi}{d\lambda} < 0$, from equation (A.13), we know that the denominator of (A.5) has to be between 0 and 1 in a convex area of the L-curve. This means that both $\frac{d\psi}{d\lambda}$ and $-\frac{d\phi}{d\lambda}$ have to be between 0 and 1 in a positive curvature area. On the other hand it is clear that if we square these quantities we get smaller values. So we can deduce that minimizing the denominator of (A.5) we locate a convex area of the L-curve (if it exists) and the rate of change of the numerator w.r.t. lambda is smaller than the rate of change of the denominator.

It is also possible to notice that the denominator of (A.5) is strictly related to our simplified selection criterion. Indeed our suggestion is to minimize $\sqrt{(\Delta\psi)^2 + (\Delta\phi)^2}$. Extending the previous reasoning to this quantity it is clear why this minimization procedure lead to a maximum curvature point for well-behaved L-curves.

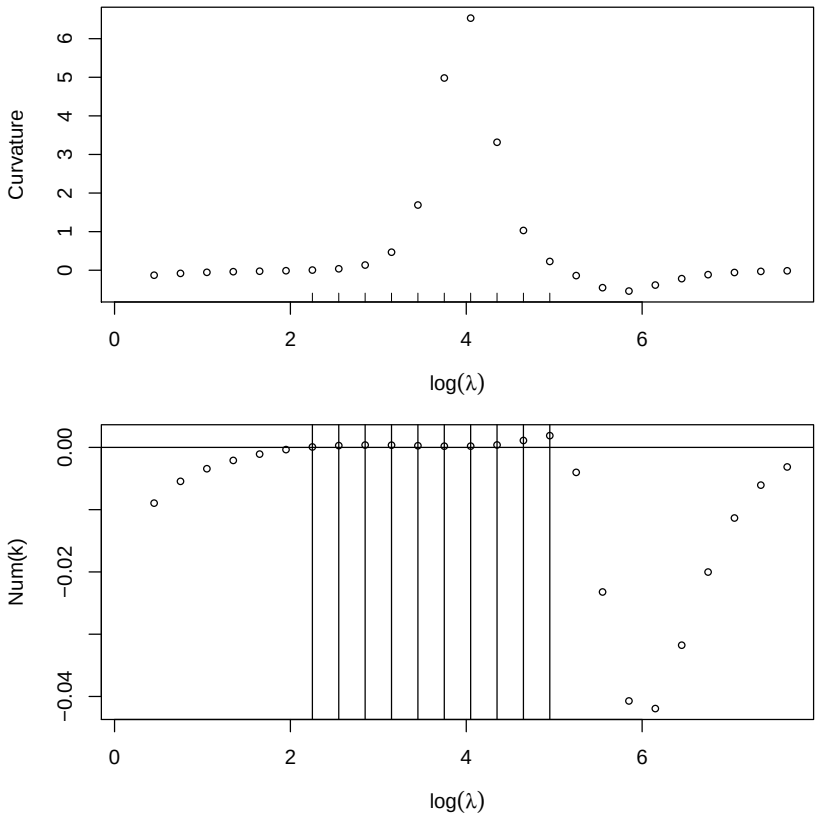


Figure 15: Curvature function and numerator of the curvature function for the example in figure 2. In the second panel the horizontal line indicates the zero abscissa level.

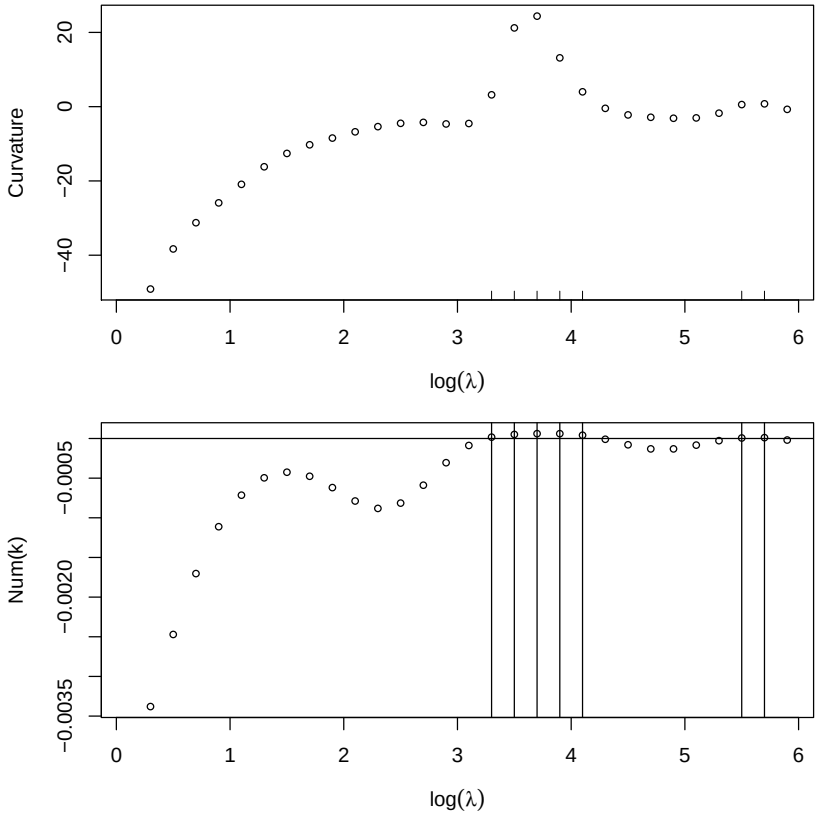


Figure 16: Curvature function and numerator of the curvature function for the mean sea level example. In the second panel the horizontal line indicates the zero abscissa level.