Universitá degli studi di Napoli Federico II

Dottorato in matematica per l'analisi economica e la finanza (XXV ciclo)

Splines, differential equations, and optimal smoothing

Gianluca Frasso

Aprile, 2013

CONTENTS

GENERAL INTRODUCTION AND OVERVIEW 1

1.1 Introduction

In many scientific areas it is of primary interest to describe the dynamics of a system, that is, how it evolves over time and/or space. In the simple one-dimensional case the state of a system at any time can be represented by a function $u(t)$ which' values track the evolution of a given phenomenon over time. It is also possible to consider phenomena evolving in time and space by using a function $u(t, x)$ depending on two independent variables.

Thus knowing t and/or x it is possible to evaluate the state of the system at at a given point in space and/or time. One way to obtain $u(\cdot)$ is taking measurements at different values of the independent variable(s) and fit the data in order to estimate a formula for u. This is the point of view exploited in statistical data analysis. Overparametric regression (smoothing) techniques are usually applied in this kind of studies.

On the other hand it is clear that such a model would tell us how the system evolves but is not able to clarify why the system behaves as has been observed. Therefore we try to formulate mathematical models summarizing the understanding we seek. Often these models are dynamic equations that relate the state function to one or more of its derivatives w.r.t. the independent variable(s). Such equations are called differential equations (or abbreviated as DE). Differential equations are common analytic tools in physics and engineering.

The first approach we have cited has the advantage to make the description of an observed phenomenon really flexible, being able to exploit all the information provided by the observed measurements. It becomes clear if we consider the applicability of overparametric smoothing techniques. These approaches, by the way, do not allow for a physical interpretation of the ob-

served dynamics. On the other hand, the main advantage of a differential modeling point of view is to highlight the physical determinants of a given phenomenon but completely ignore what has been observed.

The aim of this work is to present a flexible way to combine the statistical and the dynamic modeling points of view. To reach this goal we combine in a convenient way the flexible data description provided by a semiparametric regression analysis and the physical interpretability of dynamics summarized by differential equations.

1.2   *Differential equations*

A differential equation is an equation for an unknown state function u that connects it to some of its derivatives. A distinction is usually made between ordinary and partial differential equations. An ordinary differential equation (ODE) describes the relationship between a state function depending on a unique independent variable and some of its derivatives. On the other hand a partial differential equation (PDE) relates the state function depending on more than one independent variables and its partial derivatives. We refer to the specialized literature about differential equations for a more extensive description (see Coddington and Levinson, 1984 and Evans, 1998).

Differential equations are usually classified according to three characteristics: order, functional form and scalar versus systems of DEs. The first classification is according to the order of the equation. The order is defined to be the order of the highest derivative in the equation. Another classification divides DEs into two groups: linear and nonlinear. An equation is linear if the formulation relating the derivatives of u is a linear function of the unknown state function and its derivatives. Nonlinear equations are usually further subclassified according to the type of nonlinearity. In this work we mostly deal with linear DEs even if some of the concepts that are illustrated in the further chapters can be applied to non-linear problems, as we will see. Another classification distinguish between constant and varying coefficient DEs according to the fact that the parameters involved in the equation vary over time/space or ar fixed.

A function $u = u(t)$ is a solution of an ODE on an interval $I : a < t < b$ if it is differentiable on I and if satisfies it identically for all $t \in T$, when

substituted into the equation:

$$u^{(d)} = f(t, u(t)) \ \forall \ t \in I.$$

The function $u(t)$ that satisfies the equation above is not unique. The family of solutions of a DE is usually named "general solution". A unique solution of a differential problem can be found if some conditions are imposed to the general solution. This is true also for PDEs having a general solution that is a function of more than one independent variable.

An initial value problem (or compactly IVP) is a problem in which the differential equation has to be solved taking into account one or more conditions imposed to the state function and/or its derivative(s) at a given (initial) value of the domain. The initial condition determines the starting state of the system. This information becomes fundamental when we apply step-based numerical procedures such as the Runge-Kutta method (Golub and Ortega, 1992). These procedures approximate the state function using the numerical approximation to u and its derivatives from the previous step. On the other hand, a boundary value problem (BVP) is a differential problem that has to be solved taking into account a set of constraints called boundary conditions. These conditions are imposed at the boundaries of the domain of the solution.

Analytic solutions of differential equations may exist in a restricted number of cases. In many circumstances it is necessary to approximate the state function through some numerical procedure. These methods usually suggest to solve the differential problem substituting the continuous variables involved in it by some discretized version (see for example Golub and Ortega, 1992). Thus the continuum problem represented by the DE is transformed into a discrete problem in a finite set of variables.

In this work we will concentrate on the collocation procedure. Collocation belongs to the more general class of projection methods. The rationale behind all those methods is to approximate the solution of a differential problem by a finite linear combination of basis functions. Conceptually, the solution is represented as lying in some appropriate functional space and a projection method attempts to obtain an approximation on a finite-dimensional subspace defined by basis functions. The projection of the solution onto this subspace gives the numerical solution. In other words, if we define a basis

function $\phi_j(t)$, a projection method suggests to approximate the state function and the initial/boundary conditions by a linear combination of the ϕ_js:

$$u(t) = \sum_j^n c_j \phi_j(t). \tag{1.1}$$

It is clear that, in order to obtain a numerical solution, we need to determine the coefficients in equation (1.1). The collocation approach suggests to look for a vector of coefficients guaranteeing that the numerical solution approximates the exact one on a grid of values $t_1, ..., t_n$ not necessarily equally spaced defined "collocation points". The numerical procedure reduces to the solution of a system of linear equations in the case of linear DE problems. Differently from explicit multi-step approaches the collocation method attempts to approximate the state function over the entire domain at once. In the coming chapters, we will discuss a collocation procedure based on B-splines.

1.3   *Data description and dynamic models*

In the previous section we introduced the collocation scheme for the numerical solutions of DEs. Consider now a different problem. Suppose to observe a set of data which signal is compatible with a dynamic system described by a differential equation. We might know the DE and its coefficients, or just the DE and be interested in estimating them from the raw observations. Is it possible to exploit these information to obtain an appropriate description of the observed measurements?

For example, we may suppose to be interested in summarizing a set of data showing an oscillatory behavior. This dynamics is described by an ordinary differential equation even if our reasoning can be generalized to phenomena evolving in more than one direction.

A possible approach to describe what we observe is to define a suitable ODE and use its state function. On the other hand we can suppose to ignore where the measurements come from and so to ignore the differential law describing them. In this case an appropriate description of the observed data dynamics could be provided, for example, by a nonlinear regression tool.

In both cases the data dynamics is efficiently separated from the erratic component and a model describing the dynamic is approximated. Which one is more appropriate? The statistical approach has the advantage to be really

flexible. Indeed, as will be also shown in the coming chapters, a smoothing function is able to summarize the data trend imposing a limited number of hypotheses on its functional form. A strong disadvantage of this approach is that the information extracted by smoothing function is hard to interpret.

The state function solving the appropriate second order ODE gives a really compact and easy to interpret summary of the observed dynamics. If, for example, our oscilloscope indicates an harmonic behavior with a period close to 2.1 the phenomenon under consideration could be described solving the following BVP:

$$\frac{d^2y(t)}{dt^2} - 9y(t) = 0,$$
$$y(\min(t)) = b_1, y(\max(t)) = b_2.$$

One disadvantage of this approach is its rigidity. The final estimates are strictly related to the form of the differential problem. Even small changes of the DE parameters or of the boundary conditions lead to really different results. This becomes an important issue if we consider that in many cases we have a vague or incomplete knowledge about a differential model suitable to describe the observed dynamics. It becomes clear that a framework exploiting the flexibility of the statistical approach and the interpretability of the differential equation formulation would be ideal.

A first attempt could be done defining flexible prescribed conditions associated to the differential equation. Consider to have a precise idea about the differential equation describing the measured dynamics but to ignore the boundary values of the state function. The collocation procedure can be used to impose "soft conditions" on the state function. Those are conditions defined w.r.t. the observed data behavior and that have to hold in a least squares sense. In this case our problem could be solved estimating, from the raw observations, boundary conditions providing a numerical solution able to describe appropriately (in a least squares sense) the observed data. The result will be a curve describing the exact solution of the DE taking into account the estimated conditions.

A further generalization is also possible. Consider to know the differential equation that approximately describes the data behavior. We want to approximate the state function in such a way to give an appropriate description, in a least squares sense, of what we observe. This goal can be achieved

through a smoothing approach penalized by the differential operator $V(w, t)$ we have in mind. This approach is known in the statistical literature as L-spline smoothing (see for example Schumaker, 1981, Ramsay and Silverman, 2005 and many others). Roughness penalized smoothing splines can be viewed as a special case of L-splines considering a DE solved by a straight line state function.

We are looking for a function able to balance between a fidelity to the data criterion (expresses in terms of residual sums of squares) and a collocation scheme for the solution of the differential equation:

$$\min_c \|y - \sum_j^n \phi_j(t)c_j\|^2 + \lambda\|\sum_j^n \sum_{d=0}^2 w_d\phi_j^{(d)}(t)c_j\|^2, \qquad (1.2)$$

where w is a vector of parameters defining the differential operator and is supposed to be known for the moment and the index d indicates the dth order derivative of the state function involved in the DE. This approach can be seen as a limit case of soft constrained collocation procedure where the constraint is represented by a sum of squares minimization criterion. The result in this case is a curve providing an approximation of the state function solving the DE.

The smoothing parameter λ in (1.2) balances between the goodness of fit and the fidelity to the DE solution. It represents the relative importance of the data driven criterion over the numerical one. In principle this parameter indicates the appropriateness of the hypothesized differential model in describing the observed measurements. It can be imposed a priori or selected through an automatic procedure. A high optimal smoothing parameter testifies the adherence of the observed data to the hypothesized differential dynamics. On the other hand, for $\lambda \to 0$ the estimation the estimation procedure tends to emphasize the role of the residual sum of squares criterion leading to a rough fitting function. For this reason it is of crucial interest in many case to select this parameter in an automatic manner exploiting the data information. In this work we mainly deal with two smoothing parameter selection procedure: the mixed model approach proposed by Schall, 1991 and the robust and computationally efficient L-curve procedure.

Once that the data and the differential equation have been combined further developments of this framework are possible. Indeed, it can happen that

we don't have a precise knowledge about the differential law governing the observed measurements. If the considered differential equation has an analytic solution it is simple to infer the values of the unknown parameters from the raw data. It is often the case, for example, dealing with linear ODE. A common approach is to write the solution of the equation as explicit function of the parameter which are fitted to the data. Generally this leads to a nonlinear regression problem even if the DEs are linear. Also in these simple cases the application of a collocation procedure results convenient allowing for the avoidance of nonlinear data fitting.

On the other hand the information provided by the data can also be exploited in a penalized smoothing framework in order to estimate the unknown DE parameters. Suppose to observe the harmonic oscillatory dynamics already introduced and to ignore the parameters w defining the differential operator $V(\cdot)$. The optimization problem to be solved in this case becomes:

$$\min_{c,w} \|y - \sum_{j}^{n} \phi_j(t)c_j\|^2 + \lambda\|\sum_{j}\sum_{d=0}^{2} w\phi_j^{(d)}(t)c_j\|^2. \qquad (1.3)$$

The unknowns c and w can be estimated jointly or by a generalized cascading procedure exploiting the hierarchy existing between them (see Cao and Ramsay, 2007).

In this work we present an alternative P-splines-based two stage approach for the estimation of the unknown w. We move form the consideration that the state function solving the DE describing the observed data has to be consistent with the data signal if the differential model is appropriate. The signal can be separated from the noisy measurements using a P-spline smoother. The optimal penalized spline coefficients, giving a compact representation of the state function, can be multiplied by the set of basis functions defining the penalty term in (1.3). In this way we an approximation the functions involved in the differential law is obtained. These approximations can then be used to estimate the vector w through least squares.

The reconstructed DE penalty helps in interpreting the estimated smoothing function. In the hypothetical oscillatory dynamics we are considering we expect to estimate period approximately equal to 2.1 (this means that the parameter associated to the state function in the equation has to be close to 9) and an approximately null damping ratio (the parameter associated to the

first derivative of $u(t)$ has to be close to zero). On the other hand an high level of $\widehat{\lambda}$ would prove the appropriateness of the hypothesized differential law in describing the observed dynamics.

We may also be interested in analyze grouped data series, each driven by the same differential equation. In these cases the DE parameters can be considered in a mixed model formulation. The two stage procedure introduced above can also be applied to estimate mixed differential parameters. It becomes convenient in those circumstances in which raw data are grouped in some way. The grouping factors call for the consideration of the intra class correlation. Typical examples are represented by the data referred to different subjects or collected by repeated measurements over time. In the case of dynamic systems described by a differential law these considerations call for the estimation of fixed and random differential parameters.

1.4   *Thesis outline*

This thesis consists of eight chapters. In **chapter 2** the basic building blocks of our dissertation are introduced. In particular we briefly discuss the concepts of B-spline, tensor products of B-splines and the penalized spline smoothing procedure in one and two dimensions.

In **chapter 3** we discuss some smoothing parameter selection procedures. We mostly concentrate on the introduction of a selection criterion based on the "L-curve". The applicability of the L-curve framework to smoothing problems together with a novel and simpler selection criterion based on the "V-curve" is illustrated. The L-curve procedure has some advantages if compared to the classical ones being computationally more convenient and robust to serial correlation in the noise component of the data. This approach suggests to select the λ parameter balancing the goodness of fit and the size of the penalty applied in the estimation procedure. The performance is evaluated analyzing simulated and real data examples.

Chapter 4 generalizes the L-curve procedure introducing its two-dimensional extension: the L-surface. We discuss the applicability of the L-surface framework to non-isotropic multidimensional smoothing problems and also propose a novel selection criterion based on the "V-valley". The L-surface procedure shares the advantages of its one-dimensional analogous.

Its efficiency becomes an important issue considering that the computational effort easily increases analyzing multi-dimensional data. The performances of the L-surface procedure is evaluated analyzing simulated and real data examples.

In **chapter 5** we introduce another building block useful for the present dissertation: the B-spline collocation procedure for the solution of differential equations. We focus on the numerical approximation of state functions described by linear DEs. In this chapter we consider both ordinary and partial differential equations, discussing the solution of initial and boundary value problems. A practical application of the B-spline collocation scheme we present an example based on the solution of the Black and Scholes (Black and Scholes, 1973) equation for the pricing of European options.

Chapter 6 links numerical analysis of differential problems and data smoothing. In this chapter we introduce a differential penalized smoothing approach. We focus on one-dimensional data which dynamics is approximately described by ODEs. The penalty term is defined to be consistent with a collocation representation of the differential problem. We start considering known differential penalties without taking into account the initial/boundary conditions. Then we show how known or unknown conditions can introduced in the smoothing problem as "hard constraints". Then a P-splines-based two stage procedure is introduced to estimate the optimal differential parameters to be plugged into the penalty term. We evaluate the performance of the proposed method considering simulation studies. A real data analysis of the stomach contraction dynamics is discussed as applicative example.

In **chapter 7** we introduce a different point of view for the ODE parameter estimation problem treating it as the sum of a fixed and a random (subject-specific) component. A generalization of the P-splines based two stage procedure introduced in the previous chapter is proposed for the estimation of the unknown mixed DE parameters and variance component. In particular we suggest to exploit the relationship between penalized least squares and mixed models to obtain an efficient estimation of these unknowns. The variance components are estimated through the optimization of the smoothing parameter exploiting its variance ratio interpretation (see Schall, 1991 and Pawitan, 2001). The performances of the proposed framework are again eval-

uated through simulation studies. As real data application the stomach contraction MRI data introduced in chapter 6 are analyzed.

In **chapter 8** we generalize the framework described in chapter 6, discussing a two-dimensional smoothing approach penalized by partial differential operators. We first consider the case of known DE parameters. A tensor product P-spline-based two stage approach is then introduced for the estimation of the differential parameters from the raw observations. The performance is evaluated through simulation studies.

In **chapter 9** more advanced topics are treated. We introduce a (symmetrically and asymmetrically) weighted penalized least squares approach for the solution of differential problems with unconventional constraints. We analyze example related to both ordinary and partial differential equations. We also discuss a possible generalization of the collocation procedure for the solution of nonlinear differential problems and delay differential equations. The penalized smoothing approach discussed in the previous chapter is then generalized in order to analyze dynamics summarized by these particular classes of differential problems.

Chapter 10 concludes this thesis discussing the possible future prospective of our research.

2

PRELIMINARY TOOLS AND CONCEPTS

> *This chapter introduces some preliminary concepts that will be useful in our dissertation. The concepts of B-splines and penalized splines in one and two dimensions are introduces.*
>
> ***Keywords****: B-splines, tensor products of B-splines, P-splines, tensor products of P-splines.*

2.1   A short excursus on B-splines

Basis splines (or simply B-splines) represent one of the basic building block for our further discussion. Well known references about this topic are the work by de Boor, 1978 and Dierckx, 1995.

A B-spline is a piecewise polynomial function defined in a domain spanned by a set of points called "knots". The degree of the polynomial pieces defining the basis function determines its order. A B-spline of degree p is a piecewise polynomial function built using $p + 1$ polynomial pieces of degree p joined at p internal knots. The knots can be equally spaced or not. In this work we always deal with B-splines defined on a set of equidistant knots. Taking equidistant knots does not affect our further results and, in our opinion, is convenient in many applications, as shown by Eilers and Marx (2010). At joining points the derivatives up to the degree $p - 1$ of each polynomial piece are continuous. Each basis function takes nonzero values in the domain spanned by $p + 2$ internal knots and overlaps to $2p$ polynomial pieces of adjacent bases (except at the boundaries of the domain). At a given knot point $p + 1$ B-splines assume nonzero values. A set of 2th order B-splines is depicted in figure 2.1.

Collecting these piecewise polynomial functions it is possible to define a basis matrix B: each column of this matrix contains a B-spline. The localness
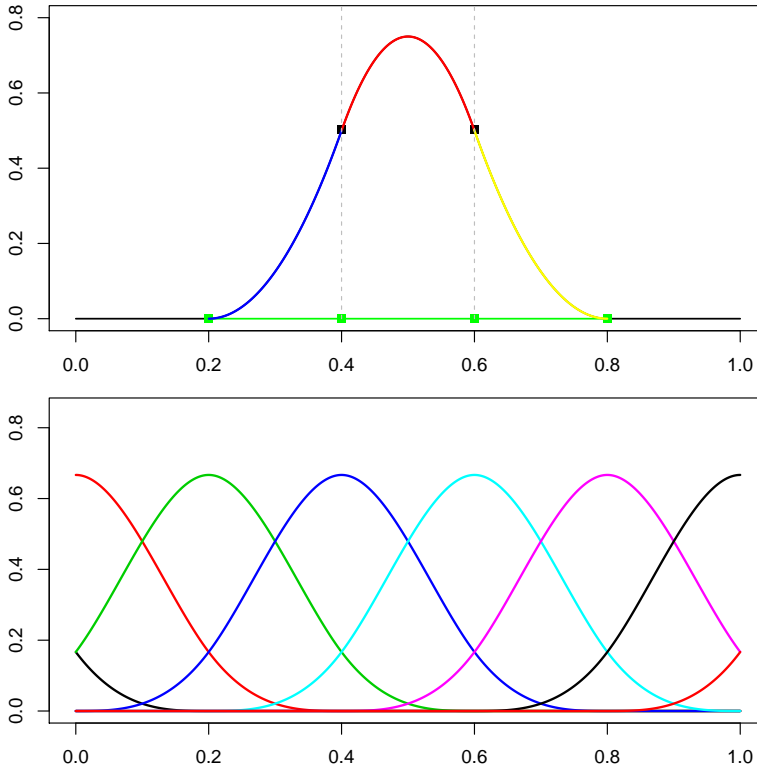
Figure 2.1: *One 3th order B-spline with its polynomial components. The second panel shown five 3th order B-splines.*

of the polynomials defining the bases makes the B matrix really sparse so that each of its row contains only few nonzero elements. This sparseness is an important feature making B-splines computationally convenient for function interpolation and approximation.

$$
B = \begin{bmatrix}
B_1(x_1) & B_2(x_1) & B_3(x_1) & ... & B_n(x_1) \\
B_1(x_2) & B_2(x_2) & B_3(x_2) & ... & B_n(x_2) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
B_1(x_n) & B_2(x_n) & B_3(x_n) & ... & B_n(x_n)
\end{bmatrix}.
$$

Several algorithms can be adopted to compute the basis spline functions forming the B matrix. A not convenient way is to evaluate the polynomial segments forming the basis functions analytically. A more practical approach is represented by the algorithm proposed by de Boor, 1978. The last alternative that can be mentioned consists in computing the spline functions using

differences between truncated power functions (Schumaker, 1981; Eilers and Marx, 2010).

The B-spline matrix B can be used to interpolate or approximate any unknown function. Suppose that we want to approximate the values y assumed by an unknown function $f(x)$ for some values of x. If we denote with $B_j(x, p)$ the value of the jth B-spline at point x we can represent y using a linear combination of B-splines $\widehat{y}(x) = \sum_j \widehat{c}_j B_j(x, p)$ where $\widehat{c}_j$ is the jth B-spline coefficient.

As shown by de Boor (1978), there is a convenient relationship between the dth derivative of a B-spline of order p and a B-spline of reduced order $p - d$. Indeed, if h is the distance between two adjacent knots, the following relation holds:

$$\widehat{y}^{(d)} = \sum_j c_j B_j^{(d)}(x, p) = \frac{\sum_j \Delta^{(d)} c_j B_j(x, p - d)}{h^d}, \qquad (2.1)$$

where $\Delta^{(d)} c_j$ is a dth order difference operator applied to the B-spline coefficients. So, if we define the dth order difference matrix $D^{(d)}$ the dth derivative basis matrix can be defined as:

$$B^{(d)}(x, p) = \frac{B(x, p - d)\, D^{(d)}}{h^d}.$$

A useful issue of B-splines is their direct generalizability to higher dimensions. Higher dimensional basis matrices can be computed using tensor products of one dimensional B-splines.

Consider a $\{x, y\}$ grid of values. If we define the B-splines B_x and B_y, the two-dimensional basis matrix is obtained as Kronecker product between them:

$$B_{xy} = B_x \otimes B_y.$$

This formulation is valid only for $\{x, y\}$ values defined on a regular grid. In the opposite case the two-dimensional basis matrix can be built using row-wise tensor products $B_{xy} = (B_x \otimes 1_L^T) \odot (1_K^T \otimes B_y)$.

In analogy with the unidimensional case it is possible to exploit a useful relationship to compute partial derivatives basis splines:

$$\frac{\partial^{p+q} B_{xy}}{\partial x^p y^q} = \sum_i \sum_j c_{ij} B_{i,x}^{(p)} B_{j,y}^{(q)}, \qquad (2.2)$$
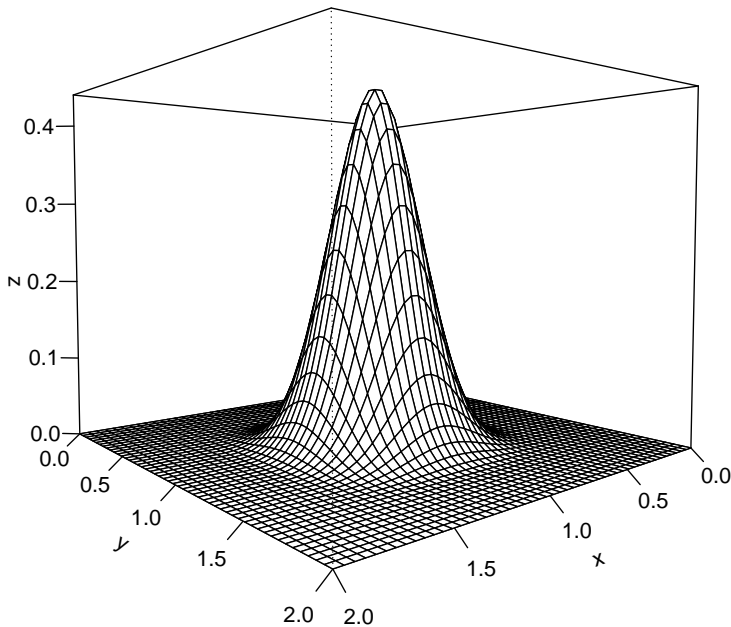
Figure 2.2: *A tensor product B Spline.*

where $B_x^{(p)}$ and $B_y^{(q)}$ represent the pth and qth order derivatives of B_x and B_y respectively. Figure 2.2 shows a single tensor product B-spline.

2.2   *Penalized splines*

P-splines have been introduced by Eilers and Marx (1996) as flexible smoothing procedures combining B-splines and difference penalties. 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 describe y through an appropriate smooth function. Denote $B_j(x; p)$ the value of the jth B-spline of degree p defined on a domain spanned by equidistant knots (in case of not equally spaced knots our reasoning can be generalized using divided differences). A curve that fits the data is given by $\widehat{y}(x) = \sum_{j=1}^{n} c_j B_j(x; q)$ where c_j (with $j = 1, ..., n$) are the estimated B-splines coefficients. Unfortunately this curve, obtained minimizing $\|y - Bc\|^2$ w.r.t.

c, shows more variation than is justified by the data if the number of spline functions is too large. To avoid this overfitting tendency it is possible to estimate c using a generous number of bases in a penalized regression framework:

$$\widehat{c} = \underset{c}{\operatorname{argmin}} \|y - Bc\|^2 + \lambda \|Dc\|^2, \tag{2.3}$$

where D is a dth order difference penalty matrix and λ is a smoothing parameter. Second or third order difference penalties are suitable in many applications. A second order difference matrix appears as follows:

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

The optimal spline coefficients follow from (2.3) as:

$$\widehat{c} = (B^T B + \lambda D^T D)^{-1} B^T y. \tag{2.4}$$

The smoothing parameter λ controls the trade-off between smoothness and goodness of fit. For $\lambda \to \infty$ the final estimates tend to be constant while for $\lambda \to 0$ the smoother tends to interpolate the observations. Figure 2.3 shows how different values of the smoothing parameter influence the estimated smoother (for brevity only four λ values are shown). The data were simulated by adding a Gaussian noise to a sine wave trend (200 observations). The B-spline matrix has 30 equidistant knots. Cubic B-splines and second order difference penalties have been used to estimate the smoothing functions.

A P-spline smoother has some desirable properties. One of them is the conservation of moments. If we define $v_{ik} = x_i^k$ with integer k, the inner product $y^T v_k$ defines the kth moment of y. A nice property of penalized splines built using a mth order difference penalty is that, for $0 < k < m$, it is true that $y^T v_k = \widehat{y}^T v_k$ for each value of λ.

P-splines are easily generalizable to analyze multidimensional data. Suppose to have a dataset represented by a triplet $\{x, y, Z\}$ where x and y indicate a grid of $\{x, y\}$ values defining the domain over which the matrix Z have been observed. Our aim is to efficiently summarize this 2D cloud of points. To reach this purpose we could use a tensor product of B-splines
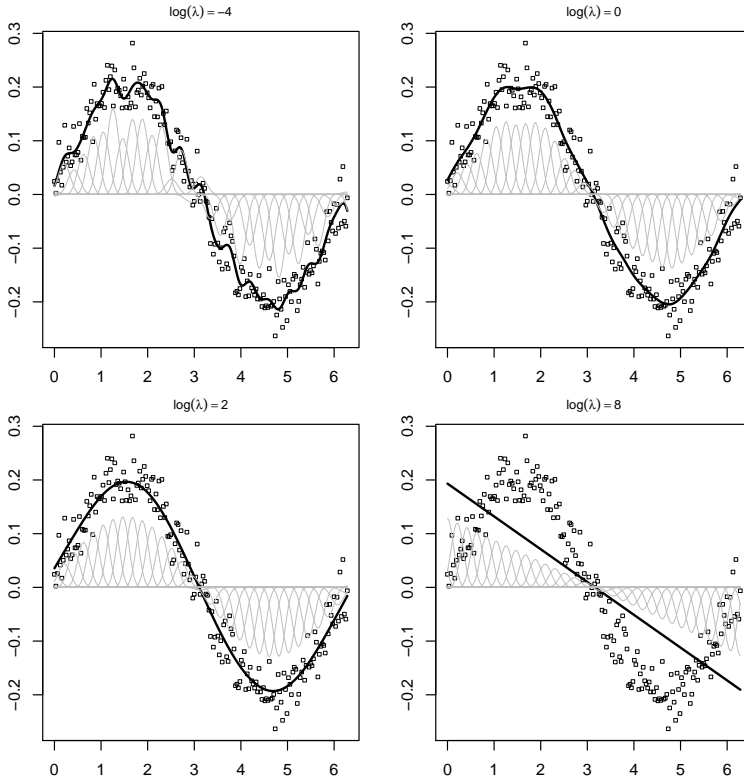
Figure 2.3: *Influence of the smoothing parameter on the P-spline smoother.*

defined on the x and y directions. This approach would be not appropriate giving a surface showing an unjustified roughness being influenced by local noise. To overcome this drawback we can add penalty terms to control for the smoothness of the estimated surface. Marx and Eilers (2005) suggest to introduce two difference penalties, one for each direction of the grid. The strength of the penalties defining the smoothness of the final estimates is tuned by a couple of independent λ parameters (non-isotropic smoothing). A simpler alternative is represented by isotropic smoothing considering $\lambda_1 = \lambda_2 = \lambda$. In case of non-isotropic smoothing the estimation problem becomes:

$$\underset{c}{\operatorname{argmin}} \|\operatorname{vec}(Z) - (B_x \otimes B_y)c\|^2 + \lambda_1 \|P_1 c\|^2 + \lambda_2 \|P_2 c\|^2, \qquad (2.5)$$

where B_x and B_y represent the B-splines built on the x and y direction respectively. For scattered observations the two-dimensional basis matrix can be built using row-wise tensor products as mentioned above $B_{xy} = (B_x \otimes 1_L^T) \odot (1_K^T \otimes B_y)$. In these cases the observations are represented by

three $\{x, y, z\}$ vectors of same length. The P_1 and P_2 penalty matrices are defined as follows:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes D_1^{(d)}$$

and

$$P_2 = D_2^{(d)} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where $D_1^{(d)}$ and $D_2^{(d)}$ are two dth order difference matrices. Let $B_{xy} = B_x \otimes B_y$ (or $B_{xy} = (B_x \otimes 1_L^T) \odot (1_K^T \otimes B_y)$ if we observe scattered observations) then the solution of the minimization problem in (2.5) is the coefficient matrix:

$$\widehat{c} = (B_{xy}^T B_{xy} + \lambda_1 P_1^T P_1 + \lambda_2 P_2^T P_2)^{-1} B^T \text{vec}(Z). \tag{2.6}$$

Figure2.4 shows a simulated non.isotropic two-dimensional P-spline smoothing example. The computational effort needed to perform the estimates described above becomes quickly prohibitive as the number of observations increases. An efficient algorithm (GLAM) has been proposed by Eilers et al. (2006). The cited algorithm have been used in every 2D applications discussed in this work.
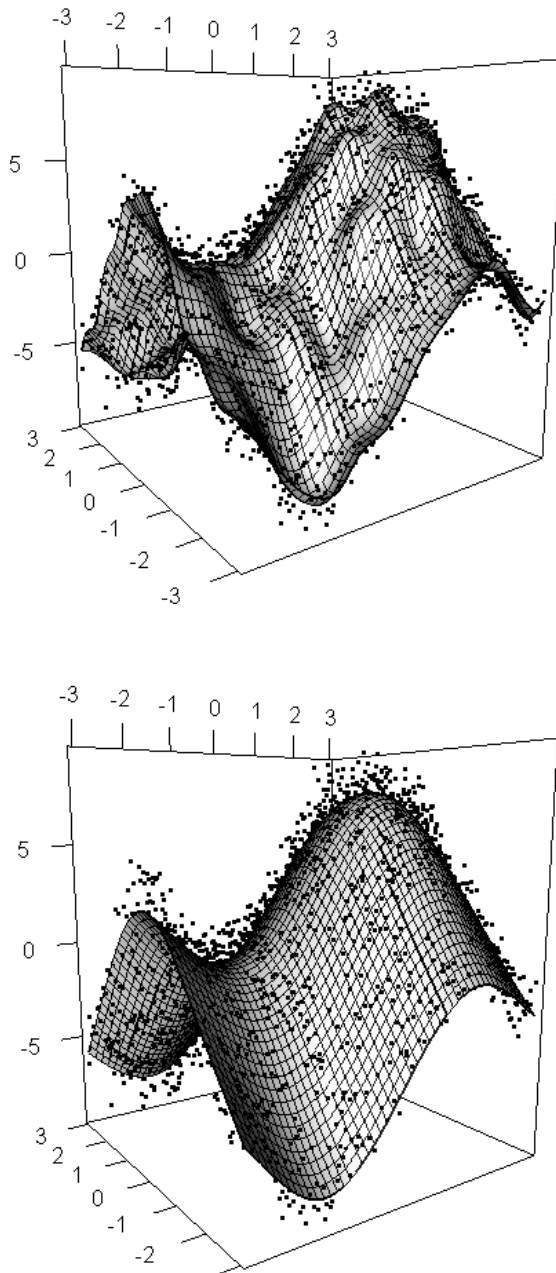
Figure 2.4: *Multidimensional smoothing of simulated data. The first panel shows the P-spline smoother obtained for $\lambda_1 = \lambda_2 = 0.01$. The second panel shows the smoother estimated using $\lambda_1 = 10$ and $\lambda_2 = 31.62$.*

3

The L-curve, adopted for the selection of the regularization parameter in ill-posed inverse problems, shows a parametric plot of the residuals vs the penalty. The corner of the L indicates the right amount of regularization. The L-curve is easy to compute and works surprisingly well also for smoothing data with correlated noise. We present the theoretical background, an alternative criterion to find the corner automatically, and applications to real data.

***Keywords**: Whittaker and P-spline smoothers, L-curve, cross validation.*

3.1 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 (2010).

P-splines (Eilers and Marx, 1996) 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 (Whittaker, 1922; Eilers, 2003), which became popular in the econometric literature as the Hodrick-Prescott filter (Hodrick and Prescott, 1997). 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. On the other hand sparse matrix algorithms allow fast computations.

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), generalized cross-validation (Wahba, 1990) 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 (Schall, 1991; Ruppert, 2003).

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 $\widehat{f}(x_k)$ is the estimated value taking into account the leave-out procedure y_k is an observation:

$$E\left[(\widehat{f}(x_k) - y_k)^2\right] = E\left[(\widehat{f}(x_k) - f(x_k))^2\right] + \sigma^2 - 2\text{Cov}[\widehat{f}(x_k), y_k],$$

which shows that the expected squared error for a cross validation term is equal to the true expected squared error plus the noise variance σ^2 minus two times the covariance between the observed data and the estimates. Even if this last term is not zero (as in the case of serial dependence in the noise component) the cross validation misspecifies it to be equal to zero. (see Carmack et al. (Carmack et al., 2012)). This leads to a smoothing function that tends to consider the correlated errors as a part of the wanted signal.

In this chapter we present and alternative approach, based on the L-curve method for ill-posed inverse problems (Hansen, 1992; Hansen and O'Leary, 1993; Hansen, 2000). The L-curve is a plot of the logarithm of the magnitude of the penalty term against the log of the sums of squares of the residuals, parameterized by the regularization parameter λ. In the case of inverse problems a very pronounced L-shape is obtained and a good value of the regularization parameter is found in its corner. 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 the smoothing parameter 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 as already noticed by Hansen for Tikhonov regularization applications. This is illustrated in figure 3.1, showing historical data of the price of orange juice (Stock and Watson, 2003), with very strong serial correlation. For the upper panel the amount of smoothing was chosen automatically by LOO-CV, while for the lower panel the L-curve was used. It is quite clear that no meaningful trend can be obtained with LOO-CV; GCV, 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 but it is significant that the L-curve that does not rely on any statistical model. This method shows excellent performance in practice, so we like to share our experiences with the statistical community.

Krivobokova and Kauermann (2007) approach smoothing as a mixed model with correlated noise and present a REML estimation algorithm. We compare their results with those obtained using the L-curve and find that the latter performs better and faster (especially when complex covariance structure are needed to describe the data). It is necessary to clarify that the evaluations about the computational efficiency of the methods discussed are obtained considering an high-level programming language such as R and using its stand alone functions.

This chapter is organized as follows: in section 3.2 we introduce the smoothing procedures that will be used in our discussion, together with some standard smoothing selection procedures, section 3.3 describes in more
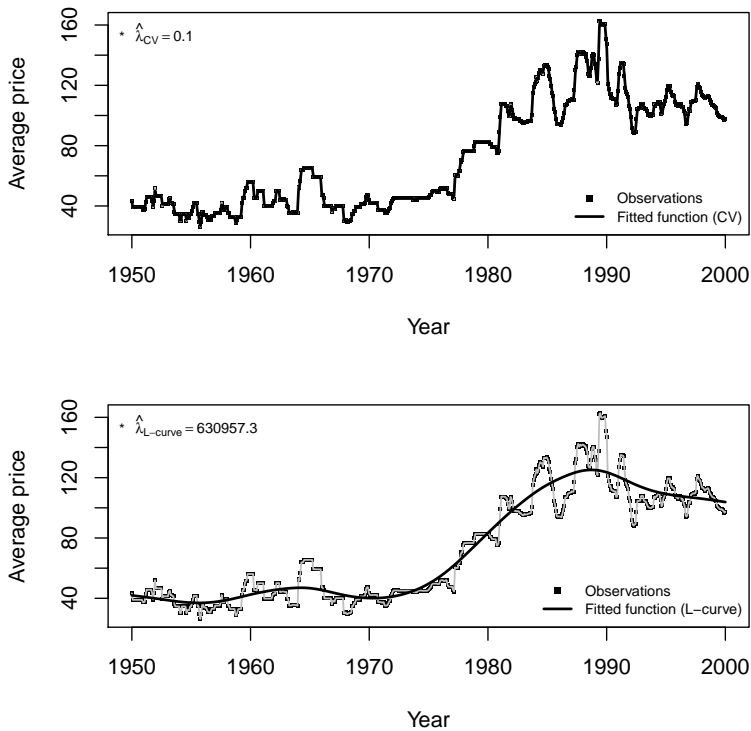
Figure 3.1: *Whittaker smoother applied to the orange juice dataset (see section 3.5). The smoothing parameter have been selected by cross validation (upper panel) and by L-curve procedure (lower panel). In both cases the smoothing parameter was considered in the range* $\log_{10}(\lambda) \in [-4, 6]$.

details the L-curve method, in section 3.4 we analyze the shape of the L-curve through a simulation study verifying how it is influenced by the characteristics of the data, and finally, section 3.5 compares the L-curve with LOO-CV and REML-based approaches using real data.

3.2 *P-splines, Whittaker smoother and Hodrick-Prescott filter*

P-splines were proposed by Eilers and Marx (1996). 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 Eilers and Marx, 2010). A curve that fits the data 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 phenomenon it is possible to estimate $\widehat{z}$ using a penalized linear regression approach:

$$\widehat{z} = \underset{z}{\operatorname{argmin}} \|y - Bz\|^2 + \lambda \|Dz\|^2, \tag{3.1}$$

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 (3.1) for the spline coefficients we get:

$$\widehat{z} = (B^T B + \lambda D^T D)^{-1} B^T y. \tag{3.2}$$

The Whittaker smoother (Whittaker, 1922) can be viewed as a special case of the P-spline smoother. It arises when the observations are located on an equally-spaced grid, a knot is placed at each abscissa point, and $B = I$, the identity matrix. In the econometric literature this smoother is also known as the Hodrick-Prescott filter (Hodrick and Prescott, 1997). 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 a $\lambda = 1600$ as a good choice. Eilers (2003) uses a leave-one-out cress-validation. Kauermann et al. (2011), work in the other direction, replacing the H-P filter with a penalized spline smoother.

Popular methods for smoothing parameter selection are: the Akaike Information Criterion, Cross Validation. AIC estimates the predictive log likelihood, by correcting the log likelihood of the fitted model (Λ) by its effective dimension (ED): $AIC = 2ED - 2\Lambda$. Following Hastie and Tibshirani (1990) we can compute the effective dimension as $ED = \operatorname{tr}[(B^T B + \lambda D^T D)^{-1} B^T B]$ for the P-spline smoother while $ED = \operatorname{tr}[(I + \lambda D^T D)^{-1}]$ is the effective dimension of the Whittaker smoother, and

$$2\ell = -2n \ln \hat{\sigma}^2 \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}_0^2},$$

where $\hat{\sigma}$ is the maximum likelihood estimate of σ. But $\hat{\sigma}^2 = \sum_i (y_i - \hat{y}_i^2)^2/n$, so the second term of ℓ is a constant. Hence the AIC can be written as:

$$AIC(\lambda) = 2ED + 2n \ln \hat{\sigma}. \tag{3.3}$$

The optimal parameter is the one that minimizes the value of $AIC(\lambda)$. LOO-CV chooses the value of λ that minimizes:

$$CV(\lambda) = \sum_{i=1}^{n} \left[\frac{y_i - \widehat{y}_i}{1 - h_{ii}} \right]^2, \tag{3.4}$$

where h_{ii} is the ith diagonal entry of $H = B(B^T B + \lambda D^T D)^{-1} B^T$ for P-splines or $H = (I + \lambda D^T D)^{-1}$ in case of the Whittaker smoother. Analogous to CV is the generalized cross validation measure (Wahba, 1990):

$$GCV(\lambda) = \sum_{i=1}^{n} \left[\frac{y_i - \widehat{y}_i}{n - ED} \right]^2, \tag{3.5}$$

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

Related to this last method is the Generalize Correlated Cross Validation procedure proposed by Carmack et al. This method exploits the possibility to modify the definition of the degrees of freedom to be taken into account in computing the GCV including the (estimated or a priori known) correlation structure in the noise component. If we denote with C this correlation matrix and with S the smoothing matrix this modified GCV approach can be computed as follows:

$$GCCV(\lambda) = \sum_{i=1}^{n} \frac{1}{n} \left[\frac{y_i - \widehat{y}_i}{1 - \text{tr}[2SC - SCS^T]} \right]^2. \tag{3.6}$$

3.3   *L-curve selection method*

The L-curve is a parameterized curve showing the two ingredients of every regularization or smoothing procedure: the goodness of fit and the roughness of the final estimate. This approach was originally proposed by Hansen (1992) for the selection of the regularization parameter in ill-posed inverse problems. Regularization arises both in statistics and inverse problems applications even if the aim of the latter is slightly different. Indeed, in statistical modeling one posits a true data generating model and aims to estimate

it, in inverse problems one simply look for a 'good' approximate solution of a given equation without taking into account any data generating process. Also the errors are considered differently. An important parameter for the solution of inverse problems is the error upper bound level while in statistical applications the error it treated as an exogenous component (see Vogel, 1996).

In this section we study its applicability to statistical models and in particular we focus on smoothing applications.

3.3.1 *The L-curve in ridge regression*

Ridge regression is a common regularization tool in statistics. It adds a penalty term to the standard regression problem to shrink the coefficients:

$$\underset{\beta}{\text{argmin}} \, \|y - X\beta\|^2 + \lambda \|\beta\|^2$$
$$\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T y. \tag{3.7}$$

The strength of the shrinkage depends on λ. Define:

$$\{\omega(\lambda); \theta(\lambda)\} = \{\|y - X\beta\|^2; \|\beta\|^2\} \text{ and } \{\psi(\lambda); \phi(\lambda)\} = \{\log(\omega); \log(\theta)\}.$$

The L-curve is a plot of $\phi(\lambda)$ vs. $\psi(\lambda)$, parameterized by λ. the name.

Figure 3.2 shows a toy example. A sample of 200 realizations of 50 explanatory variables was drawn from a multivariate normal distribution with mean vector $\mu_i = 1$ with high correlation ($\rho \in \{0.7, 0.8, 0.9, 0.99\}$). The dependent variable was obtained as a linear combination of the 50 independent variables plus a random noise $y_i = cX + N(0, 0.2)$ (where c_j is the regression parameter associated to each independent variable randomly drawn from a uniform distribution). The shape of the curve is true to its name. It shows a corner in a region characterized by intermediate values of ψ, ϕ and λ. Hansen suggested to select the regularization parameter that corresponds to the corner, the point of maximum curvature. The curvature can be computed using:

$$k(\lambda) = \frac{\psi' \phi'' - \psi'' \phi'}{[(\psi')^2 + (\phi')^2]^{3/2}}. \tag{3.8}$$

The maximization of $k(\lambda)$ requires the computation of the first and second derivatives but the computations can be simplified in some cases as will be
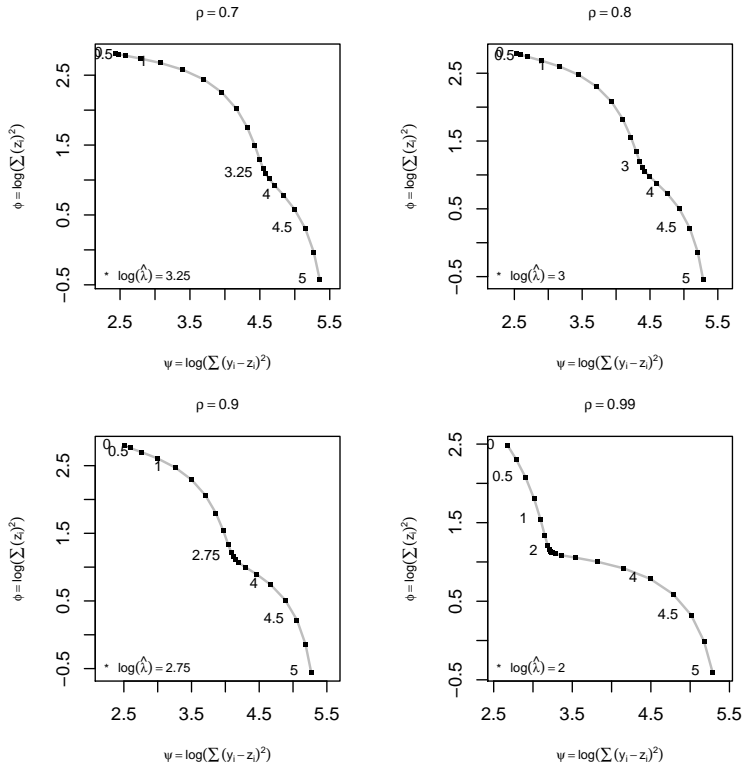
Figure 3.2: *Four L-curves obtained for simulated ridge regression examples with different degrees of correlation between the explanatory variables. The corner point represents the point of maximum curvature while the other points represent the points used to draw the curve. For some points the associated logarithmic value of the parameter is shown ($\log_{10}(\lambda) \in [0, 5]$).*

shown in what follows. It is important to notice that that while methods such as cross validation optimize the smoothing parameter minimizing a measure of the prevision accuracy, the L-curve suggests to look for the λ parameter that guarantees an optimal compromise between the residual sum of squares and the amount of penalty.

The L-curve is usually characterized by two components: a vertical and an horizontal part (consider the fourth panel of figure 3.2). The vertical part is drawn for small values of the regularization parameter. In this region the residual sum of squares is small while the amount of penalty tends to be high. The second component is the horizontal one: it is drawn for increasing values of the smoothing parameter leading to higher residual sum of squares. This

flat part of the curve is characterized by a rather constant amount of penalty. If we think about an ideal L plot we can say that for a constant amount of penalty we could tune the residual sum of squares along the horizontal part of the plot. It is an ideal situation but it holds till we reach the corner of the L. At the corner we obtain the best goodness of fit for the smallest possible price (smallest penalty).

3.3.2 The L-curve in penalized smoothing

In section 3.2 we briefly introduced the P-spline and the Whittaker smoothers. Consider a P-spline smoother. Taking $z = B\alpha$, the following quantities can be defined:

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

and the L-curve is given by:

$$L = \{\psi(\lambda); \phi(\lambda)\} = \{\log(\omega); \log(\theta)\}. \tag{3.9}$$

The L-curve for a simple smoothing problem is depicted in figure 3.4. These results were obtained from 200 observations simulated using the following scheme: $y = 5\sin(x) + N(0, 0.5)$ where $x \in [0, 2\pi]$.

The lower panel of figure 3.4 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 3.2 or the second panel of figure 3.4). 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.

3.3.3 A new selection criterion

The curvature function can be computed using (3.8) per each λ parameter on a given grid. If we set $\ell = \log(\lambda)$, the rate of change of the arc length distance between each point on the curve w.r.t. ℓ is given by:

$$\frac{ds}{d\ell} = \sqrt{\left(\frac{d\psi}{d\ell}\right)^2 + \left(\frac{d\phi}{d\ell}\right)^2}. \tag{3.10}$$

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). This means that we can simplify the selection procedure. The corner, if it exists, coincides (at least approximately) with the point satisfying:

$$\min\left\{\sqrt{(\Delta\psi)^2 + (\Delta\phi)^2}\right\}. \tag{3.11}$$

The criterion in (3.11) suggests that the best smoothing parameter can be selected minimizing the Euclidean distance between adjacent points on the L-curve. The Euclidean distance between these points describes a U-shaped as shown in the last panel of figure 3.4. Shahrak (2013) already coined the name U-curve, although with a different definition. To avoid confusion, we christened our procedure the "V-curve".

It is possible to use a more precise criterion based on an algorithm to compute the gradient $\nabla_{\ell_i}(L)$ on a really fine grid of smoothing parameter. Indeed, using the well known equation for the computation of the B-splines derivatives (de Boor, 1978), we can look for the λ satisfying (3.11) on a fine grid of candidate parameters. The optimal smoothing parameter is then the λ on the this grid that minimizes the magnitude of the gradient.

First of all we can approximate the functions ψ and ϕ using basis functions. If we define the matrix $L = [\psi\ \phi]^T$, we can estimate a matrix of suitable B-spline coefficients as follows:

$$\widehat{C} = \left(B^T B + k D^T D\right)^{-1} B^T L, \tag{3.12}$$

where, in analogy with what we already discussed in the previous sections, B is a B-spline basis matrix, D is a difference penalty matrix and we introduce a really small regularization parameter k (for instance $k = 10^{-6}$) to ensure the well-poseness of the problem. Given the B-spline coefficients matrix $\widehat{C}$ it is possible to estimate the derivatives of ψ and ϕ w.r.t. ℓ using a derivative basis matrix $\dot{B}(\ell, p)$. This matrix can be obtained not only considering the original grid but also on a finer one. Indeed, in order to obtain $\dot{B}(\ell, p)$, we need to set only the extreme values for ℓ being possible to make the grid as fine as we prefer.

The final estimates of the derivatives are the matrix $G = \dot{B}\widehat{C}$. The rows of this matrix represent the gradient of the L-curve for each value of ℓ: $G =$

Wood surface: squared gradient components

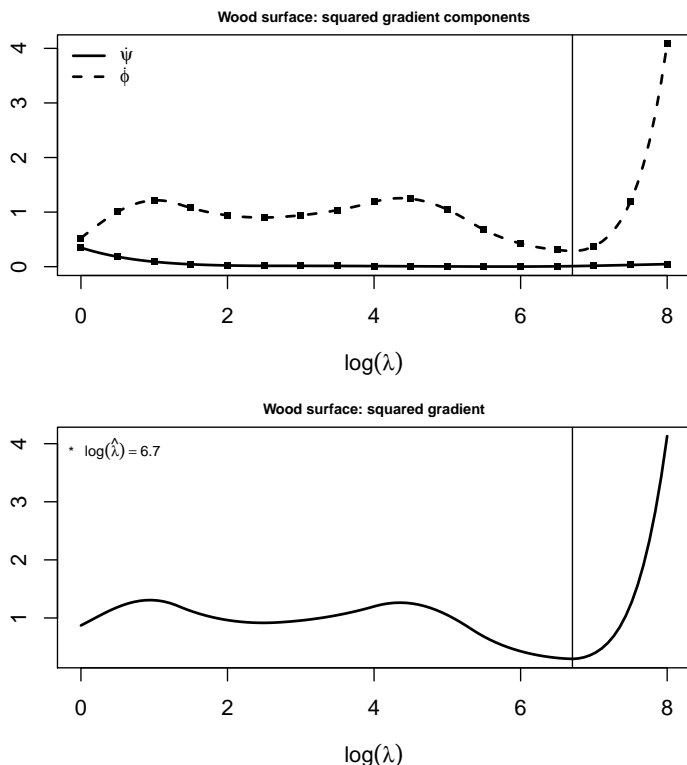**Wood surface: squared gradient**

Figure 3.3: *B-spline based gradient computation for the wood data (see first example of section 3.5.*

$\nabla_{\ell_i}(L)$. The optimal smoothing parameter, the ℓ that minimizes (3.10), is the one corresponding to the gradient vector with minimum length.

Figure 3.3 shows the results obtained using this procedure for the wood data (see first example section 3.5). The upper panel shows the squared components of the gradient vectors for each ℓ value in the normal grid (dot points) and considering the finer grid (lines). The lower panel shows the distance function and the optimal smoother parameter selected using the finer grid of smoothing parameter.

3.4 The shape of the L-curve

Using simulated data we will now show how the characteristics of the data impact on the shape of the L-curve. Let us start by considering figure 3.6. The first panel shows the L-curve for a Whittaker smoother applied to data simulated using the following scheme: $y = 10^{c_j} \sin(x_i) + N(0, 1)$ with $x_i = 1, ..., 2\pi$
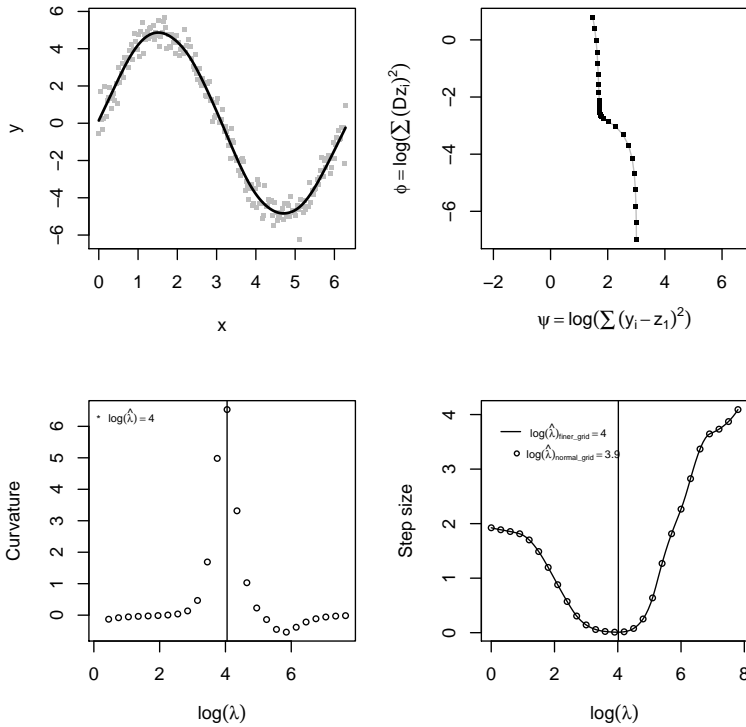
Figure 3.4: *P-spline smoothing of simulated data using an L-curve approach. The first panel shows the obtained smoothing function (line) and the second the associated L-curve. The lower panels show the curvature function and Euclidean distance between adjacent points of L-curve. The values of these functions are plotted against different values of $\log_{10}(\lambda)$.*

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 3.5 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 3.7 reproduces the distributions of residual mean square errors computed comparing the fitted values and the underlying trend component considering data with random noise showing different standard deviations (we considered now 1000 simulated observations). As we can notice the mean deviations for each variability level is close to zero even if those with a less variable noise are smaller.

The L-curve turns out to be particularly useful when smoothing data with autocorrelated noise. However the shape of the curve depends on the strength of the serial correlation of the error component. Figure 3.8 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. Slightly 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 3.8 shows some 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 3.9 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

Figure 3.5: *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 $\log_{10}(\lambda) \in [-2, 9]$. The trend component was obtained considering $x \in [0, 2\pi]$ and $y_0 = \sin(x)$ for $i = 1, ..., 200$.*
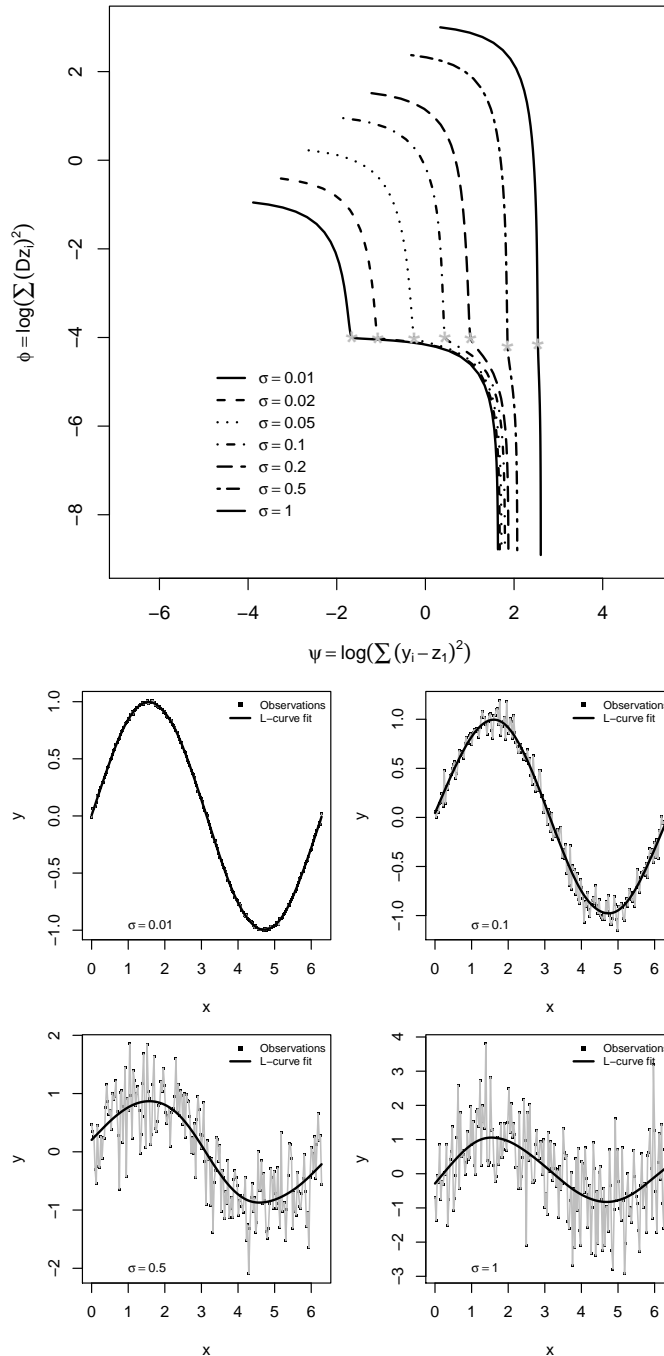
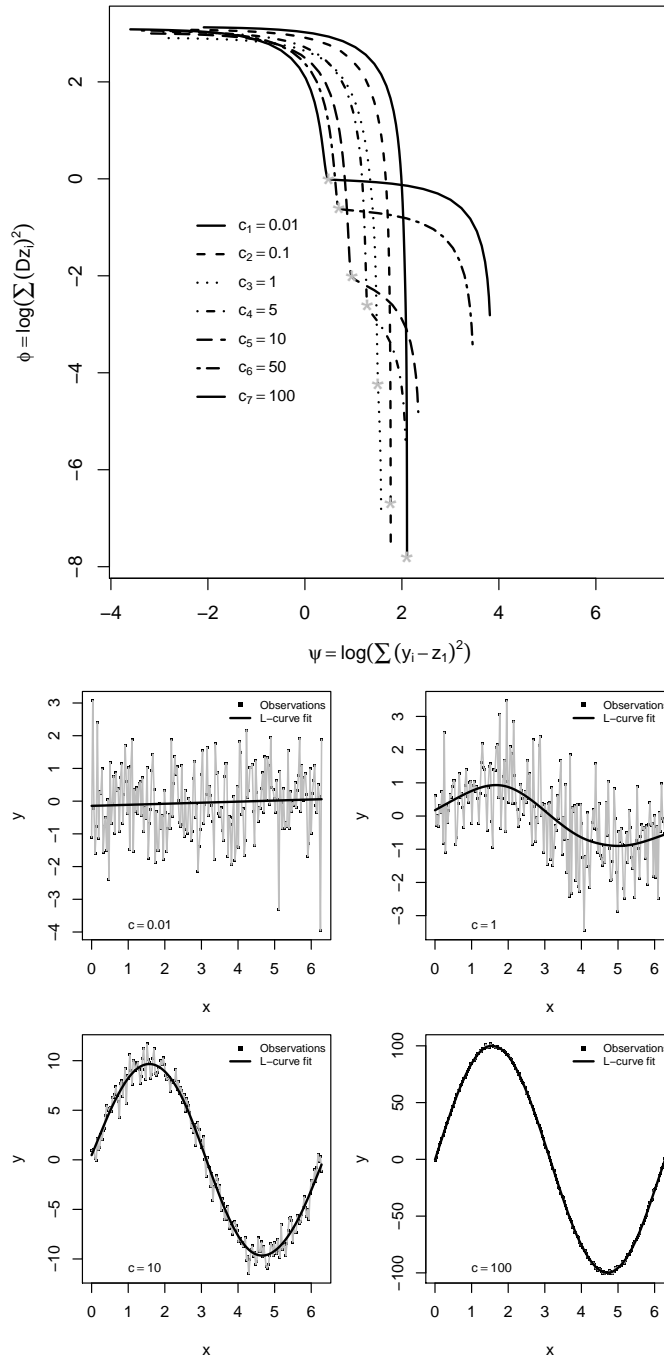**Figure 3.6:** *The first panel shows seven L-curves obtained for a Whittaker smoother estimated on data with different weights for the signal component while, the lower panels show four smoothers obtained considering data simulated using different weights for the this component (indicated in the lower legend of each plot). All these results were obtained considering* $\log_{10}(\lambda) \in [0, 9]$.
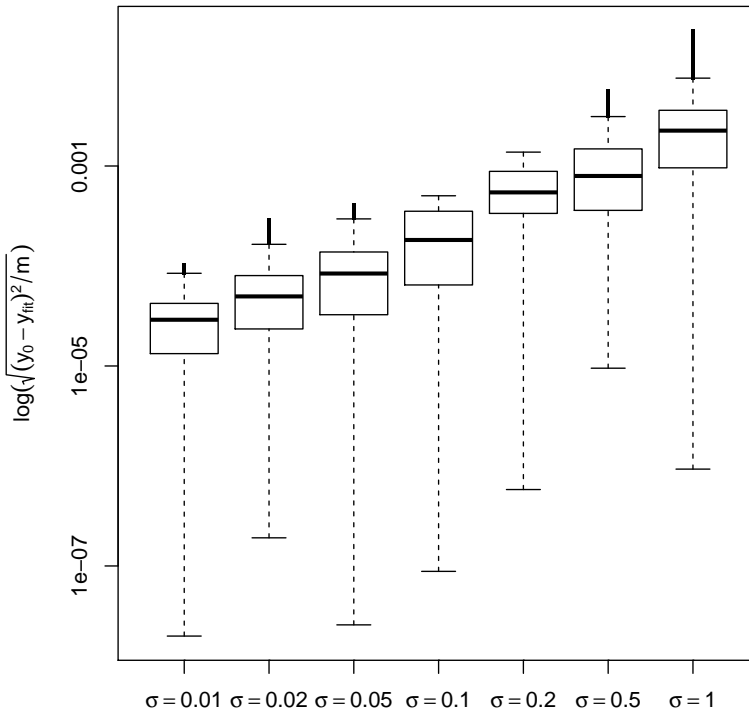
Figure 3.7: *Distributions of RMSE computed on the estimated values and the underlying trend considering data with noise components characterized by different degrees of variability. These results were obtained considering $\log_{10}(\lambda) \in [-2, 9]$ and simulating 1000 observations.*

the L-curve seems to be less sharp in smoothing spline regression than in the applications proposed in the literature (typical 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.

3.5   *Applications*

In this section we would like to test the performances of the L-curve criterion using real datasets coming from different scientific fields. We compare here the L-curve procedure with a cross validation approach and with the criterion proposed by Krivobokova and Kauermann (2007). For this last procedure we exploit what is shown in the Appendix B of the cited paper based on the use
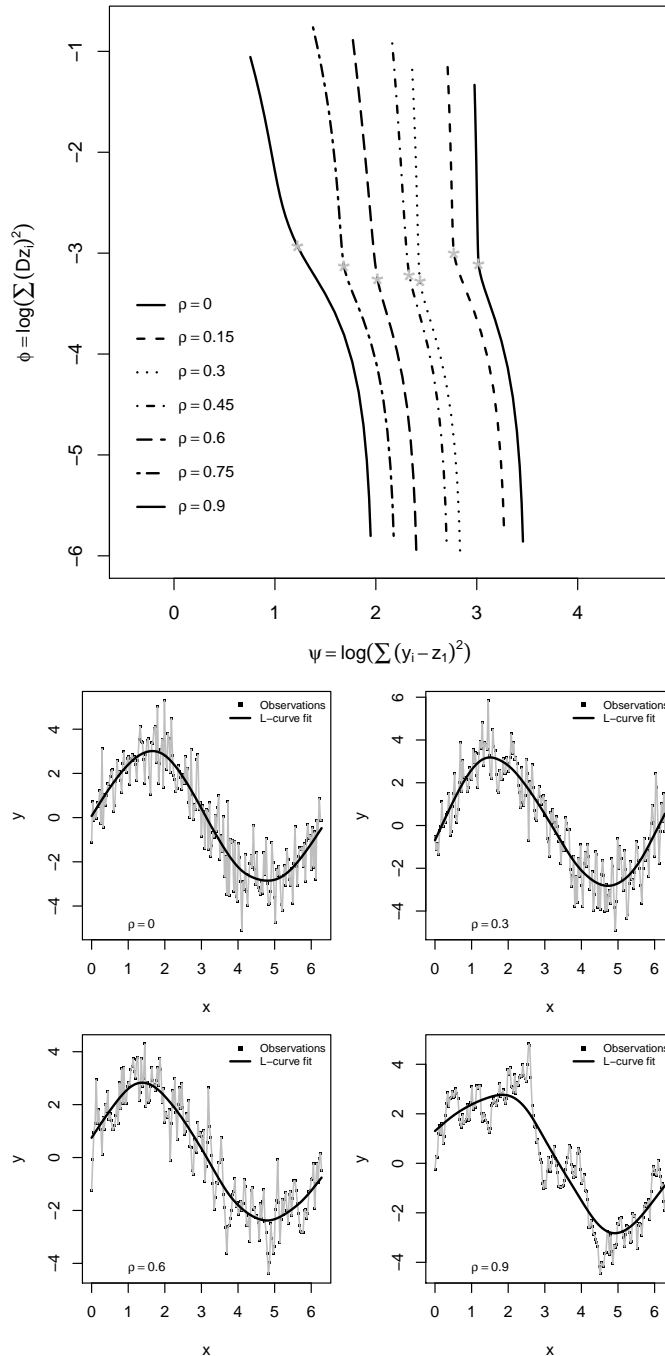
Figure 3.8: *The first panel shows seven L-curves for a Whittaker smoother estimated on data with an increasing autocorrelation coefficient for the noise component while, the lower panels show the smoothing functions associated to four L-curves depicted in the upper panel. All these results were obtained considering $\log_{10}(\lambda) \in [-2, 9]$.*
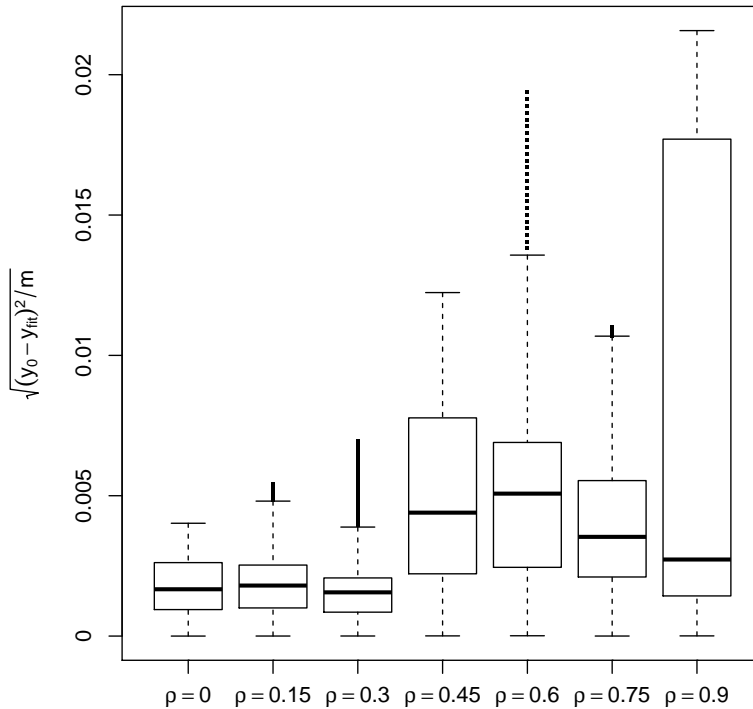
Figure 3.9: *Distributions of the RMSE between estimated values and underlying function for different autocorrelation coefficients for the noise component of the data. These results were obtained considering $\log_{10}(\lambda) \in [-2, 9]$ and simulating 1000 observations.*

of the R function `nlme`. This criterion is consistent with the work of Currie and Durban (Currie and Durban, 2002). The first application that we would like to show concerns the smoothing of the wood data. This dataset was originally proposed by Pandit and Wu (Pandit and Wu, 1993). It describes 320 measurements of a block of wood that was subject to grinding. In figure 3.10 the profile height at different distances is drawn. The profile variation follows a curve determined by the radius of the grinding stone. We can 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 3.11. We also compare the Whittaker smoother obtained using the L-curve and the cross validation selection criterion with the filter suggested by Krivobokova
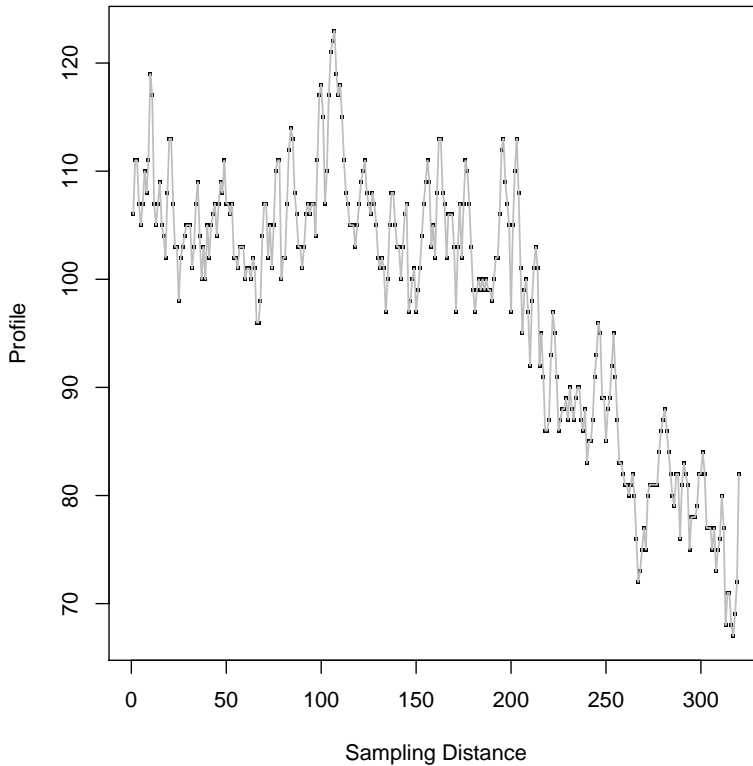
Figure 3.10: *Profile of a block of wood subject to grinding.*

and Kauerman considering both $AR(1)$ and $AR(2)$ correlation structures. In particular, we use piecewise linear basis function to build the smoother defined on 31 knots. 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. On the other hand the third proposed filter (we will call it K&K filter for brevity up to now) also guarantees satisfactory results. As second 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 the web repository `http://www.psmsl.org/data` and are summarized in figure 3.12.

As in the previous cases the cross validation procedure gives a rough smoothing function while the Whittaker smoother obtained tuning the smoothing parameter with the L-curve catches the trend in the data. On
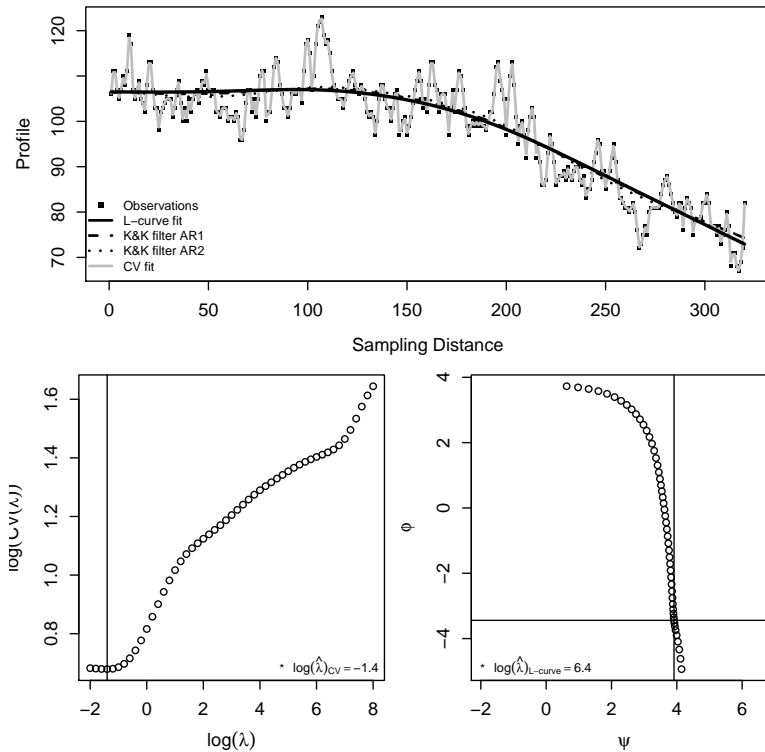
Figure 3.11: *Whittaker smoothing of wood data. The upper panel shows the results obtained selecting the parameter by the L-curve, the cross validation and using a K&K filter defined using piecewise linear basis functions on 31 knots. The lower panels represent the L-curve and the cross validation functions and the selected parameters. These results were obtained considering $\log_{10}(\lambda) \in [-2, 8]$.*

the other hand the results obtained considering the strategy suggested by Krivobokova and Kauerman seems to efficiently reproduce the data behavior using both $AR(1)$ and $AR(2)$ correlation structures. 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 P-spline smoothers. Welham and Thompson (2009) showed the possibility of bimodality in the the smoothing parameter log likelihood profile. Figure 3.13 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
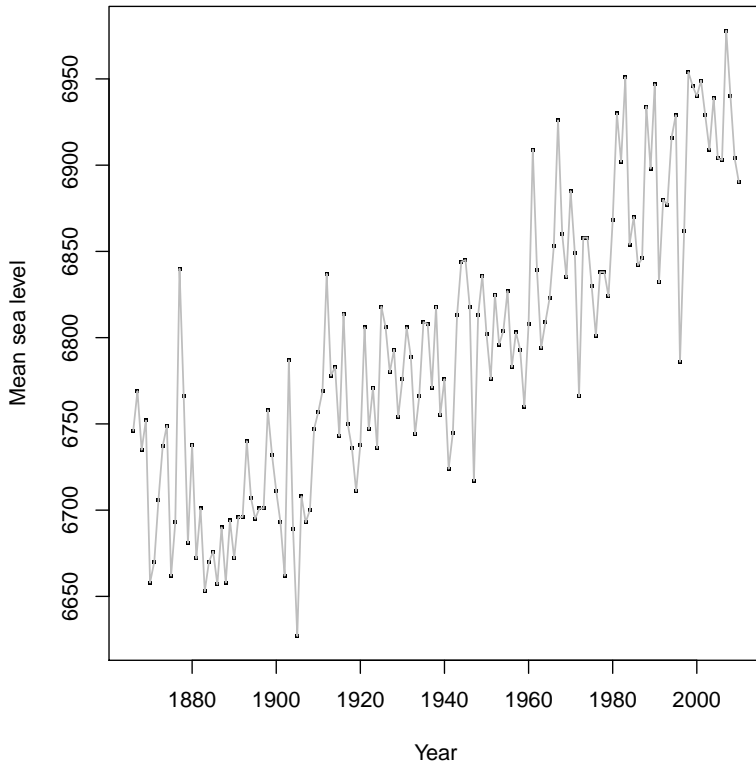
Figure 3.12: *Annual mean sea level registered in Delfzijl for the period 1865-2010.*

cross validation criterion. The last example that we would like to discuss is the orange juice price data already introduced in section 3.1. 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. From figure 3.1 it appeared clear that this dataset represents a really hard smoothing exercise. We again compare the L-curve procedure with a cross validation approach and with the criterion proposed by Krivobokova and Kauerman. Following the cited paper we use piecewise linear basis functions defined on 63 knots to build the smoother. In figure 3.14 the results obtained smoothing the orange juice prices time series with a Whittaker smoother and the K&K procedure are shown together
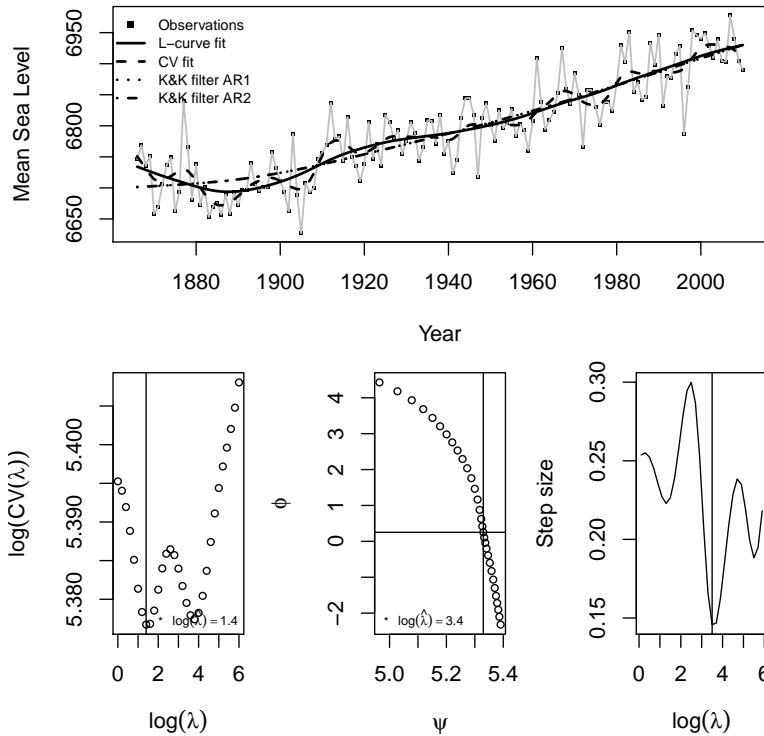
Figure 3.13: *Whittaker smoothing and K&K filter of mean sea level data. The upper panel show the results obtained selecting the parameter using the L-curve and the cross validation ($\log_{10}(\lambda) \in [0, 9]$) and the smoothing functions obtained using K&K filters considering an $AR(1)$ and $AR(2)$ correlation structures. The lower panels reproduce the L-curve, the cross validation curves and the selected parameters.*

with and the cross validation and L-curve profiles. The smoothing parameters selected by these three procedures lead to very different smoothing functions. The 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 Whittaker smoother is able to reproduce the trend behind the data. The results obtained using the filter proposed by Krivobokova and Kauerman are not satisfactory suggesting a smoother that completely miss the data behavior. These last results were obtained considering an $AR(1)$ correlation structure while, following the computational strategy suggested in the cited paper, considering an $AR(2)$ correlation structure the R model crashes. In order to compare the discussed smoothing parameter selection
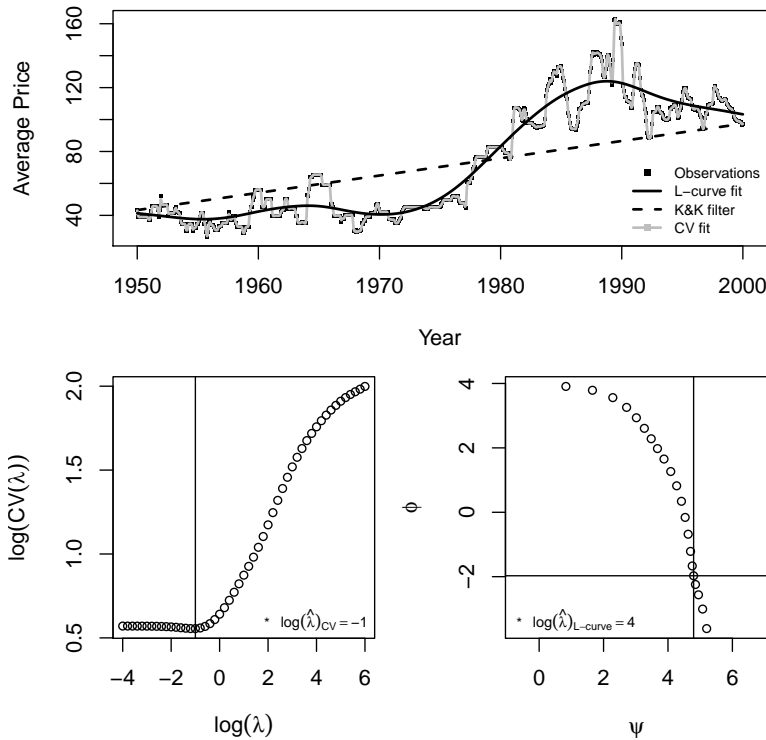
Figure 3.14: *The upper panel compare the results obtained selecting the λ parameter of a Whittaker smoother using the cross validation, the REML approach (K&K filter) and the L-curve method. The lower panels show the cross validation and L-curve profiles and indicate the selected smoothing parameters. For both selection methods we considered $\log_{10}(\lambda) \in [-4, 6]$.*

	Wood data (320)			Mean sea level (145)			Orange juice (642)		
L-curve	0.640	0.020	0.670	0.160	0.002	0.150	2.840	0.560	3.110
Cross Validation	1.140	0.110	1.300	0.310	0.010	0.310	6.420	1.280	7.050
K&K AR(1)	0.800	0.010	0.810	0.210	0.002	0.210	3.010	0.510	5.510
K&K AR(2)	7.800	0.010	7.630	1.120	0.020	1.240	n.a.	n.a.	n.a.

Table 3.1: *Computational times related to each selection procedure used for real data examples. The user, system and elapsed times are reported for each procedure and example. The number of observations per each dataset is indicated within brackets.*

procedures in terms of computational efficiency table 3.1 shows the computing times for the real data examples discussed here. The measurements have been obtained using the R function `proc.time()`.

3.6   Conclusions

In this chapter we introduced an L-curve procedure for the selection of the smoothing parameter in a P-splines framework. This approach selects the smoothing parameter through a direct comparison between the goodness of fit and the smoothness of the estimates. It was showed that the optimal smoothing parameter can be selected by locating the point of maximum curvature, i.e. the corner, of the L-curve as was originally argued in the works of Hansen (Hansen, 1992; Hansen and O'Leary, 1993). We also proposed an alternative selection procedure based on the minimization of the Euclidean distance between the adjacent points of the curve: the V-curve criterion.

Using several simulation strategies it has been explained how the shape of the L-curve is influenced by the underlying characteristics of the data and how the selection procedure performs.

A comparison between the L-curve and the cross validation approaches was examined using real data. It was highlighted that, selecting the smoothing parameter with a LOO-CV strategy, can suggest a too small λ parameter leading an under-smoothing of the data. This is true especially in the case of data with a noise component showing serial correlation. In such cases, on the other hand, the L-curve procedure was found to be a robust selection procedure.

The same comparison was made to the strategy proposed by Krivobokova and Kauerman. It was found that the L-curve ensures results consistent with those obtained using the K&K filter in two well behaved examples over three. On the other hand it fails in a really extreme cases (i.e. the orange juice example). We also found that the L-curve procedure guarantees good results and a relatively lower computational effort in the analysis of long data series.

Other REML-based approaches have been proposed in the literature. Currie and Durban (2002) also 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 dataset modeling the correlation structure through a $AR(2)$ model showing results really close to what we obtained in section 3.5. However we believe that our approach, besides being computationally more convenient, guarantees good performances also in extreme

situations in which a suitable correlation structure is hard to define (such as the orange juice price data).

From our discussion in this chapter appears also that the L-curve is a computationally efficient selection procedure if compared with standard procedures. Indeed it does not require the computation of the effective dimension of the smoother at each step. In addition in some cases it is also possible to simplify the selection procedure minimizing the Euclidean distance between adjacent points of the L-curve. This allows to avoid the computation of the first and second derivatives that appear in the curvature function. These factors make this procedure computationally faster than those based on a leave-out or a mixed model framework. This is not a secondary feature especially in those cases in which the amount of data tends to increase the computational cost of the entire statistical analysis.

It is necessary to underline that in some cases, the L-curve criterion could not lead to reliable results. It happens smoothing data that approach to a pure white noise or when the signal component under the data tends to disappear. As shown in figure 3.15 and in the discussion of section 3.5 in these cases the L-curve does not show a convex area and selecting the smoothing parameter through the minimization of the Euclidean distances leads to light smoothing parameters. In these cases the curvature approach has to be preferred because it tends to suggest a λ as close as possible to the optimal one. For this reason we recommend to use this procedure after a preliminary (also visual) inspection of the data. The L-curve procedure offers margins 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 the L-curve can also be generalized. Our future research will concentrate on an L-curve criterion for multivariate smoothing analyzes and on an L-curve based procedure suitable for expectile smoothing problems.
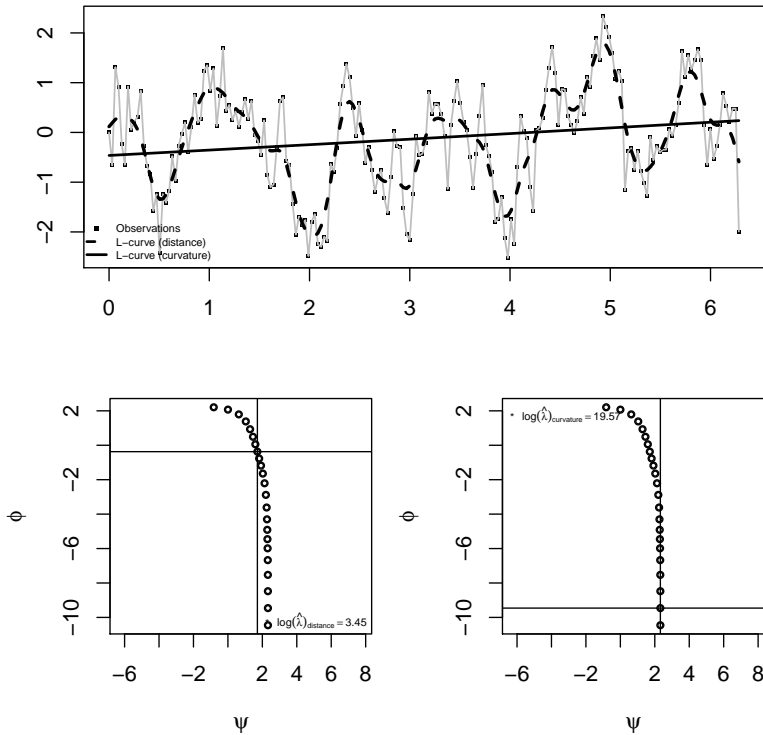
Figure 3.15: *Whittaker smoothing of a white noise. The upper panel shows the results obtained selecting the parameter using the L-curve and the cross validation ($\log_{10}(\lambda) \in [0, 9]$). The lower panels reproduce the L-curve selections based on the curvature and on the Euclidean distance criteria.*

4 THE L-SURFACE FOR NON-ISOTROPIC OPTIMAL SMOOTHING

The L-surface is an attractive two-dimensional generalization of the L-curve procedure for optimal smoothing. However, one has to locate the point of maximal curvature of the surface, which is not easy. We introduce a simplified procedure that replaces maximum curvature by a distance to be minimized.

Keywords: *P-splines, L-curve, tensor products, penalties*

4.1 Introduction

The L-curve criterion was proposed by Hansen (1992) and Hansen and O'Leary (1993) for selection of the regularization parameter in ill-conditioned inverse problems. In chapter 2 we have shown that it is an efficient and robust method to select the penalty parameter in smoothing applications. The L-curve does not require the computation of the effective model dimension and is insensitive to correlated noise.

Here we discuss a generalization, the L-surface, for non-isotropic two-dimensional smoothing. It is a plot of the logarithm of the residual sum of squares against the logarithms of the sizes of the penalty terms, parameterized by the smoothing parameters. It appears as a surface with a nook and, in analogy with the L-curve, the optimal pair of smoothing parameters is located in the deepest point of the nook. This is the point of maximum Gaussian curvature, as defined in differential geometry. Unfortunately, locating this point is a non-trivial computational task.

In the nook changes of parameters lead to small displacements on the L-surface. We compute the surface for a grid of values of the two smoothing parameters. Over this grid we search for the points on the surface that are closest to each other. If we do this for the L-curve, in the case of one-

dimensional smoothing, a curve with a U-shape is obtained. Shahrak (2013) already coined the name U-curve, although with a different definition. To avoid confusion, we christened our procedure the V-valley. We describe it for two dimensions, but it works as well in one too.

The L-surface shares most of the desirable properties characterizing the L-curve procedure. First of all it represents a computationally efficient alternative to classical selection procedure due to the fact that it does not require the computation of the effective model dimension. This issue becomes particularly important if we consider that the computational effort needed for the estimation of the final fits increases easily considering smoothing problems in more than one dimension. On the other hand the L-surface seems to be robust to serial correlation in the data noise component in analogy with its 1D equivalent.

Figure 4.1 shows the results obtained smoothing simulated data with with a non isotropic tensor product P-spline procedure (see chapter 1) and selecting the smoothing parameters by cross validation and the L-surface respectively. It is possible to notice from figure 4.1 that the estimates obtained selecting the smoothing parameters by cross-validation are quite rough while the estimated surfaced obtained using the parameters by the L-surface appears more appropriate. These data count 2500 simulated observations obtained as $z_{ij} = \sin(x_i) + \cos(y_j) + \epsilon$. with $x \in [-3,3]$ and $y \in [-3,3]$. The observations have been simulated so that the noise component shows serial correlation. This correlation has been induced considering as noise component a surface fitted through an isotropic tensor product P-spline smoother with a not too large λ.

In this chapter we describe the L-surface as a procedure for the selection of the smoothing parameters in two dimensional problems. In particular we will consider tensor product P-spline smoothers described by Eilers and Marx (2003). For all the simulated and real data applications discussed here the fast GLAM algorithm introduced by Eilers et al. (2006) have been used.

The rest of the chapter is organized as follows: in section 4.2 the L-surface is introduced highlighting its connections with the L-curve procedure, in section 4.3 the novel V-valley strategy for the selection of the optimal pair of smoothing parameters is briefly introduced, section 4.4 discusses some aspects related to the shape of the L-surface taking into account simulated data

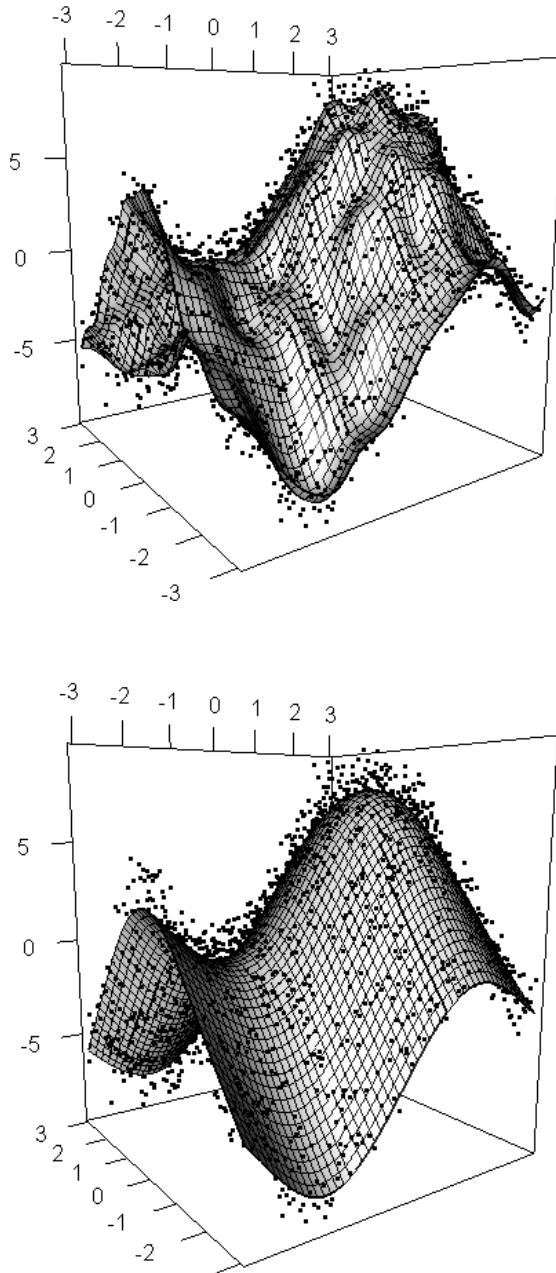**Figure 4.1:** *Multidimensional smoothing of simulated data. The first panel shows a non isotropic tensor product P-spline smoother obtained selecting the λ parameters by cross validation ($\widehat{\lambda_1} = \widehat{\lambda_2} = 0.01$). The second panel shows the smoother obtained selecting the tuning parameters through the L-surface procedure ($\widehat{\lambda_1} = 10$ and $\widehat{\lambda_2} = 31.62$).*

with different characteristics while section 4.5 compares the performances of the proposed selection procedure with those obtained using the well known cross validation criterion on the base of simulated and real examples.

4.2   *The L-surface*

Consider 2D non-isotropic smoothing with tensor product P-splines. We have data vectors x, y and z, where the latter is the dependent variable. Define the following quantities:

$$\psi = \log(\|z - B\beta\|^2); \ \phi_x = \log(\|P_x\beta\|^2); \ \phi_y = \log(\|P_y\beta\|^2),$$

B is the tensor product basis, based on x and y and β the vector of spline coefficient; P_x and P_y are penalty matrices, working on the coefficients in x and y direction. Although not shown explicitly here, these quantities are parameterized by λ_x and λ_y (which get chosen value on a 2D grid). Plotting these quantities in a 3D Cartesian system we obtain points on a surface with a nook (see figure 4.2, second panel). The profiles of the surface represent L-curves. Each L-curve shows a corner region and all curves together define a non-regular grid. The pair of optimal smoothing parameters can be located looking for the point of maximum curvature in the convex nook of the surface.

In analogy with the L-curve framework this point can be located maximizing a suitable curvature measure (see Belge et al., 1998): the Gaussian curvature. For 2D applications the Gaussian curvature can be computed as:

$$k = \frac{\det(II)}{\det(I)} = \frac{LN - M^2}{EG - F^2}. \tag{4.1}$$

The quantities II and I in (4.1) represent, respectively, the second and the first fundamental forms. The first fundamental form allows for the computation the arc length of a curve laying on a surface patch while the second fundamental form measures the degree of curvature of this curve. The computation of the fundamental forms requires the knowledge (or approximation) of the Jacobian and Hessian matrices.

A more generic definition of the Gaussian curvature can be useful in analyzes involving more than two dimensions where a L-hypersurface can be

Figure 4.2: *L-surface computed for the simulated data in figure 4.1. The right panel shows the L-surface while the left one depicts a L curve obtained taking $\log_{10}(\lambda_y) = -2$ and varying $\log_{10}(\lambda_x)$ over the surface.*

used to select the optimal tuning parameters. The so called manifold Gaussian curvature is a suitable measure to locate the optimal smoothing parameters on a generic L-hypersurface:

$$k_M = \frac{\Omega(\dot{L}; \ddot{L})}{\operatorname{tr}(L)^{\frac{(n+1)}{2}}},$$ (4.2)

where $\dot{L}$ and $\ddot{L}$ represent, respectively, the Jacobian and the Hessian matrices and the operator $\Omega(\dot{L}; \ddot{L})$ stays for the sum of each element of the Jacobian times the cofactor of the corresponding element of $\ddot{L}$. Figure 4.2 shows the L-surface related to the simulated example of figure 4.1.

4.3 *The V-valley procedure*

From our discussion appears that the location of the optimal pair of smoothing parameters on the L-surface represents a non-easy task from a computational point of view if it is faced maximizing the Gaussian curvature. On the other hand, looking carefully at the characteristics of the surface plotted in the second panel of figure 4.2 a easier criterion can be deduced.

As we already mentioned the profiles of the L-surface represent one dimensional L-curves. Each L-curve shows a corner and all together define a non regular grid of points on the L-surface. The points on this grid become
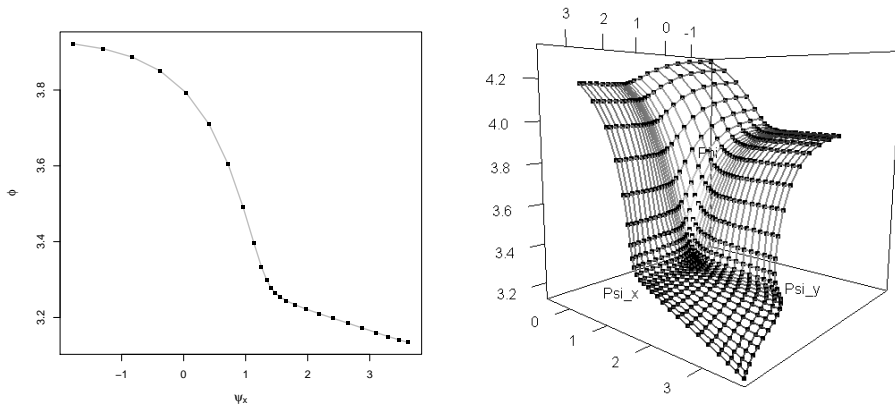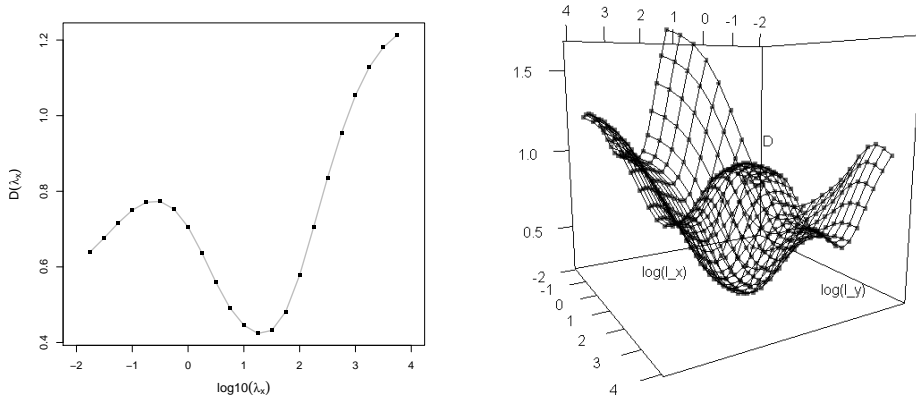
Figure 4.3: *V-valley computed for the simulated data in figure 4.1. The right panel shows the V-valley while the left one depicts a V-curve obtained taking $\log_{10}(\lambda_y) = -2$ and varying $\log_{10}(\lambda_x)$ over the surface.*

closest in the corner (right panel of figure 4.2). Plotting the distances between these adjacent points, a V-shaped curve is obtained (left panel of figure 4.3). All V-curves combined define the "V-valley" shown in the right panel of figure 4.3. The optimal pair (λ_x, λ_y) is located at the bottom of the valley. So an approximation of the maximum curvature point can be obtained searching for the minimum of (4.3). The V-valley procedure have the advantage to avoid the tedious and some times demanding computations related to the numerical estimation of the Jacobian and Gaussian matrices.

$$D(\lambda_x, \lambda_y) = \sqrt{(\nabla_\lambda \psi)^2 + (\nabla_\lambda \phi_x)^2 + (\nabla_\lambda \phi_y)^2}. \tag{4.3}$$

4.4 *The shape of the L-surface*

The L-surface depicted in figure 4.2 shows a clear area characterized by a positive curvature where the optimal smoothing parameters have been located. We can investigate, through simulations, how the shape of the L-surface depends on the characteristics of the data. For all the simulation showed in this section we used tensor product P-splines built considering cubic basis functions defined on 20 equally spaced knots and second order difference penalties in each direction. First of all we evaluate how the L-
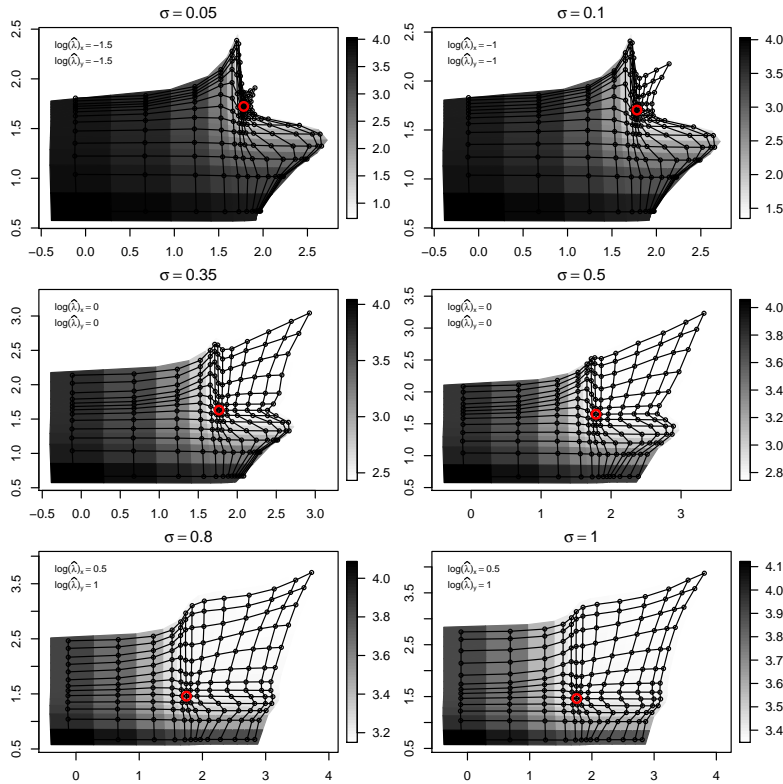
Figure 4.4: *Six projected L-surfaces obtained for different degrees of variability of the noise component, $\sigma_k \in \{0.05, 0.1, 0.35, 0.5, 0.8, 1\}$. The smoothing parameters were selected in a range $[-3, 3]$ for both $\log(\lambda_x)$ and $\log(\lambda_y)$.*

surface shape is related to the variability of the data noise component and, consequently, how the goodness of the final estimates is influenced. Consider the following simulation settings: $z_{ij} = 4(\cos(x_i) + \sin(y_j)) + N_{ij}(0, \sigma_k)$ with $\sigma_k \in \{0.05, 0.1, 0.35, 0.5, 0.8, 1\}$ and $i = j = 1, ..., 50$. Figure 4.4 shows the projections of the estimated L-surfaces for each σ parameter. The red dots represent the optimal smoothing parameters located using the V-valley procedure. The shape of these surfaces is clearly influenced by the standard deviation of the error component. Indeed the area of positive curvature (the nook) tends to become less evident when the variability of the white noise component increases. The impact of the noise variability on the final estimates can be evaluated considering figure 4.5. It shows the contours (solid lines) of the estimated smoothing surfaces obtained for four of the six standard deviation parameters used in figure 4.4. The estimates are compared

with the raw data (background of each plot). It appears that, even if the associated L-surfaces are different, the smoothing surfaces efficiently catch the data behavior. In order to better evaluate the impact of the variability of the noise component on the goodness of the final estimates we can use a bigger simulation and evaluate the distributions of the (log) standardized deviations of the estimated surfaces from the simulated data trend. Figure 4.6 shows these distributions obtained for a grid of 10000 noisy points per each σ parameter mentioned above. The mean deviations in figure 4.6 are all close to zero but the variability of the distributions seems to increase together with the variability of the noise component.

In analogy with the L-curve method, its multidimensional analogous, is robust to data with correlated noise (an example has been shown in figure 4.1). In order to evaluate this feature we can consider a simulated experiment: generate 2500 observations according to $z_{ij} = 4\cos(x_i) + 4\sin(y_j) + E_{ij}(\rho_k)$ where $x = y = -3, ..., 3$ and $E_{ij}(\rho)$ represents the noise term. This noise component has been simulated in such a way to show serial correlation (in analogy with what we did for the example of figure 4.1). The correlation structure has been induced considering smooth surface obtained using a isotropic tensor product P-spline smoother. Here $\rho_k = \{0, 0.1, 0.3, 0.5, 0.75, 0.9\}$ indicates the strength of the penalty imposed in the smoothing procedure used to obtain the noise component. The projections of the L-surfaces resulting from these simulations are depicted in figure 4.7. All the surfaces of figure 4.7 show an area characterized by a clear nook where the optimal smoothing parameters have been located by a V-valley procedure (red dots). In addition, the smoothing surfaces in figure 4.8 show how the estimated smoothers reproduce the underlying behavior for four of the simulated data clouds. In order to evaluate how the goodness of the final estimates is influenced by the degree of serial correlation in the noise component, in analogy with what we did before, we can consider a larger simulation procedure. Figure 4.9 shows the distributions of the standardized deviations of the final estimates from the underlying data trend for different degree of serial correlation of the noise component. These boxplots have been obtained considering a grid of 10000 observations. The standardized mean deviations are all close to zero and the shape of the distributions is almost equal for each ρ parameter.
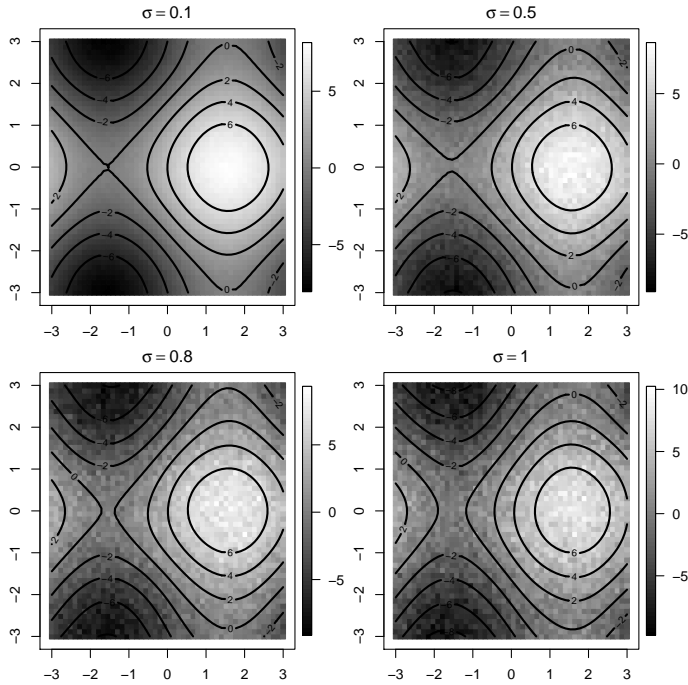
Figure 4.5: *Smoothing surfaces obtained for $\sigma_k \in \{0.1, 0.5, 0.8, 1\}$ and selecting the smoothing parameter using the V-valley procedure.*
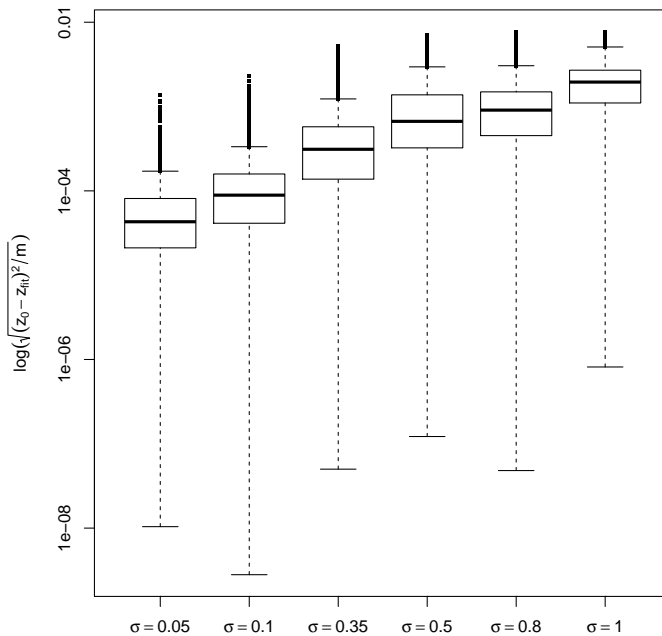
Figure 4.6: *Distributions of the deviations from the underlying trend for smoothing surfaces estimated using data with different variability of the white noise component.*
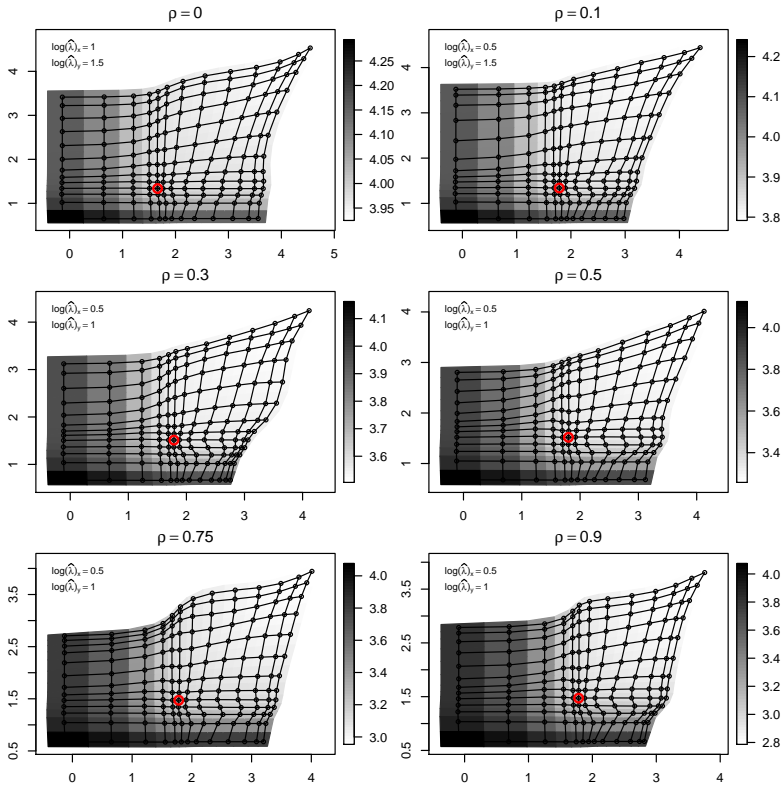
Figure 4.7: *Six L-surfaces obtained for different degrees of serial of the noise component. The smoothing parameters were selected in a range $[-2, 4]$ for both $\log(\lambda_x)$ and $\log(\lambda_y)$.*

The strength of the trend component is another factor that influences the shape of the L-surface. To evaluate its impact we can analyze data simulated as : $z_{ij} = c_k(\cos(x_i) + \sin(y_j)) + N_{ij}(0, 0.5)$ where $c_k \in \{0.01, 0.5, 1, 5, 10, 20\}$ and $N_{ij}(0, 0.5)$ represents a white noise component ($i = j = 1, ..., 50$). The projected L-surfaces obtained for the data with different c_k coefficients are shown in figure 4.10. The effect of the strength of the trend component on the L-surfaces is quite clear. When the data approach to a white noise the region of positive curvature tends to disappear leading to unreliable selection of the smoothing parameters. On the other hand this area becomes more evident for high trend coefficients. This phenomenon has an impact on the shape of the estimated smoothing surface. Figure 4.11 shows four smoothing surfaces obtained for $c_k = [0.5, 2, 5, 10]$.
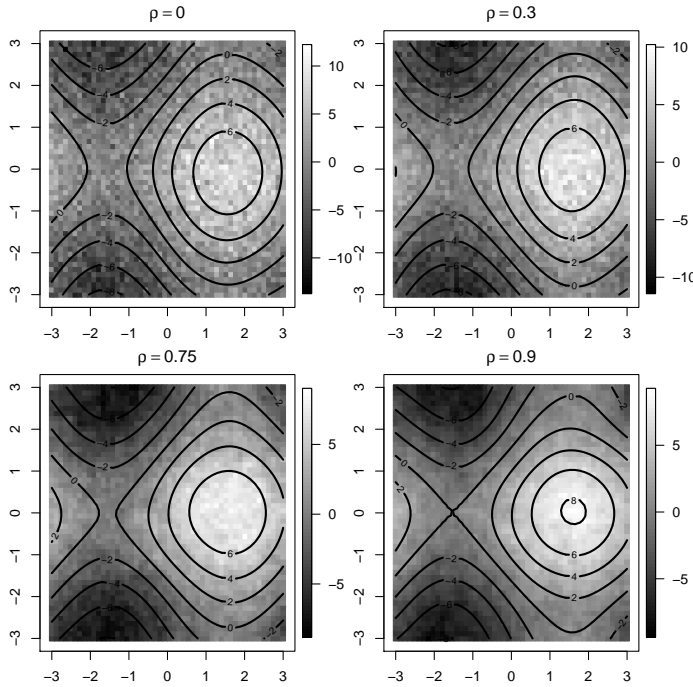
Figure 4.8: *Contours of the smoothing surfaces obtained for $\rho_k \in \{0, 0.3, 0.75, 0.9\}$. The background of each plot represents the raw data.*

Figure 4.9: *Distributions of the deviations from the underlying trend for smoothing surfaces estimated using data with different serial correlation of the noise component.*

Figure 4.10: *Six L-surfaces obtained for different trend coefficients, $c_k \in \{0.01, 0.5, 1, 5, 10, 20\}$. The smoothing parameters were selected in a range $[-2, 5]$ for both $\log(\lambda_x)$ and $\log(\lambda_y)$.*
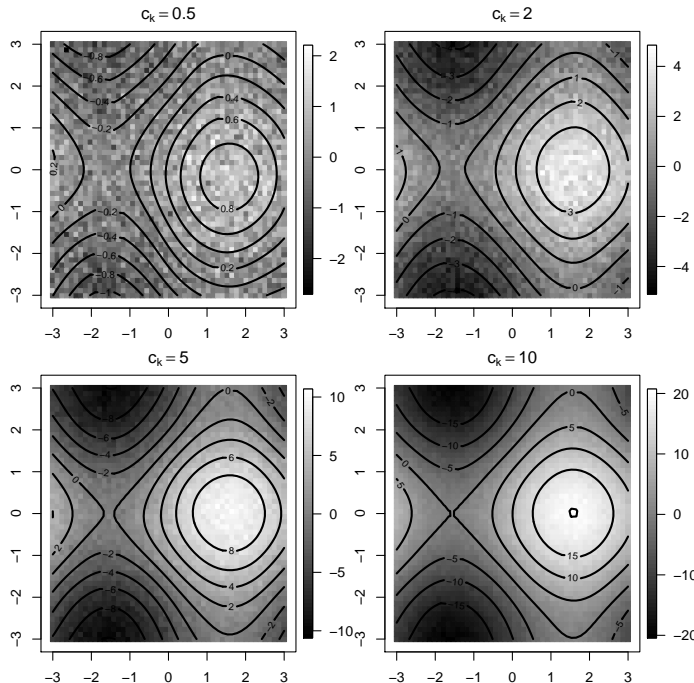
Figure 4.11: *Contours of the smoothing surfaces obtained for $c_k \in \{0.5, 2, 5, 10\}$. The background of each plot represents the raw data.*

4.5   Some examples

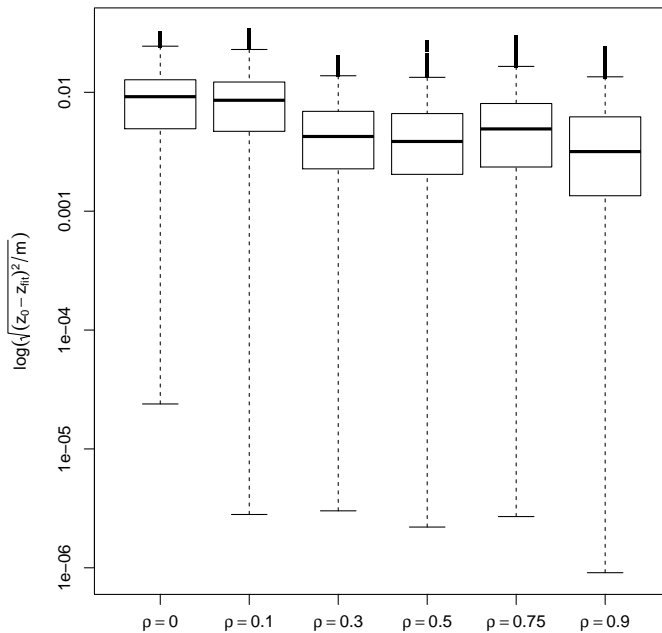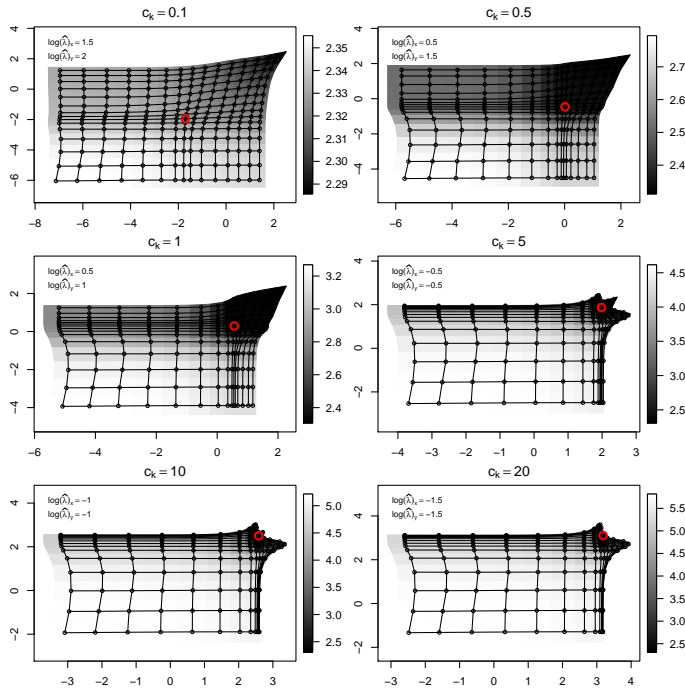In this section some application of the L-surface selection procedure are shown. We consider both simulated and real data examples. As first application we consider the de-noise of a simulated image. The left upper panel of figure 4.12 depicts a a grid of 90×90 pixels describing a quarter of circle to whom a random noise have been added. This example is analogous to the experiment proposed by Eilers et al. (2006). The height of the simulated surface is indicated by the intensity of the image with a light color indicating a greater height. The right upper panel of the same figure shows the enhanced image obtained using a P-spline smoother built on a grid of 22×22 tensor product B-splines and applying third order penalty matrices to both directions. The optimal smoothing parameters have been selected using the L-surface criterion: $\log(\lambda_x) = 1.5$ and $\log(\lambda_y) = 1.5$ (see right panel of the lower part of the figure obtained considering a $\log(\lambda_x) = \log(\lambda_y) \in \{-1, 3\}$ grid of possible smoothing parameter). The left lower panel shows the residuals between the original and the smoothed image. The results obtained selecting the smoothing parameter by cross validation on the same grid of candidate λs are shown in figure 4.13. The two selection procedures give comparable results given that the cross validation suggests $\log(\lambda_x) = \log(\lambda_y) = 1$.

As second example we analyze the monthly total precipitation (mm) for April 1948 in the contiguous United States described in a dataset discussed in Johns et al. (2003). The data and the smoothed precipitation over the US are showed in figure 4.14. The upper panels of this figure show the raw data (left panel) and the smoothed data obtained without considering the boundaries. In the lower panels the smoothed data and the projected L-surface associated are shown. The smoothed image have been obtained considering 10 cubic bases on the x and y directions and the penalties were built using 3th order difference matrices. The optimal smoothing parameter on the L-surface has been selected using the V-valley criterion described above computed over a grid of candidate parameters defined in the interval $\log(\lambda_x) = \log(\lambda_y) \in \{-1, 4\}$. The selected smoothing parameters were $\lambda_x = 0.5$ and $\lambda_y = -0.5$. Figure 4.15 shows the results obtained for the same data smoothing the raw observations using the λ parameters selected by cross validation. As we can see the cross validation surface appears flat
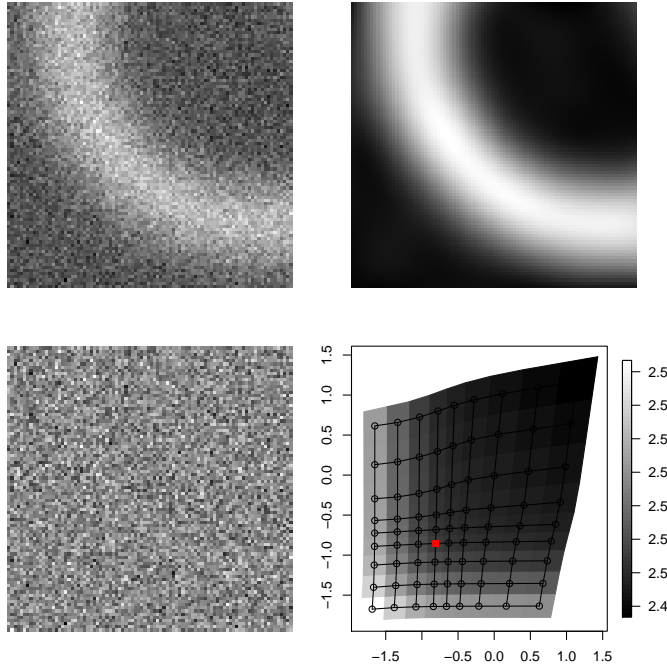
Figure 4.12: *Smoothing of a noisy image using the L-surface smoothing parameter selection criterion. The upper right panel shows the raw data, the left upper panel depicts the result of the smoothing procedure. In the lower panels the residuals and the projected L-surface are shown.*

suggesting small smoothing parameters ($\lambda_x = \lambda_y = -1$). The cross validation surface presents a large flat area and there is not a clear minimum. The overfitting effect produced by the cross validation estimates could be due to the presence of spatial correlation in the noise component of the raw observations. The estimated image overfits the raw data. Also in this case the estimates have been obtained using tensor products of 10 cubic basis functions per each direction built on equally spaced knots and third order difference penalties.

As last example we analyze the well known ethanol data. The data count 88 observations of three variables: concentration of nitrogen oxides (*NOx*, the dependent variable), compression ratio of the engine and equivalence ratio (*C* and *E*). Figure 4.16 shows results for the V-valley and for cross-validation. The cross-validation surface does not show a clear minimum and suggests small λs leading to a wiggly surface. We suspect that this is caused by serial correlation of the errors, along the E direction. The V-valley shows a clear
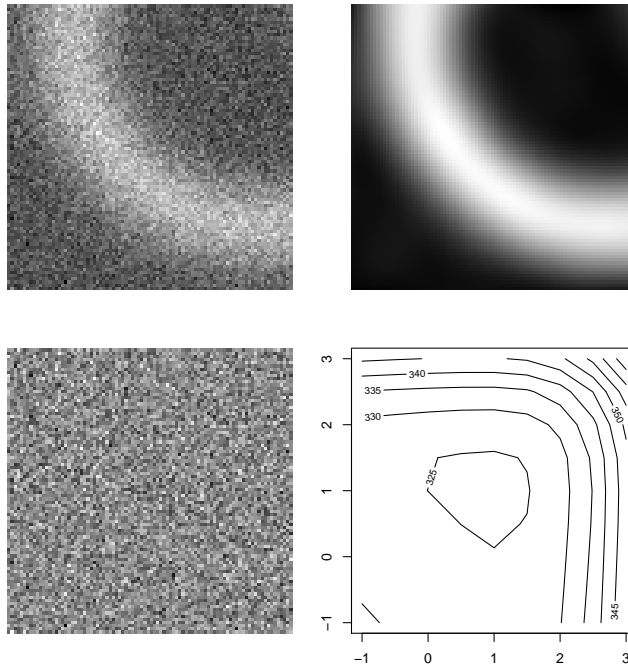
Figure 4.13: *Smoothing of a noisy image using the cross validation smoothing parameter selection criterion. The upper right panel shows the raw data, the left upper panel depicts the result of the smoothing procedure. In the lower panels the residuals and the cross validation surface are shown.*

minimum and leads to a smooth surface with intuitive appeal.

4.6 Conclusions

We have shown how the concept of the L-curve can be extended to two-dimensional smoothing with tensor product P-splines. A natural two dimensional generalization of the L-curve is represented by the L-surface. The L-surface is defined by a set of 1D L-curve computed in the x and y directions. Each L-curve has a convex area and the points defining them become closer and in the proximity of the corner. The Euclidean distance between adjacent points on a single L-curve is described by a V-shaped curve and the single dimensional optimal smoothing parameter is located at the minimum of the V-curve.

This combination of L-curves generates the characteristic shape of the L-surface. The L-surface has a nook in which the right balance between the residual sum of squares and two penalties is found. This nook corresponds
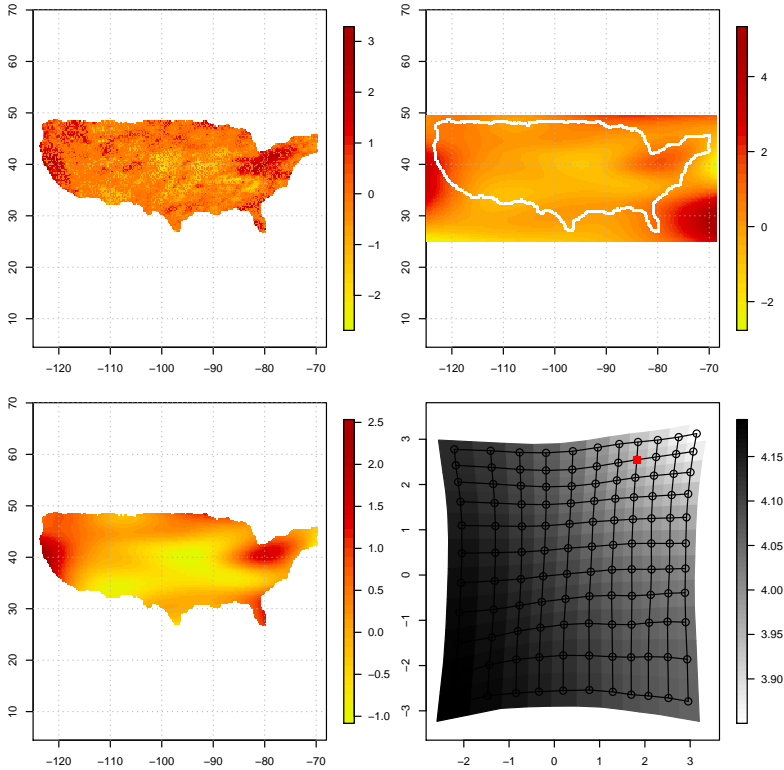
Figure 4.14: *Smoothing of the monthly total precipitation (mm) for April 1948 in the contiguous United States. These results where obtained using the L-surface smoothing selection procedure.*

to the point of maximum Gaussian curvature. The computation of the pointwise Gaussian curvature presents some difficulties given that the (approximated) Jacobian and Hessian matrices related to the components of the L-surface are required.

Instead of using local curvature to locate the right spot, we suggest to use a derived measure, based on distance, leading to simple computations. It is essentially based on the same principle of the V-curve. The distance between adjacent points defining the L-surface in the 3D Cartesian system becomes lower when we are in the proximity of the nook. The Euclidean distance between these points describes a V-shaped surface that can be considered as a generalization of the one-dimensional V-curve. We christianized this surface V-valley. The pair of optimal smoothing parameters is located in the deepest point of the valley.
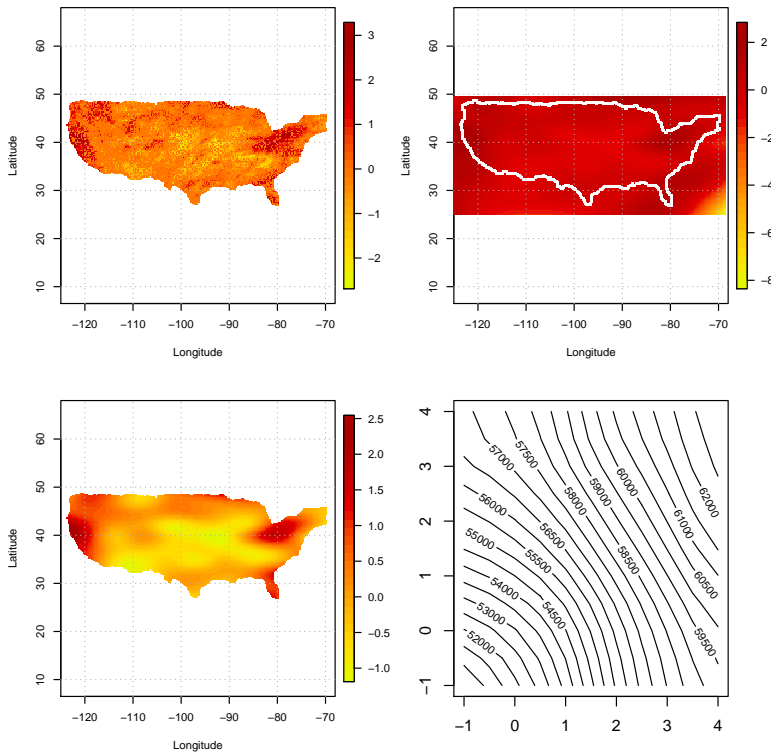
Figure 4.15: *Smoothing of the monthly total precipitation (mm) for April 1948 in the contiguous United States. These results where obtained using a cross validation smoothing selection procedure.*

The L-surface seems to be quite insensitive to serial correlation. This appears clearly from the analysis of the ethanol and US precipitation anomalies data as well as from the simulation studies presented in this chapter. The L-surface criterion in general, and the V-valley selection procedure in our particular case, tends to efficiently separate the trend from the noise component even if the latter is serially correlated.

It appears that our approach opens the road to smoothing in more dimensions, e.g. space and time. More research is needed. A more minimization method will be needed then, because brute-force exploration of a three-dimensional grid of smoothing parameters locating is time-consuming.

There are many opportunities for further research like smoothing of non Gaussian data and penalty terms defined by vector norms different from L_2.

As a final remark we emphasize the crucial role of (the size of) penalties.
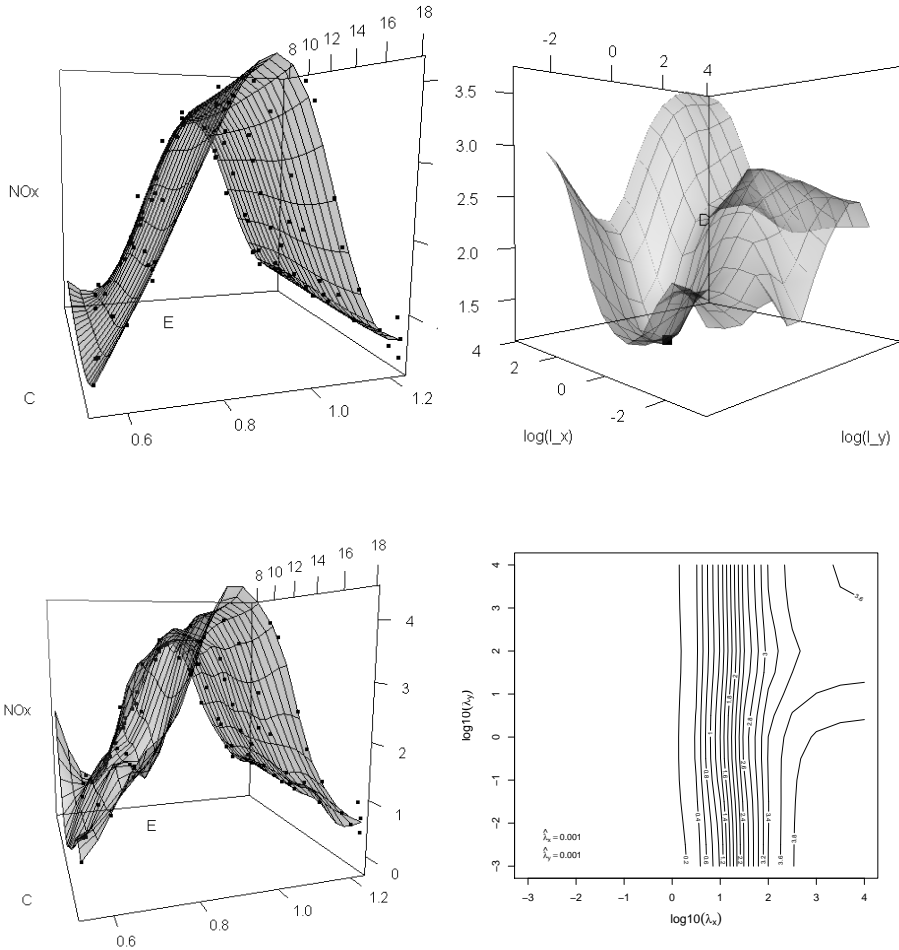
Figure 4.16: *Smoothing of the ethanol data. Top left: the smooth surface obtained after selecting the λ parameters on the L-surface (upper right panel) using the V-valley criterion. The lower left panel shows the smoother estimated using cross validation while the panel on the left shows the cross validation function computed for the grid of smoothing parameters. 23 Cubic B-splines and third order penalties have been used for each dimension.*

Smoothing methods like kernels and local likelihood do not have an equivalent and so they cannot profit from the L-curve and its extensions.

THE B SPLINE COLLOCATION PROCEDURE

<div style="text-align: right;">

5

</div>

> *In this chapter we introduce the B-spline collocation procedure. We treat both ordinary and partial differential problems with initial and boundary conditions. The collocation scheme is applied to the Black and Scholes equation for the pricing of European option contracts. The approximated prices are compared with the analytic solution of the cited differential problem to evaluate the performance of the proposed numerical scheme.*
>
> ***Keywords****: Ordinary/partial differential equations, collocation, B-splines, European option pricing.*

5.1 Introduction

Differential equations are powerful tools for describing system's dynamics. In many circumstances it is advantageous to have at our disposal a procedure to numerically approximate the solution. This is the case when dealing with problems that cannot be solved analytically or if we are interested in inferring the characteristics of a phenomenon which' dynamics is described by observed data, as introduced in chapter 1.

Several numerical procedures have been proposed in the literature to solve ordinary and partial differential problems. The main idea of a numerical method is to replace the DE, formulated for one unknown real valued function, by a discrete equation in finitely many unknowns. The discrete problem defines the so called "numerical scheme". For a complete overview about the different numerical methods proposed in the literature a well known reference is Ascher and Petzold (1998).

Here we introduce the collocation method and discuss it in the case of linear differential problems. The collocation method belongs to the more gen-

eral class of projection methods. The rationale behind all those methods is to approximate the solution of a differential problem by a finite linear combination of basis functions. If we define a basis function $\phi_j(t)$, a projection method suggests to approximate the state function and the initial/boundary conditions by a linear combination of the ϕ_js:

$$u(t) = \sum_j c_j \phi_j(t_i). \tag{5.1}$$

It is clear that, in order to obtain an approximation of the solution, we need a way to determine the coefficients in equation (5.1). The collocation approach suggests to look for a vector of coefficients that guarantees a solution satisfying the differential equation on a grid of values $t_1, ..., t_n$ not necessarily equally spaced. These values define a set of "collocation points".

The selection of optimal collocation points is not straightforward and influences the appropriateness of the approximated solution. The computational cost of the collocation method depends on the number of collocation points. On the other hand, as the number of collocation points increases, we satisfy the differential equation at more points. A possible solution of this trade-off is to chose a limited number of collocation points satisfying a certain optimality criterion. For example, many authors suggest to select them according to a Gauss-Legendre scheme (de Boor and Swartz, 1973). An alternative is to select the collocation points in correspondence of the maximum value of the basis functions. This leads to the so called "smoothest collocation procedure". A comparison between the Gauss-type collocation and this last strategy found in Botella (2002). In this chapter and in the rest of this work we don't discuss this issue and locate the collocation points in such a way to obtain a relatively dense coverage of the domain.

Our discussion is organized as follows: section 5.2 introduces the B-spline collocation scheme for the solution of ordinary differential equations considering both initial and boundary value problems, in section 5.3 we show how to solve partial differential equations using a collocation scheme built on tensor product B-splines. In section 5.4 the collocation approach is evaluated solving the Black and Scholes equation for the pricing European options.

5.2   *Collocation solution of ordinary differential equations*

To clarify how a B-spline collocation procedure provides a numerical approximation of the state function solving an ordinary differential problem, it is useful to show some examples. Consider the following first order order IVP:

$$y' = 2y, \ t \in \{0, 2\}$$
$$y(t = 0) = 2,$$

having the following analytic solution:

$$y^*(t) = \exp(2t + \log(2)).$$

We approximate this solution using a system of 50 B-splines of 3th degree. Consider $t_i \in \{0, \ 2\}$ with $i = 1, ..., 200$. The differential equation is approximated in terms of basis splines as follows:

$$\sum_j c_j B'_j(t_i, p) - 2 \sum_j c_j B_j(t_i, p) = e_i$$
$$\text{s.t. } \sum_j c_j B_j(t_i = 0) = 2, \tag{5.2}$$

where e_i is the approximation error evaluated on the ith collocation point, B and B' represent matrices of B-splines and their derivatives respectively. In order to obtain a numerical solution of the differential problem we need to compute the B-spline coefficients c_j under the constraint represented by the initial value condition. These coefficients can be computed minimizing $\sum_i e_i^2$ (a least squares criterion) solving a linear system with Lagrange multipliers:

$$\begin{bmatrix} V^T V & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} c \\ l \end{bmatrix} = \begin{bmatrix} V^T 0 \\ b \end{bmatrix}, \tag{5.3}$$

where V is a (200×53) matrix (200 domain points and 50 cubic basis splines built on equidistant knots) defined as $V_{ij} = B'_j(t_i, p) - 2 B_j(t_i, p)$, c is a vector of B-splines coefficients of length 53, l is a Lagrange multiplier, $K = B_j(t_i = 0, p)$ is the (1×53) vector of constraints and $b = 2$ is the imposed initial value for the state function. Figure 5.1 shows the numerical results obtained for this IVP and compares them with the analytic solution.

   The number of collocation points and their location play a crucial role in determining the quality of the approximate solution.

   The same procedure can be applied to boundary value problems with
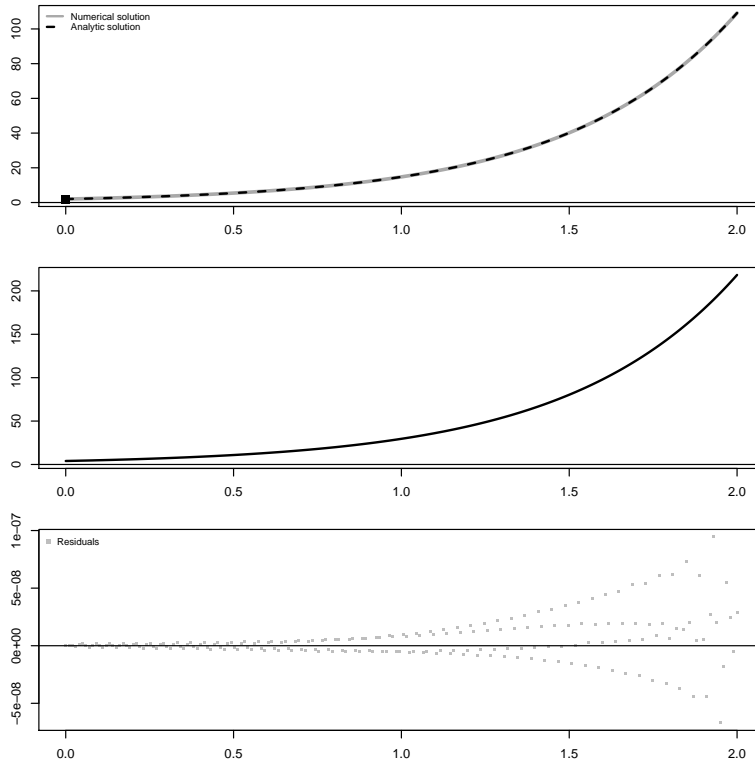
Figure 5.1: *B-splines collocation solution of an ordinary IVP. The first panel shows the analytic and the numerical solution of the IVP. The second panel shows the first derivative, while the third panel shows the residuals between the analytic and the numerical solutions. The squared black dot represents the initial value.*

mixed conditions. For example, consider the following second order BVP:

$$y^{''} + y^{'} + 2y = 0, \; t \in \{0, 2\pi\}$$
$$y(0) = 0, \; y^{'}(2\pi) = 1,$$

with analytic solution $y^*(t) = \exp(-\frac{t}{2})(-22.4\sin(\frac{\sqrt{7}}{2}t))$. The numerical solution is obtained solving a linear system similar to (5.3) with $V_{ij} = B_j^{''}(t_i) + B_j^{'}(t_i) + 2B_j(t_i)$ and $K = [B(t_i = 0) \; B^{'}(t_i = 2\pi)]^T$. It is shown in figure 5.2.

This last example gives us the possibility to discuss a simple generalization of the collocation procedure. Indeed, it is possible to adapt equation (5.3) to solve systems of differential equations. In many circumstances differential problems involve systems of equations. Any nth order differential problem can be expressed in terms of a system of lower order differential equations,
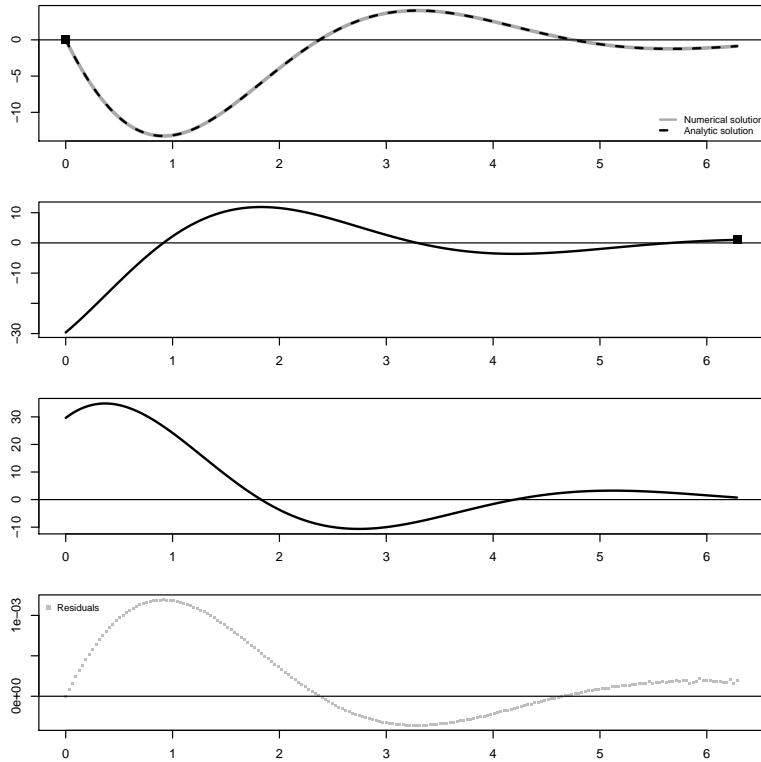
Figure 5.2: *B-splines collocation solution of an ordinary second order BVP. The first panel shows the analytic and the numerical solution of the differential problem. The second and third panels show the first and second derivatives respectively, while the fourth panel shows the residuals between the analytic and the numerical solutions. The squared blacks dot indicate the boundary values. The approximate solution has been obtained using 50 equally spaced cubic B-splines.*

and in many cases it is convenient to deal with lower order derivatives. The BVP above can be translated in the following system:

$$\begin{cases} x_1' = x_2 \\ x_2' = -2x_1 - x_2 \\ x_1(0) = 0, \ x_2(2\pi) = 1, \end{cases} \tag{5.4}$$

where the change of variables $y = x_1$, $y' = x_2$ has been applied.

In this case the least squares system in (5.3) can be generalized. Each of the two top equations in (5.4) leads to residuals as in (5.2) that have to be minimized through leasr squares. One way to compute the optimal spline coefficients is to define the linear system in terms of Kronecker products. In
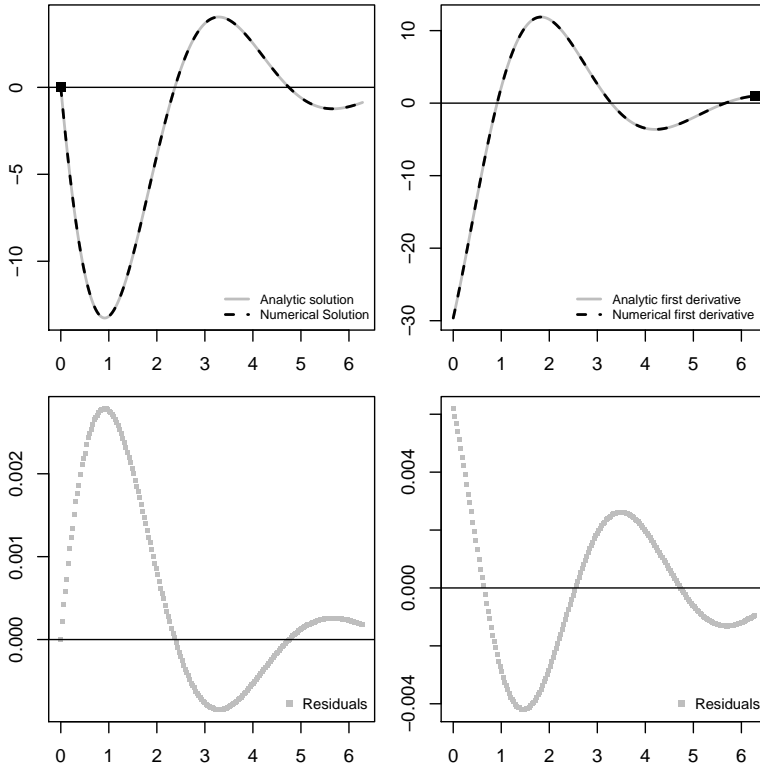
Figure 5.3: *B-splines collocation solution of a system of ODEs. The upper panels show the analytic and the numerical solutions of the differential problems. The lower panels show the residuals between the analytic and the numerical solutions. The squared blacks dot indicate the prescribed values. These results have been obtained using 50 cubic B-splines defined on equally spaced knots selected over 200 collocation points.*

this case the components of (5.3) are: $V_{ij} = \underline{B}'_j(t_i) - \underline{B}_j(t_i)$ and $K = [\breve{B}_j(t_i = 0) \; \breve{B}_j(t_i = 2\pi)]$ with $\underline{B}_j(t_i) = I_2 \otimes B'_j(t_i)$, $\breve{B}_j(t_i) = I_2 \otimes B_j(t_i)$, $\underline{B}_j(t_i) = M \otimes B_j(t_i)$ and $M = \begin{bmatrix} 0 & 1 \\ -2 & -1 \end{bmatrix}$. The analytic and numerical solutions of this differential problem are shown in figure 5.3.

Our last example is related to the first order differential equation shown above with the difference that now we consider an integral condition:

$$y' = 2y, \; t \in \{0, 2\}$$
$$\int_0^{\frac{2}{3}} y(x)dx = 1,$$

with solution $y^*(t) = \frac{2}{3\exp(\frac{2}{3}) - \frac{1}{2}} \exp(2t)$. We again reduce the problem to
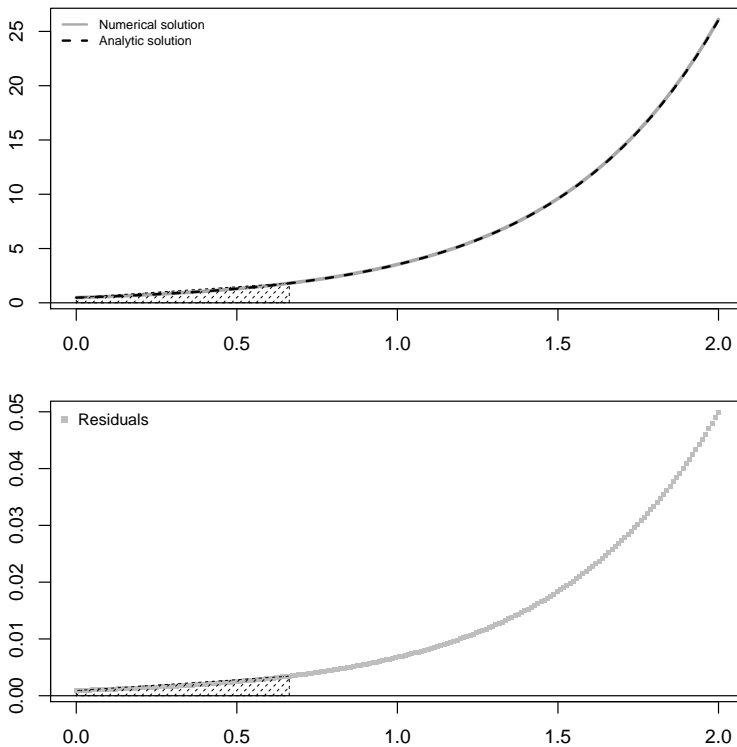
Figure 5.4: *B-splines collocation solution of an ODE with integral constraint. The upper panel shows the analytic and the numerical solutions of the differential problem. The lower panel shows the residuals between the analytic and the numerical solutions. The shaded areas indicate the prescribed integration interval and condition. The numerical solution has been computed using 50 cubic bases built on equally spaced knots selected over 200 domain points.*

the solution of a linear system of equations. We only need to define the K vector in such a way it determines the integral of the B-spline matrix over the imposed integration interval. The results for this example are depicted in figure 5.4.

5.3 Collocation solution of partial differential equations

Many PDEs don't have an analytic solution and a numerical procedure has to be applied to approximate the unknown state function. On the other hand numerical procedures can also be useful in those cases where we deal with observed two-dimensional dynamics and we want to infer their characteristics from the available measurements. We suggest to generalize the one-dimensional B-spline collocation method, using tensor product B-splines.

To clarify this point it is convenient to consider some examples. The first example we would like to introduce is a first order partial differential equation describing a diffusion phenomenon:

$$2u_x + u_y + 3u = 0, \ x \in \mathbb{R}, \ y \in \mathbb{R}^+$$
$$u(x,0) = \frac{1}{1+x^2},$$

where x_x and $_y$ are the first derivative of the state function w.r.t. x and y respectively. According to the collocation framework it is possible to approximate the state function $u(x,y)$ and its partial derivatives using a linear combination of basis functions. Dealing with PDEs we need to choose a grid of collocation points. We locate them on a regular grid corresponding to the maxima of the two-dimensional basis functions.

If we define two B-spline matrices B_x and B_y on x_i, $i = 1, ..., n$ and y_j, $j = 1, ..., m$ collocation points located in the x and y directions respectively, the state function and its derivatives are approximated by:

$$\widehat{u} = B_x C (B_y)^T,$$
$$\widehat{u}_x = B'_x C (B_y)^T,$$
$$\widehat{u}_y = B_x C (B'_y)^T,$$

where C is a matrix of unknown spline coefficients. These coefficients can be computed through least squares, setting:

$$\sum_i \sum_j (B_{y_j} \otimes B'_{x_i}) c_{ij} + \sum_i \sum_j (B_{y_j} \otimes B'_{x_i}) c_{ij} + 3 \sum_i \sum_j (B_{y_j} \otimes B_{x_i}) c_{ij} = e_{ij}$$
$$\text{s.t.} \ \sum_i \sum_j (B_y(y_j = 0) \otimes B_{x_i}) c_{ij} = \frac{1}{1+x_i^2},$$

and minimizing the magnitude of the approximation error $\sum_i \sum_j e_{ij}^2$.

For the present IVP the spline coefficients follow from:

$$\begin{bmatrix} V^T V & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} c \\ l \end{bmatrix} = \begin{bmatrix} V^T 0 \\ b \end{bmatrix},$$

where $V = [2B'_{y_j} \otimes B_{x_i} + B_{y_j} \otimes B'_{x_i} + 3B_{y_j} \otimes B_{x_i}]$ is the B-spline approximation of the differential equation, $K = [B_y(y_j = 0) \otimes B_{x_i}]$ is a constraint matrix, c is obtained vectorizing the columns of a spline coefficient matrix, l is a vector of
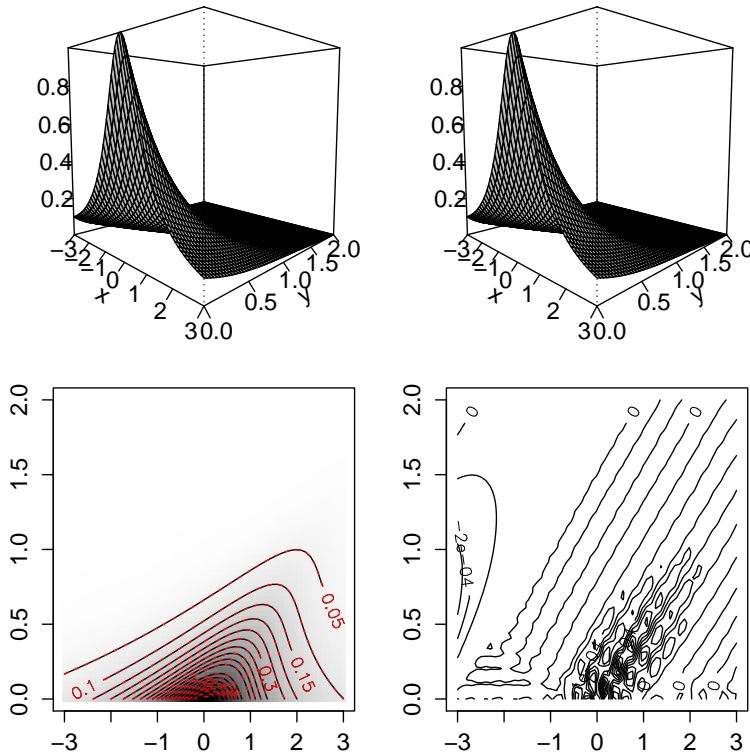
Figure 5.5: *Analytic and numerical solution of a first order diffusion PDE, the right lower panel shows the error surface. The approximate solution has been obtained using 20 cubic B-splines built on equidistant knots selected over 30 domain points in each direction.*

Lagrange multipliers and $b = \frac{1}{1+x_i^2}$ is the vector of initial values. The analytic and approximate solution of the cited PDE are shown in figure 5.5 together with the error surface.

A more complex example is represented by a Cauchy problem with mixed conditions:

$$u_{yy} - u_{xx} = 1, \; x \in \mathbb{R}, \; y \in \mathbb{R}$$
$$u(x,0) = x^2, u_y(x,0) = 1,$$

where u_{yy} and u_{xx} represent, respectively, the second derivatives of the state function $u(x,y)$ w.r.t. y and x. This differential problem has analytic solution
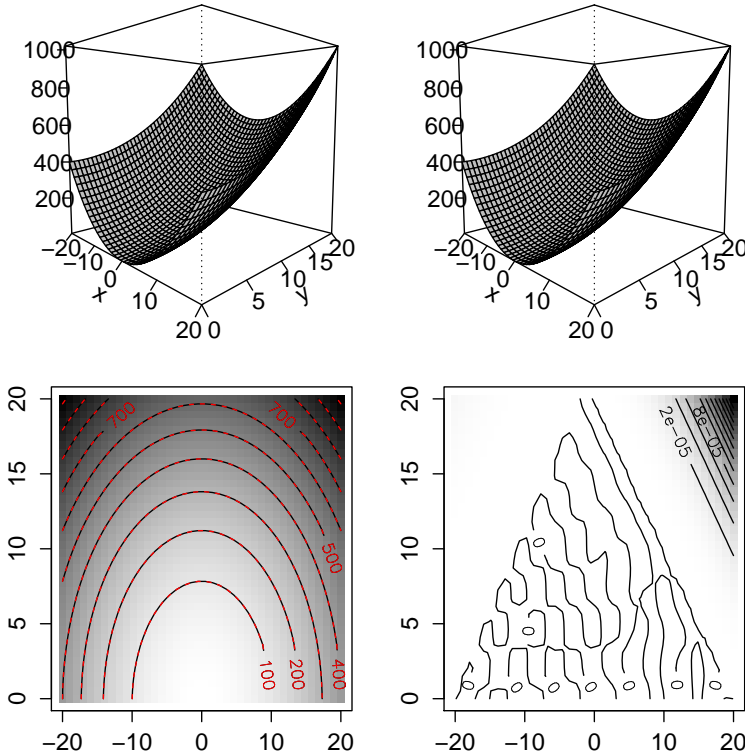
Figure 5.6: *Analytic and numerical solution of a second order Cauchy PDE, the right lower panel shows the error surface. The state function has been approximated using fourth order basis splines defined on 15 equally spaced knots in each direction.*

$u^*(x, y) = x^2 + y + 1.5y^2$. We now have:

$$\widehat{u} = B_x C(B_y)^T,$$
$$\widehat{u}_{xx} = B_x'' C(B_y)^T,$$
$$\widehat{u}_{yy} = B_x C(B_y'')^T,$$
$$\widehat{u}_y = B_x C(B_y')^T.$$

The optimal spline coefficients c are computed solving a system of equations similar to (5.3), with $V_{ij} = [B_{y_j}'' \otimes B_{x_i} - B_{y_j} \otimes B_{x_i}'' - 1]$, $K = [B_y(y_j = 0) \otimes B_x \; B_y'(y_i = 0) \otimes B_{x_i}]^T$, $b = [x^2 \; 1]^T$. Figure 5.6 compares the analytic and numerical solutions.

5.4   *Option pricing through B-spline collocation*

To evaluate the performance of the B-spline collocation scheme introduced in the previous section we propose to solve the Black and Scholes equation (Black and Scholes, 1973) and compare the approximated and the exact solutions.

The B&S equation is the most famous mathematical pricing model for European style financial options. Quoting Cox et al. (1979), an option is a contract that gives the right, but not the obligation, to buy or sell a risky asset at a predetermined fixed price within a specified period. This contract allows to bet about the future evolution of the price of the underlying asset. The underlying can be a financial instrument (a stock for example), a financial index and also a commodity. Options and other related contracts are often defined financial "derivatives". It is because their price is derived from the market value of the underlying risky asset.

Two different kinds of option contracts are distinguished according to the right that they give to the holder. A "call" is a contract that gives the right to buy the underlying asset while a "put" gives the right to sell it. These characteristics can be summarized in the payoffs function at expiration (maturity) time T associated to these two different contracts. If we indicate with $C(S,T)$ the payoff of a call at maturity and with $P(S,T)$ the payoff of a put, the definition above can be summarized as follows:

$$C(S,T) = \begin{cases} 0 & \text{if } S_T \leq E \\ S_T - E & \text{if } S_T > E, \end{cases}$$

$$P(S,T) = \begin{cases} 0 & \text{if } S_T \geq E \\ E - S_T & \text{if } S_T < E, \end{cases}$$

where with E we indicate the strike price (the predetermined price of the underlying at expiration time) and S_T represents its market price at maturity.

The Black and Scholes model moves from the assumption that the dynamics of the price of the underlying risky asset follows a geometric Brownian motion in the instantaneous time dt with mean μ and standard deviation $\sigma\sqrt{dt}$. Consider a option written on a share. If, in analogy with the previous notation, we indicate with S the stock price this assumption implies that:

$$\frac{dS}{S} \sim \phi\left(\mu dt, \sigma^2 dt\right), \tag{5.5}$$

where dS is the stock price variation at instant dt and $\phi(\cdot)$ indicates a Gaussian probability density function. If S_T is the stock price at maturity time this dynamics implies that:

$$\ln(S_T) - \ln(S_0) \sim \phi\left[\left(\mu - \frac{\sigma^2}{2}\right)T, \sigma^2 T\right],$$

$$\ln(S_T) \sim \phi\left[\ln(S_0) + \left(\mu - \frac{\sigma^2}{2}\right)T, \sigma^2 T\right]. \tag{5.6}$$

This implies that S_T follows a log-normal distribution such that $E(S_T) = S_0 \exp\{\mu T\}$ and $\text{Var}(S_T) = S_0^2 \exp\{2\mu T\}[\exp\{\sigma^2 T\} - 1]$. The dynamic of S under these assumptions are then described by:

$$dS = \mu S dt + \sigma S dz. \tag{5.7}$$

Let $V(S,t)$ be the value of the derivative contract. Following the definition of an option contract its price depends on: the maturity date T (or better of the time to maturity $t - T$), the risk free interest rate r, the volatility of the underlying asset σ, the strike price E and the spot price of the underlying S. Those variables are all included in the B&S model. Indeed, this model states that the mathematical law governing the value of the option can be written as the PDE:

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV. \tag{5.8}$$

The solution depends on the boundary conditions that define the value of the derivative contract for extreme values of S and t. The boundary conditions are defined to be consistent with the payoff functions. Consider a plain vanilla (which price does not depends on accessory rights) European call option written on a stock that does not pay dividends. In this case the boundary conditions for the differential problem (5.8) are:

$$C(0,t) = 0$$
$$C(S,t) = S - E\exp\{-r(T-t)\} \text{ for large } S$$

On the other hand the terminal condition, the boundary condition defined for $t = T$, can be written as:

$$C(S,T) = \max(S - E, 0) = (S - E)^+$$

Analogous terminal and boundary conditions hold for a vanilla European put option. Given those conditions the solution of (5.8), for a call and a put option respectively is:

$$C(S,t) = S_0 \Phi(d_1) - E \exp\{-rt\} \Phi(d_2)$$

$$(5.9)$$

$$P(S,t) = E \exp\{-rt\} \Phi(-d_2) - S_0 \Phi(-d_1)$$

where $\Phi(\cdot)$ indicates the standard normal cumulative distribution function and:

$$d_1 = \frac{\ln(S_0/E) + (rt\sigma^2/2)T}{\sigma\sqrt{T}},$$
$$d_2 = d_1 - \sigma\sqrt{T}.$$

5.4.1 Pricing a plain vanilla European call option

An European vanilla call option is a contract giving to the owner the right to buy a share of a specific common stock (whose price is $S(t)$ at time t) at a fixed price E at a certain date T. In the previous introduction we have defined the Black and Scholes equation for the value $C(S,t)$ of an European call option on an asset of value of value with a given volatility and interest rate. As we showed there exists a closed form solution of the B&S equation for this kind of option contract. It gives us the possibility to test the quality of the numerical solution achieved through the proposed B-spline collocation scheme.

As an example, we consider an European vanilla call option with exercise price $E = 100$, volatility $\sigma^2 = 0.3^2$, risk free rate $r = 0.1$, and exercise time 1 year. The change of variables $U(x,\tau) = \frac{1}{E}V(E\exp(Lx),k)$ with $\tau = T-t$, $x \in [-1,1]$, $k = 2\tau/\sigma^2$ and $L = -r(T-t)$ leads to the following reformulation of (5.8):

$$\frac{\partial U}{\partial \tau} = -L^{-2}\frac{\partial^2 U}{\partial x^2} - (k-1)L^{-1}\frac{\partial U}{\partial x} + kU, \qquad (5.10)$$

with terminal condition $U(x, \sigma^2 T/2) = \max(\exp(Lx)-1, 0)$ and the boundary conditions $U(-1,\tau) = 0$, and $U(1,\tau) = \exp(L) - \exp(-k(T\sigma^2/2 - \tau))$.

Following the framework described in the previous sections we compute the basis matrices B_{x_i}, B_{τ_j}, B''_{τ_j}, B''_{x_i}, B'_{τ_j} and $B_{x_i,\tau_j} = B_{\tau_j} \otimes B_{x_i}$ and the partial derivatives basis matrices $U_x = B_{\tau_j} \otimes B'_{x_i}$, $U_{xx} = B_{\tau_j} \otimes B''_{x_i}$, $U_t = B'_{\tau_j} \otimes B_{x_i}$.
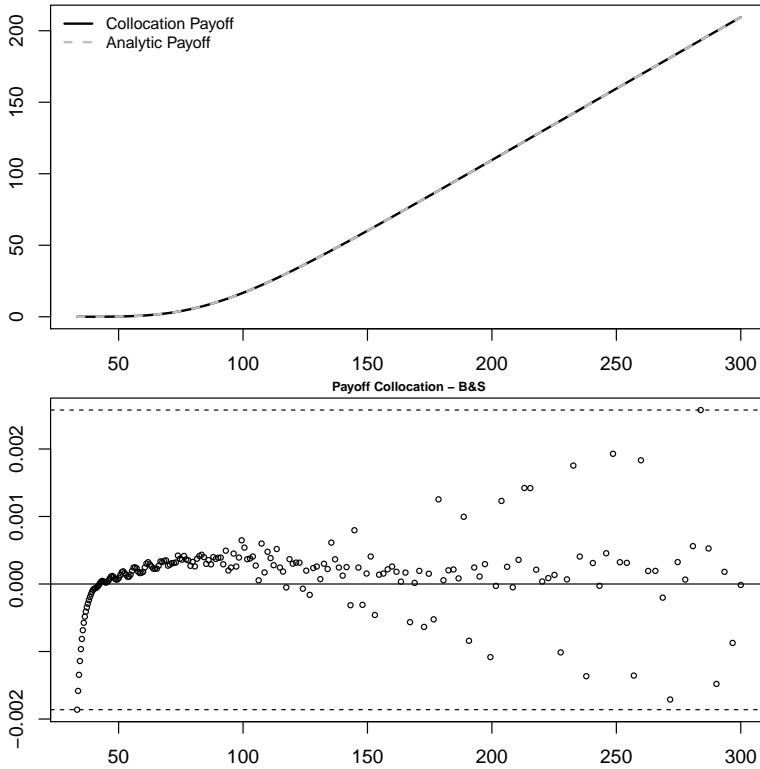
Figure 5.7: *The top panel shows the Numerical and analytic payoffs at maturity. The lower panel shows the differences between the numerical results and the analytic ones.*

This implies that the differential problem can be written stacking the differential equation and the related conditions in a linear system of equations:

$$
\begin{bmatrix}
U_\tau + L^{-2}U_{xx} + (k-1)L^{-1}U_x - kB_{x,\tau} \\
B_\tau \otimes B_x(x=-1) \\
B_\tau \otimes B_x(x=1) \\
B_\tau(\tau = T\sigma^2/2) \otimes B_x
\end{bmatrix}
\begin{bmatrix}
c \\
l
\end{bmatrix}
=
\begin{bmatrix}
0 \\
0 \\
e^L - e^{-k(T\sigma^2/2-\tau)} \\
\max(e^{Lx}1,0)
\end{bmatrix}, \quad (5.11)
$$

where l is a vector of Lagrange multipliers and c is the spline coefficient vector to be optimized. Figure 5.7 compares the price to maturity of the considered toy option contract computed through B-spline collocation and the analytic solution of the B&S model. The approximation error is quite small. Figure 5.8 shows the numerical and the analytical solutions for different times to maturity. Also in this case it is possible to notice how close the two solutions are.
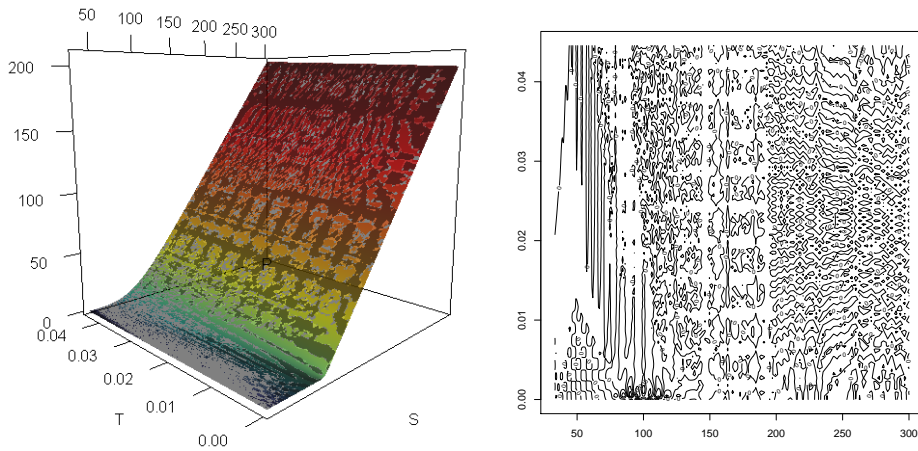
Figure 5.8: *The left panel plots the numerical solution for a European call option (colored surface) and the B&S price (gray surface) for different times to maturity. The right panel shows the approximation error contours. These results have been obtained using 15 cubic basis splines built on equidistant knots selected on the x and τ directions.*

5.5   Conclusions

In this chapter we introduced the B-spline collocation procedure for the solution of linear ordinary and partial initial/boundary value problems. Collocation is an efficient of one-step numerical procedure. All the discussed examples were related to differential problems with analytic solutions so that it was possible to evaluate the quality of the results.

The selection of the number and the location of the collocation points becomes an issue in many numerical application. The accuracy of the approximation and the computational cost required by the collocation scheme depend on these aspects. One way is to choose a modest number of collocation points and locating them according to some optimality condition (for example following a Gauss-Legendre scheme). An alternative is to select the collocation points in correspondence of the maximum value of the basis functions. Botella (2002) defines this collocation procedure the "smoothest collocation" scheme and compares it with a Gauss-based one. In this chapter and in the rest of this work adopt a rather intuitive strategy locating the collocation points in such a way to obtain a relatively dense coverage of the domain. The accuracy in terms of error analysis of our choice was not investigated and

will be part of our future research.

As practical application we have shown how to solve the Black and Scholes differential problem for pricing European options using a B-spline collocation procedure. A plain vanilla call option has been considered. The numerical approximation of the price to maturity has been compared with the analytic solution of the B&S model. The same comparison has been performed computing the price of an analogous option contracts with different times to maturity. In all the cases the price approximated by collocation has been found really close to the price computed using the analytic solution of the differential problem.

Further generalizations of the collocation scheme are possible and will be illustrated in the coming chapters. We will show, for example, that it is possible to define a collocation procedure able to take into account inequality constraints imposed on the state function or on one (or more) of its derivative(s). Analogously, using a B-spline collocation procedure, it is possible to approximate state functions imposing irregularities on its domain or codomain. Finally, the collocation approach is flexible enough to be adopted to approximate the solution of different classes of differential problems such as time varying parameter, non linear and delayed equations.

6

DIFFERENTIAL PENALIZED SMOOTHING

Smoothing procedures driven by differential equations have gained popularity in the last few years. These methods combine basis functions and differential operators in a penalized estimation setting. The results of the analysis are interpretable in relation to the dynamics described by the differential law taken into account. In this chapter we show an approach for functional smoothing able to take into account not only the differential operator but also initial/boundary conditions. We also propose a two-stage procedure based on P-splines for the estimation of the unknown DE parameters.

Keywords: *B-splines, collocation method, P-splines, Ordinary Differential Equations, generalized conditions.*

6.1 Introduction

Penalized smoothing techniques have a prominent place in modern over-parametric regression analysis. The penalty term determines the characteristics of the final fit, balancing between smoothness and fidelity to the data of the estimated curves. In this chapter we formulate a penalized smoothing problem based on B-spline functions guaranteeing a certain degree of flexibility to the final estimates.

Different definitions of the penalty term have been proposed in the literature. Smoothing splines are based on a penalty term representing the integrated second derivative of the final fit. On the other hand penalized splines (P-splines), introduced by Eilers and Marx in 1996, define a penalty term obtained as differences between the spline coefficients.

More general penalties are possible. L-spline smoothing techniques consider penalty terms combining the derivatives of the final estimates defining

a differential operator. The idea is to smooth the raw data penalizing the least squares problem taking into account a model for the underlying dynamics. Examples of this kind of approach have been discussed by Schumaker (1981), Ramsay and Silverman (2005), Welham et al. (2006), Wood et al. (2008) and many others. In this framework the observed data can be viewed as a sort of generalized conditions applied to a numerical scheme for the numerical solution of a differential equation.

This approximation is obtained by balancing between two conflicting aims: the fidelity to the data (measured by the residual sum of squares) and the fidelity to the approximate solution of the differential problem behind them (obtained integrating the differential operator included in the penalty):

$$\min_c \|y - x(t,c)\|^2 + \lambda \int [L(x(t,c),w)]^2 dt,$$

where c is the vector of of basis coefficients determining the smoothing function, $L(x(t,c),w)$ is a differential operator that depends on the differential parameters vector w and λ is a smoothing parameter determining the emphasis given to the numerical scheme in the estimation procedure.

In practical applications the parameters w have to be estimated from the available observations. A possibility is to estimate them minimizing a joint penalized log-likelihood. The penalized minimization problem can be simplified, exploiting the hierarchy existing between the unknowns and considering the basis coefficients as nuisance parameters, leading to the so called "generalized profiling technique" proposed by Cao and Ramsay (2007).

In this chapter we propose to estimate the unknown DE parameters using a P-spline-based two-stage approach moving from the consideration that, if the differential problem represents a good description of the observed measurements, the signal has to be consistent with the state function.

This chapter is organized as follows: in section 6.2 the differential penalized smoothing approach is introduced in the case of known differential operators. We also discuss the introduction of known and unknown initial/boundary conditions defining the differential problem clarifying the role of Lagrange multipliers in the estimation procedure. In section 6.3 we focus on the role of the smoothing parameter introducing an EM-based algorithm (see Schall, 1991) for its optimization. In section 6.4 we deal with smoothing problems driven by unknown linear differential operators introducing

a P-spline-based two-stage approach to estimate the unknown DE parameters. The performance of this procedure is evaluated through two simulation studies. Finally, in section 6.5, a real data example is analyzed.

6.2   *Smoothing with ordinary differential penalties*

Suppose we observe a set of noisy data described, at least approximatively, by a differential model and have an idea about the ODE describing the dynamic underlying the observations What happens if our knowledge about the differential equation describing the observed dynamics is not precise or even not correct? In this case the solution of the differential problem would provide an inadequate description of the observed measurements. It is possible to overcome this problem forcing the solution of the DE we have in mind to give an appropriate description of the data. This can be achieved using a smoothing approach penalized by a differential operator approximated using a collocation scheme. If we define $\widehat{u}(t) = \sum_i^n B_j(t)\widehat{c}_j$ the approximated state function computed using a B-spline collocation procedure and $\breve{V}(w,t) = \sum_j^n \sum_{d=0}^m w_d \breve{B}_j^{(d)} \widehat{c}_j$ the approximated differential operator, our estimation problem is summarized by:

$$\min_c \|y - \sum_j^n B_j c_j\|^2 + \lambda\| \sum_j^n \sum_{d=0}^m w_d \breve{B}_j^{(d)} c_j\|^2, \tag{6.1}$$

where λ is a smoothing parameter balancing between the residual sum of squares and the collocation approximation of the observed dynamics, c is a vector of unknown spline coefficients, w is a vector of m DE parameters that we consider, for the moment, known (m indicates the order of the differential operator that we consider to be linear in this chapter).

In the formulation above we used the breve accent to distinguish two sets of points on which the basis functions are evaluated. The B-splines involved in the penalty term are evaluated on an "enlarged" set of N collocation points. On the other hand the basis functions in B are evaluated at observed domain points. This definition of the penalty term is convenient in those cases a moderate number of observations is available providing a more precise definition of the collocation solution of the hypothesized differential problem. This issue is particularly relevant when we are interested in estimating the DE parameters from the data, as we discuss later.

To illustrate how the spline coefficients are computed, we present some examples assuming to know the DE parameters. The first simulated example is represented by the data depicted in the first panel of figure 6.1 (squared gray dots). This plot shows 200 observations $y_j(t)$ generated adding a Gaussian noise to the solution of the following differential problem:

$$y' = 2y, \ t \in \{0, 2\}$$
$$y(t = 0) = 2.$$

The optimal spline coefficients can be estimated solving:

$$(B^T B + \lambda \breve{V}^T \breve{V})c = B^T y,$$

with $\breve{V} = \breve{B}' - 2\breve{B}'$. The optimal λ parameter has been selected using an EM-type procedure (see section 6.3). The large optimal smoothing parameter testifies the appropriateness of the imposed collocation scheme in describing the observed dynamics.

In the previous example the initial value condition has not been considered in the smoothing procedure. Lagrange multipliers offer an easy way to include this kind of "hard constraints" in the penalized smoothing problem. Figure 6.2 shows a set of 200 data points simulated adding a Gaussian noise to the analytic solution of the following second order BVP:

$$y'' + y' + 2y = 0, \ t \in \{0, 2\pi\}$$
$$y(0) = 0, \ y'(2\pi) = 1.$$

The estimation problem can be summarized as follows:

$$\min_c \|y - \sum_j^n B_j c_j\|^2 + \lambda \|\sum_j^N (\breve{B}_j'' + \breve{B}_j' + 2\breve{B}_j)c_j\|^2,$$
$$\text{s.t. } \widehat{u}(t = 0) = 0, \ \widehat{u}'(t = 2\pi) = 1.$$

The optimal spline coefficients follow from:

$$\begin{bmatrix} B^T B + \lambda \breve{V}^T \breve{V} & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} c \\ l \end{bmatrix} = \begin{bmatrix} B^T y \\ b \end{bmatrix}, \qquad (6.2)$$

where $\breve{V} = \breve{B}'' + \breve{B}' + 2\breve{B}$, $K = [B(t = 0) \ B'(t = 2\pi)]^T$, c is a vector of spline coefficients, l is a vector of Lagrange multipliers and b is a vector of boundary
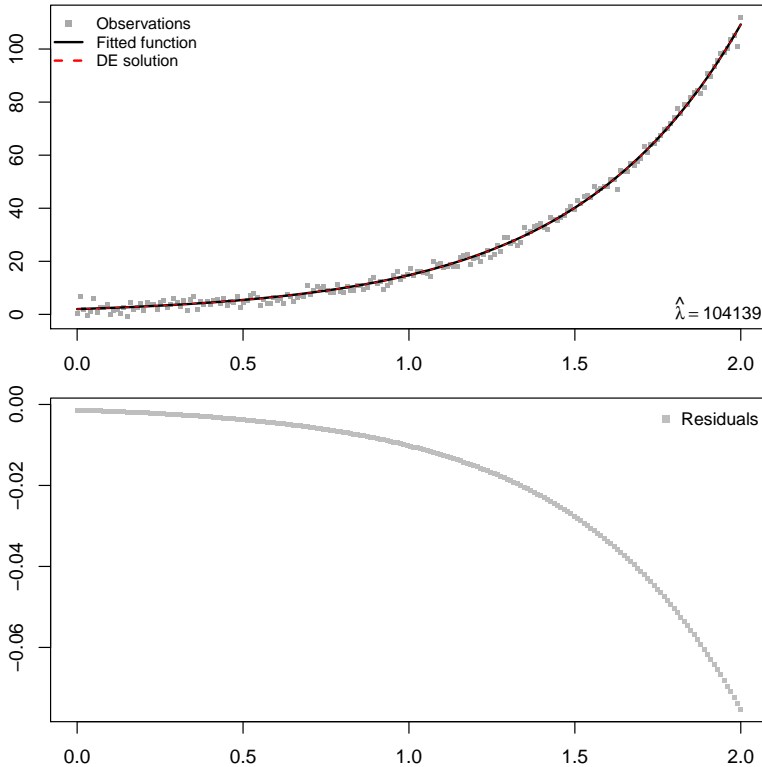
Figure 6.1: *Smoothing with known differential penalties. The data have been simulated taking into account a linear first order ODE. The upper panels shows the raw data and the estimated smoothing function while the lower panel shows the residuals between the smoother and the solution of the differential problem. The estimates have been obtained using the penalized smoothing approach introduced in equation* (6.1) *using cubic B-spline functions built on 50 equally spaced knots. The bases included in the penalty term have been evaluated on a set of 500 points.*

values. The Lagrange multipliers can be interpreted as the shadow price that the smoothing procedure has to pay to respect these hard constraints. For this reason we expect small Lagrange multipliers if the imposed conditions are consistent with the data dynamics.

It is also possible to introduce unknown prescribed conditions and estimate them from the available data. The procedure is analogous to what we have described in (6.2) with the difference that now the vector b is unknown. The optimal differential conditions are inferred from the observed measurements exploiting a least squares procedure. Using the same notation introduced above, we define $Q = \breve{V}^T \breve{V} + K^T K$ and $B(x)Q^{-1}M^T = G$. The
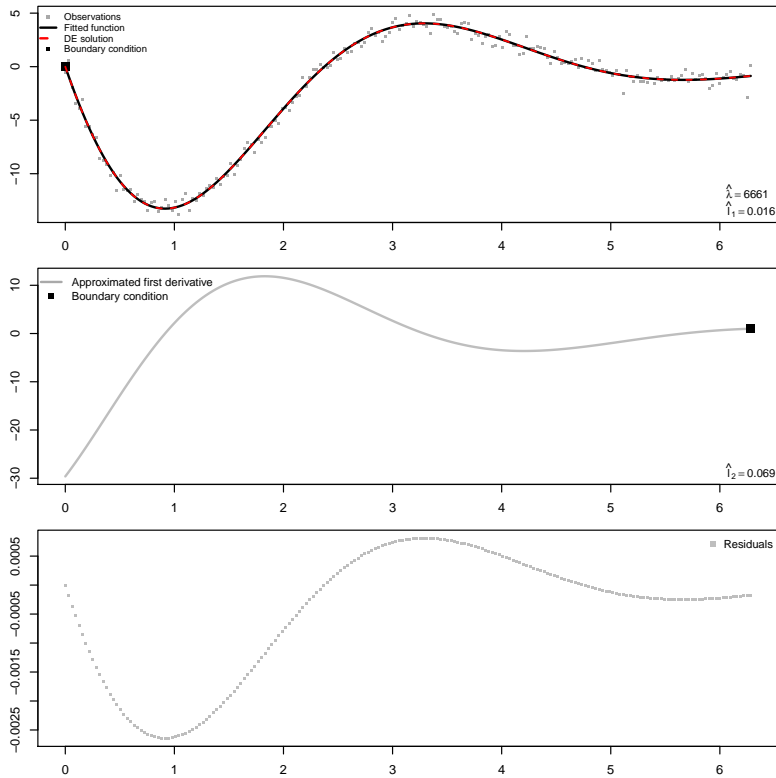
Figure 6.2: *Smoothing with known differential penalties and boundary conditions. The data have been simulated taking into account a linear second order ODE. The upper panels shows the raw data and the estimated smoothing function while the lower panel shows the residuals between the smoother and the solution of the differential problem. These estimates have been obtained considering 50 cubic spline functions built on equally spaced knots. The bases involved in the penalty term have been evaluated on a grid of 500 domain points.*

optimal prescribed values follow from:

$$(G^T G)b = G^T y.$$

Once that these values have been estimated they can be plugged in (6.2) to find the spline coefficients minimizing the constrained penalized smoothing problem.

Figure 6.3 shows the results obtained smoothing a series of data described by the system of ODEs:

$$\begin{cases} x_1' = x_2 \\ x_2' = -x_1 - 0.5x_2 \\ \text{s.t. } x_1(0) = 0, \ x_2(0) = 2 \end{cases}$$
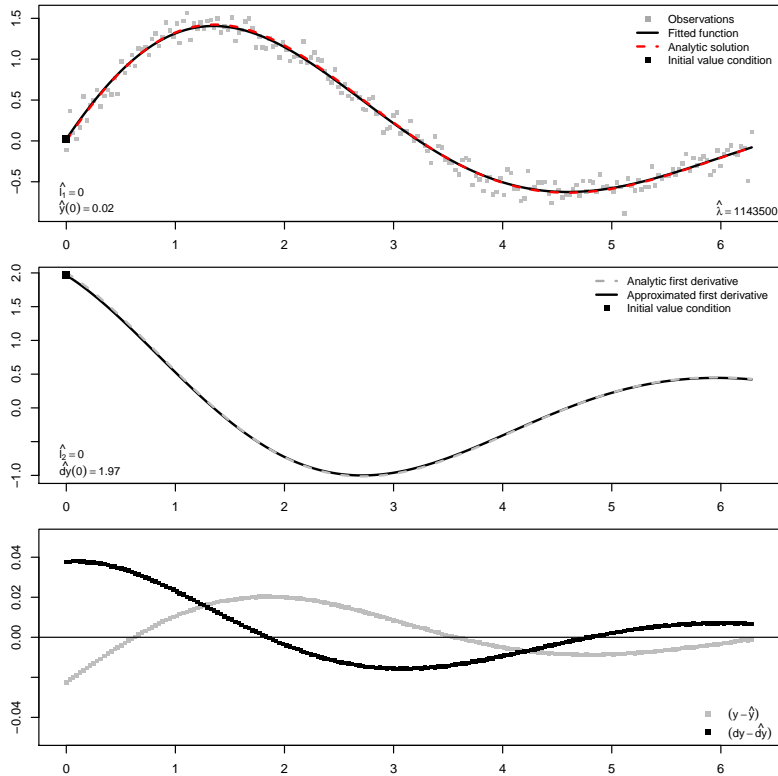
Figure 6.3: *Smoothing with known differential penalties and unknown boundary conditions. The data have been simulated taking into account a system of linear ODEs. The upper panels shows the raw data and the estimated smoothing function, the second plot compares the analytic first derivative of the state function with the approximated one and, finally, the lowest panel shows the residuals between the smoother and the solution of the differential problem. These estimates have been obtained considering a set of 50 cubic basis functions built on equally spaced knots. The initial values have been estimated equal to 0.022 and 1.965 respectively.*

The high value of the optimal smoothing parameter (selected again through an EM-type algorithm) testifies that the differential problem penalizing the smoothing procedure is appropriate to describe the observed dynamics. The same aspect can be evaluated considering the residuals between the analytic solution and the smoothing function plotted in the lowest panel of the figure.

6.3 *The role of the smoothing parameter*

The smoothing parameter λ plays a crucial role in the evaluation of the appropriateness of the differential operator included in the penalty term. This parameter sets the balance between to the data and the fidelity to the DE.

For $\lambda \to \infty$ fidelity to the DE becomes all-important and the data provide only generalized boundary conditions. On the other hand when $\lambda \to 0$ we get a B-spline fit which may lead to trouble because support might be missing for some basis functions.

It is attractive to use the data to set an appropriate value of λ. In principle, all the automatic selection criteria introduced in chapter 3 can be used. Here we introduce a different procedure based on an EM-type algorithm that will be crucial for the mixed model framework we will present in the next chapter.

This algorithm was originally proposed by Schall (1991) as an iterative estimation procedure for mixed effects and dispersion components in generalized linear models with random effects (mixed models). This approach moves from the formal equivalence between penalized least squares and mixed models (we will discuss this topic in more details in the next chapter). The same equivalence can be exploited to select the optimal smoothing parameter.

The optimization algorithm consists of two alternating steps. In the first step the optimal spline coefficients are estimated considering a given value of λ. In a second step the variance components of the errors (σ_ϵ^2) and the penalty variance (σ_p^2) are estimated. Then the value of the smoothing parameter is updated taking it equal to the ratio between the current estimates of these variances: $\lambda = \sigma_\epsilon^2/\sigma_p^2$. The procedure is iterated until convergence of the smoothing parameter.

The two variance components are functions of the effective model dimen-
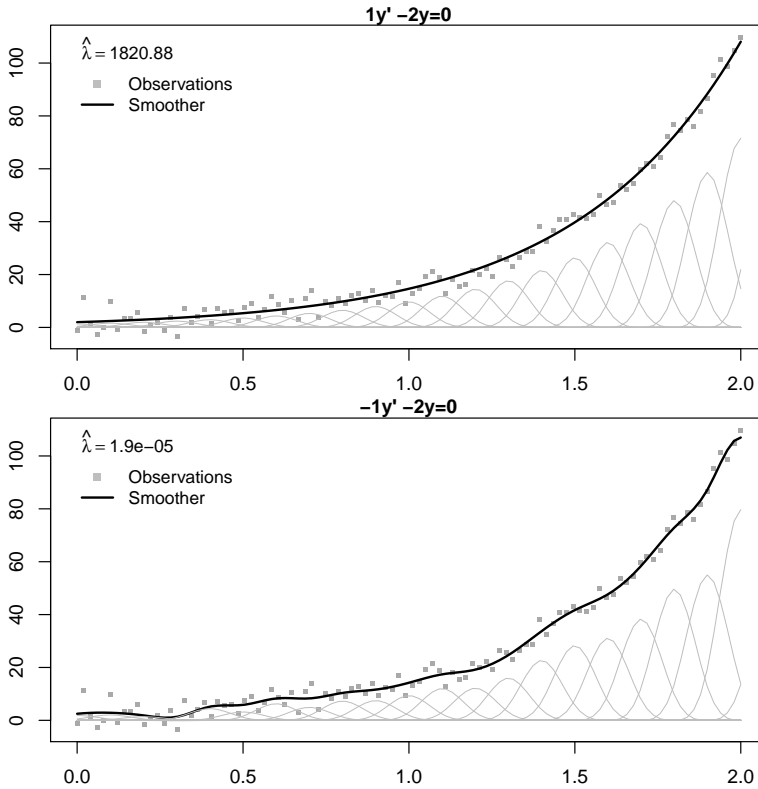
Figure 6.4: *smoothing data described by a linear first order ODE (exponential growth). Each plot shows the results obtained for different DE parameters. In the legend of each panel the optimal smoothing parameter is indicated.*

sion $ED = \text{tr}[B(B^TB + \lambda\breve{V}^T\breve{V})^{-1}B^T]$:

$$\sigma_\epsilon^2 = \frac{\|y - Bc\|^2}{n - ED}$$

$$\sigma_p^2 = \frac{\|\breve{V}(w)c\|^2}{ED} \tag{6.3}$$

Figure 6.4 clarifies the impact of the parameters of the DE on the optimal value of λ selected through an EM-algorithm. If the penalty is close to the DE the optimal λ is large; otherwise it is small.

6.4 Smoothing with unknown differential penalties

It is possible to estimate the unknown DE parameters. A common approach is to write the (analytic or numerical) solution of the equation as explicit function of the unknown parameters to be estimated from the data. This generally

leads to a nonlinear regression problem even if the DE is linear. An alternative approach is to incorporate the differential equation as a penalty in a smoothing framework as suggested by Ramsay and Silverman (2005). The optimal DE parameters and smoothing coefficients are then estimated minimizing the penalized joint log-likelihood. A more convenient estimation procedure exploiting a generalized profiling likelihood criterion has been proposed by Cao and Ramsay (2007). Here we propose a two-stage P-spline-based estimation of the unknown DE parameters.

We start from the consideration that the state function solving the DE describing the observed will be consistent with the data signal if the DE is appropriate. First the signal is extracted from the noisy measurements using a P-spline smoother. The optimal amount of smoothing is determined by the EM-algorithm. The estimated spline coefficients are then used to estimate the parameter vector w through least squares.

Once the differential parameters have been estimated we define the penalty term $\check{V}$ and use it to estimate the optimal spline coefficients:

$$\widehat{c} = (B^T B + \lambda \check{V}^T \check{V})^{-1} B^T y. \tag{6.4}$$

Figure 6.5 shows an example based on 200 observations generated taking into account the analytic solution of following the system of first order ODEs:

$$\begin{cases} x_1' = x_2 \\ x_2' = -x_1 - \frac{1}{2}x_2 \\ x_1(0) = 0; \ x_2(0) = 2. \end{cases}$$

6.4.1 Simulations

We propose two simulation studies related to a first and a second order DE. We want to evaluate the influence of the sample size, the number of subjects and the variance of the noise component on the estimates.

The first order differential problem that we take into account is:

$$y' = y$$
$$y(0) = 2.$$

This IVP has analytic solution $y^*(t) = 2\exp(-x)$. We perform our simulations for 4 different sample sizes: $N = \{20, 50, 100, 200\}$ and 4 noise standard

$\hat{w}_1 = 0$
$\hat{w}_2 = 1$
$\hat{w}_3 = -0.98$
$\hat{w}_4 = -0.52$

$\hat{\lambda} = 1208164$
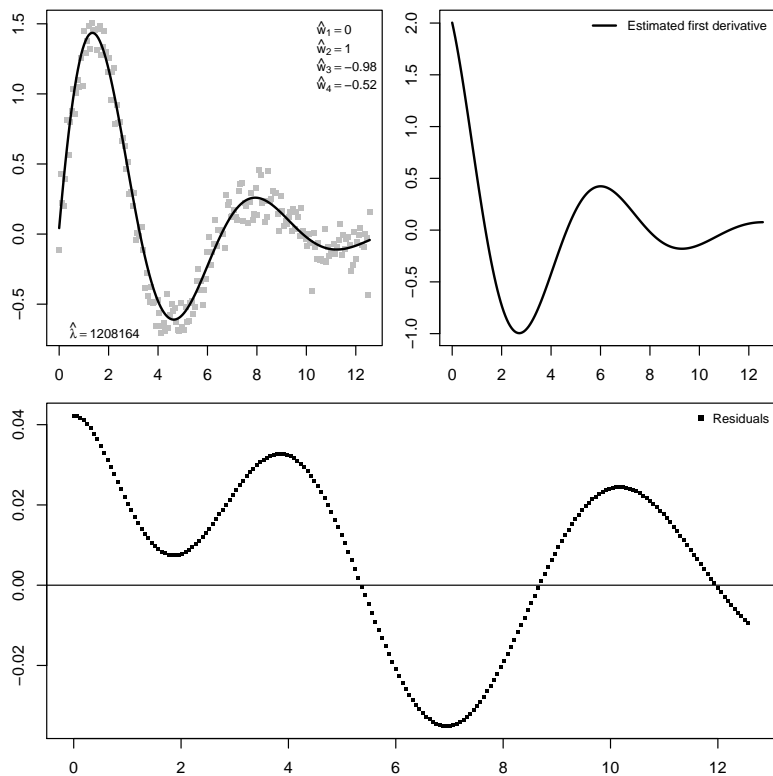
Figure 6.5: *Smoothing with penalty defined by a system of ODEs. The upper panel shows the raw data and the estimates while the lower panel shows the residuals between the estimated smoother and the analytic solution. The estimates have been computed considering 50 B-splines of fourth degree defined on a set of equally spaced knots. The bases used to define the penalty have been evaluated over 500 points.*

deviations $\epsilon \sim N(0, \sigma_i = \{0.05, 0.1, 0.2, 0.3\})$. An example is shown in figure 6.6.

Table 6.1 shows the bias, standard deviation and RMSE of the estimates estimated for 100 replications of the experiment per each sample size and noise standard deviation. The bias of the estimates tends to decrease when the sample size increases for different values of σ. The same is true for the STDs and the RMSEs. On the other hand it is clear that, for a given sample size, increasing the variability of the noise components the quality of the estimates tends to decrease.

We simulate now different data sets described by a second order BVP
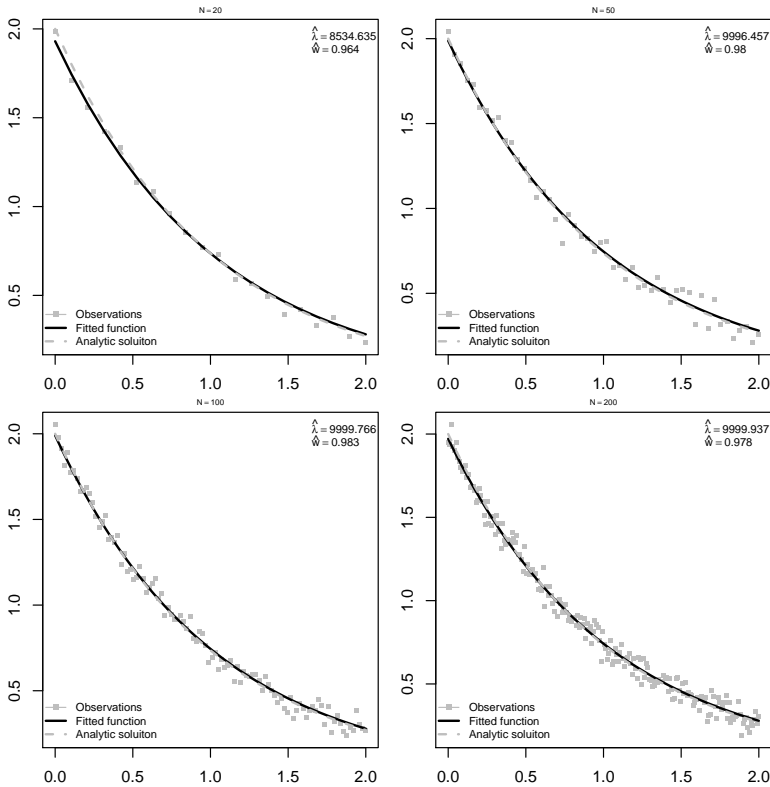
Figure 6.6: *Smoothing with soft constraints defined by a first order ODE with different sample sizes. The data have been simulated taking $\sigma = 0.1$.*

True $w = 1$	$N = 20$			$N = 50$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma = 0.05$	-0.060	0.025	0.065	-0.051	0.028	0.058
$\sigma = 0.10$	-0.066	0.043	0.078	-0.062	0.028	0.068
$\sigma = 0.20$	-0.065	0.090	0.111	-0.068	0.061	0.091
$\sigma = 0.30$	-0.060	0.146	0.158	-0.076	0.090	0.117
True $w = 1$	$N = 100$			$N = 200$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma = 0.05$	-0.040	0.029	0.049	-0.020	0.018	0.027
$\sigma = 0.10$	-0.061	0.029	0.068	-0.058	0.027	0.064
$\sigma = 0.20$	-0.062	0.041	0.075	-0.065	0.032	0.072
$\sigma = 0.30$	-0.064	0.055	0.084	-0.063	0.047	0.078

Table 6.1: *Bias, standard deviations and root mean squared errors of the estimates of the unknown DE parameters of a first order differential problem.*
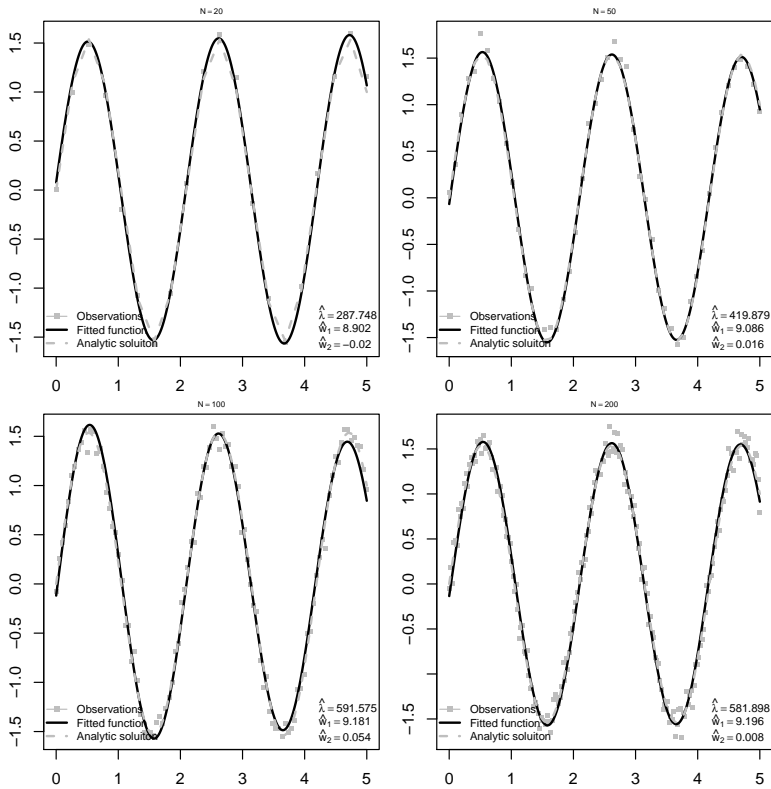
Figure 6.7: *Smoothing with soft constraints defined by a second order order ODE with different sample sizes.*

(harmonic oscillatory system):

$$y'' + 9y = 0$$
$$y(0) = 0, \; y(5) = 1.$$

Figure 6.7 shows the estimates obtained generating noisy data of 4 sample sizes from the analytic solution of this BVP: $y^*(t) = \sin(3t)/\sin(15)$. The error component was considered normally distributed with standard deviation equal to 0.1.

For the simulations we used the same settings discussed in the previous example and the same number of replications. The results are shown in table 6.2. An increasing sample size tends to guarantee more precise, less variable and more robust estimates. Also the amount of variability of the error component has an impact on the quality of the final estimates. As we expected, for a given sample size the robustness, the variability and the pre-

cision of the final estimates improve if the variability of the noise decreases.

	True	$N = 20$ BIAS	STD	RMSE	BIAS	$N = 50$ STD	RMSE
$\sigma = 0.05$	$w_1 = 9$	0.077	0.107	0.132	0.060	0.107	0.122
	$w_2 = 0$	0.026	0.074	0.079	0.035	0.053	0.064
$\sigma = 0.1$	$w_1 = 9$	0.124	0.210	0.243	0.091	0.174	0.195
	$w_2 = 0$	0.051	0.108	0.119	0.060	0.084	0.104
$\sigma = 0.2$	$w_1 = 9$	0.158	0.337	0.371	0.076	0.284	0.293
	$w_2 = 0$	0.115	0.189	0.221	0.124	0.137	0.185
$\sigma = 0.3$	$w_1 = 9$	-0.069	0.464	0.467	0.064	0.393	0.396
	$w_2 = 0$	0.165	0.211	0.267	0.130	0.191	0.231
	True	$N = 100$ BIAS	STD	RMSE	BIAS	$N = 200$ STD	RMSE
$\sigma = 0.05$	$w_1 = 9$	0.056	0.102	0.116	0.058	0.083	0.101
	$w_2 = 0$	0.027	0.050	0.056	0.030	0.051	0.059
$\sigma = 0.1$	$w_1 = 9$	0.087	0.152	0.175	0.089	0.123	0.152
	$w_2 = 0$	0.049	0.068	0.084	0.089	0.075	0.086
$\sigma = 0.2$	$w_1 = 9$	0.129	0.253	0.283	0.091	0.229	0.245
	$w_2 = 0$	0.102	0.136	0.169	0.077	0.116	0.139
$\sigma = 0.3$	$w_1 = 9$	0.066	0.289	0.295	0.138	0.242	0.277
	$w_2 = 0$	0.107	0.152	0.186	0.115	0.134	0.176

Table 6.2: *Bias, standard deviations and root mean squared errors of the estimates of the unknown DE parameters of a second order differential problem.*

6.5   *Application: stomach contractions*

As application of our methodology we propose an example motivated by the analysis of the stomach contraction dynamics during digestion. The entire dataset counts 241 .jpeg images forming a video of the stomach movements during digestion. Figure 6.8 shows four MRIs sampled from the whole set of images. Our aim is to estimate a model that approximately describes this dynamic. We will analyze the data related to a single slice even if our analysis could be repeated for each available profile. The raw data are presented in figure 6.9 showing ten noisy series. The lengths of cross-sections were mea-

sured at 10 positions along the stomach, at each point in time. This gives 10 time series of 300 measurements showing a rather oscillatory behavior. We take into account the dynamics described by the linear second order ODE:

$$y'' + w_1 y' + w_2 y = 0.$$

This model is really simple but gives us the possibility to evaluate two important characteristics of the observed data: the period $(\widehat{\pi})$ of the oscillations and the damping factor $(\widehat{\zeta})$ associated to them. These characteristics are defined by the free response of the system, defined by the DE. It is a damped cosine function. The period is two times the distance between zero crossings. The damping is the ratio of two peaks separated by one period. The period of oscillation is the smallest interval of time in which an oscillating system returns to the sate it was at a time arbitrarily chosen. If the amplitude of the oscillations decreases in time the system is defined "damped". Damping effects are usually induced by frictional forces and are measured by the damping ratio. The damping ratio is the dimensionless measure describing how oscillations decay after that a source of disturbance is applied. These quantities can be expressed in terms of the DE parameters (w_1 and w_2) and are computed as follows:

$$\widehat{\pi} = 2\pi \sqrt{\frac{1}{w_2}}; \ \widehat{\zeta} = \frac{w_2}{2\sqrt{w_1}}.$$

The solid lines in figure 6.9 depict the final estimates obtained using the penalized smoothing procedure described above. The optimal smoothing parameters have been selected through an EM-type algorithm. The estimated DE coefficients and the smoothing parameters per each curve are shown in table 6.3 and in the legends of each panel of figure 6.9.

From table 6.3 we deduce some information about the models that have been estimated per each profile. The optimal λ parameters are high for few waves showing a large period and are relatively small for the rest of the cases. Small smoothing parameters indicate low congruence of the data with the imposed differential penalty. This seems reasonable looking at the data. It is evident from the smoothing curves that the oscillatory behavior of the data presents an asymmetric pattern: the peaks are sharper than the valleys. This asymmetry is not consistent with the linear differential model used as penalty in the smoothing procedure. Probably a nonlinear equation would
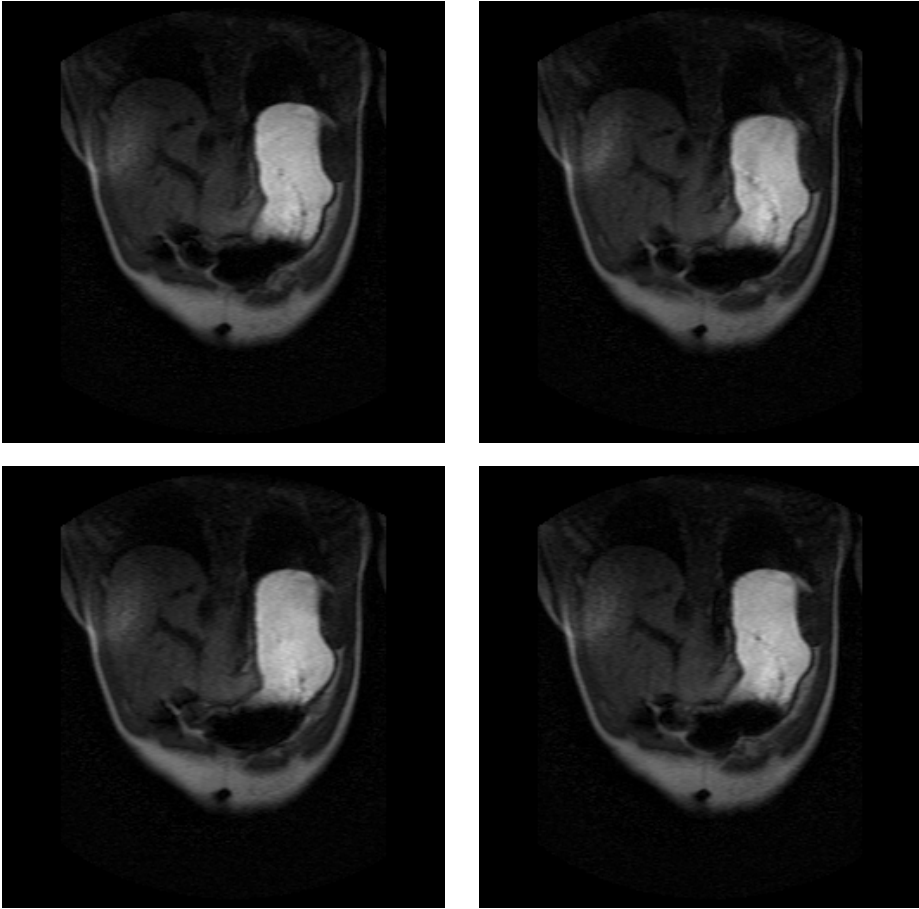
Figure 6.8: *Four abdominal MRIs.*

be more appropriate to describe the dynamics governing the raw data. On the other hand the estimated periods and damping factors seem consistent with what we observe from the measurements. The period is on average around 36 and reaches its maximum for those waves showing a roughly constant signal (see, for example the first wave on the third row of figure 6.9). On the other hand, the estimated period is smaller for those waves showing a more frequent oscillatory behavior. It is the case, for example, of the first and second waves on the first row of figure 6.9 for which the estimated periods (around 26.7) seem appropriate to describe the observed dynamics. The estimated damping factor is always close to zero and it is consistent with the behavior of the data.
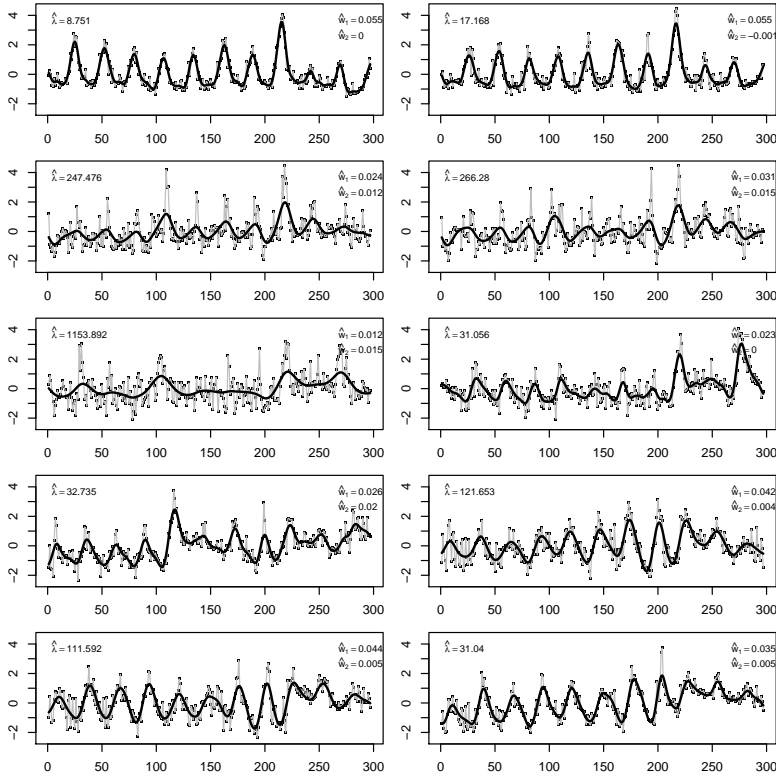
Figure 6.9: *Smoothing of 10 stomach profiles. The black lines represent the estimates obtained with the procedure described above. The estimated DE coefficients, the smoothing parameters and the Lagrange multipliers are shown in the legend of each plot. The estimates have been obtained using 100 fourth order B-splines built on equally spaced knots. The bases defining the penalty term have been evaluated over 500 domain points.*

Wave	$\widehat{w}_1$	$\widehat{w}_2$	$\widehat{\lambda}$	$\widehat{\pi}$	$\widehat{\zeta}$
1	0.055	0.001	8.7	26.7	0.001
2	0.055	-0.001	17.2	26.7	-0.002
3	0.024	0.012	247.5	40.4	0.037
4	0.031	0.015	266.3	35.5	0.043
5	0.012	0.015	1153.9	57.5	0.069
6	0.023	0.001	31.1	41.3	-0.001
7	0.026	0.020	32.7	39.1	0.061
8	0.042	0.004	121.6	30.6	0.010
9	0.044	0.005	111.6	29.9	0.011
10	0.035	0.005	31.0	33.5	0.013

Table 6.3: *Estimated DE components and optimal λ for each stomach wave.*

6.6   *Conclusions*

In this chapter we introduced an ODE-based penalized smoothing approach using a linear combination of basis spline functions for the approximation of the state function in analogy with a B-spline collocation scheme. Penalizing the smoothing procedure by a linear differential operator we imposed consistency between the estimated smoothing function and the state function approximating the solution of the ODE-based penalty.

A smoothing parameter (λ) regulates the fidelity of the final estimates to the DE. This parameter, if selected using an automatic procedure, can be interpreted as a data-based measure of the appropriateness of the differential operator penalty. We suggested to select this parameter through an EM-type algorithm exploiting the mixed model interpretation of penalized least squares.

Initially we have supposed to know exactly the form of the DE governing the observed dynamics. This is an unrealistic hypothesis in real data analyses. For this reason we introduced a two-stage P-spline-based procedure to estimate the unknown DE parameters, parameters following these steps: 1) smooth the raw data using a P-spline smoother able to separate the data trend from the noise component, 2) estimate the unknown DE parameters from the fitted curve through least squares. This simplifies, in our opinion, the estimation task separating the computation of the DE parameters from the data smoothing problem. The performances of this approach have been evaluated through simulations.

We also analyzed a real data example based on a set of time series describing the contraction movements of a stomach cross-section extracted from MRI images. A second order linear differential model has been proposed to describe this dynamics. The estimated differential parameters seem consistent with the data evidence even if the proposed linear differential model appears too simple to describe the observed asymmetric oscillatory behavior. A nonlinear differential equation should give more appropriate results but we did not evaluate yet the possibility to generalize our approach to nonlinear differential equations.

Further generalizations are possible. First of all, it is possible to allow for mixed effect DE parameters (we will discuss this topic in the next chapter).

This generalization is useful when the available observations are grouped according to some factor. On the other hand, it is possible to extend the proposed procedure to smooth data representing system dynamics evolving in more than one dimension combining tensor product smoothing splines and partial differential penalties.

DIFFERENTIAL OPERATOR PENALTIES WITH MIXED PARAMETERS

<div style="text-align: right">7</div>

Smoothing techniques able to take into account the dynamics of systems described by differential problems have been proposed in the literature by many authors. Most of these approaches consider the unknown differential parameters as fixed effects. In many applications it is interesting to study the relationship between measurements grouped according to one or more factors. This leads to a mixed model formulation. In this chapter we propose an ODE-based penalized smoothing spline approach to analyze the measurements of dynamic systems described by differential equations with mixed parameters.

Keywords: *Penalized least squares, smoothing splines, random effects, differential problems.*

7.1 Introduction

Differential equations (DEs) are powerful tools for describing dynamic systems. In practical applications the values of their parameters have to be estimated from observed data. The common approach is to write the solution of the equation as an explicit function of the parameters and fit them to the data. Generally this leads to a nonlinear regression problem even if the DEs are linear. If no explicit solution is available a similar approach can be used based on numerical solution of the DE.

An alternative is to incorporate the differential equation as a penalty in a smoothing framework (Ramsay and Silverman, 2005). In this case a joint penalized log-likelihood has to be minimized in order to estimate a smooth version of the raw data together with the unknown DE parameters. The penalized minimization problem can be simplified exploiting the hierarchy existing between the unknowns and considering the basis coefficients as nui-

sance parameters. This leads to the so called generalized profiling technique proposed by Cao and Ramsay (2007). In the previous chapter we have proposed an alternative P-spline-based two stage procedure.

In many data analyses the data are grouped in some way. Typical examples are measurements on different subjects or collected repeatedly over time. A mixed model is a natural way to analyze grouped data. In the case of dynamic systems it would be a mixed model for the DE parameters.

A classical procedure for the estimation of a random and a fixed component of the DE parameters involves a nonlinear mixed model approach. Given the analytic or numerical solution of the DE, the related parameters can be estimated using nonlinear mixed models. This is the rationale behind many well known software packages like `NONMEM` and `MONOLIX`.

An alternative is a semiparametric approach including the DE in a smoothing framework as a penalty term. The mixed DE parameters are then estimated minimizing the penalized joint log-likelihood or a restricted likelihood function. A generalized profiling approach to solve this estimation task has been proposed by Wang et al. (2012).

In this chapter we extend our discussion about differential penalized smoothing allowing for a mixed model formulation of the differential parameters. We propose to adopt a P-spline-based two-stage procedure for the estimation of the fixed and random components, exploiting the link between penalized regression and mixed model framework, in order to estimate the residual and random effects variance components (Pawitan, 2001).

This chapter is organized as follows: in section 7.2 the relationship between penalized least squares and mixed models is reviewed; in section 7.3 we introduce the two-stage procedure for the estimation of mixed DE parameters analyzing a simulated example; in section 7.4 two simulation studies are discussed in order to evaluate the performance of our proposals while in section 7.6 our approach is used to analyze the stomach MRI data introduced in the previous chapter.

7.2 *Penalized least squares and mixed models*

A linear mixed model is commonly summarized using the following functional form:

$$y = X\beta + Zb + \epsilon, \text{ with } \epsilon \sim N(0, \sigma_\epsilon^2), \ b \sim N(0, \sigma_b^2), \tag{7.1}$$

where y is the n-dimensional response vector, X is a $n \times m$ model matrix for the m-dimensional fixed effect vector β and Z is a $n \times k$ model matrix for the k-dimensional random effect vector b. The vector ϵ represents a Gaussian noise component having zero mean and constant variance σ_ϵ^2. The random coefficients are also supposed to be normally distributed but with their own variance σ_b^2 and zero mean.

One way to introduce this last hypothesis in the regression framework is through penalized least squares. The requirement of b coefficients consistent with a Gaussian probability law introduces a restriction on their codomain. These coefficients are defined in $]-\infty, \infty[$ but values around the mean (zero) are supposed to be more probable than extreme values (in module). Combining the log-likelihoods of the measurement vector y and the random effect b and dropping the terms not involving the mean parameters β and b we obtain:

$$\log(L) = -\frac{1}{2\sigma^2} \sum_{ij} (y_{ij} - \beta_i x_{ij} - b_j z_{ij})^2 - \frac{1}{2\sigma_b^2} b^T b. \tag{7.2}$$

Given σ_ϵ^2 and σ_b^2, the estimates of b and β are the minimizers of the penalized sum of squares:

$$\sum_{ij} (y_{ij} - \beta_i x_{ij} - b_j z_{ij})^2 + \lambda b^T b,$$

where $\lambda = \sigma_\epsilon^2 / \sigma_b^2$.

The joint log-likelihood based on the observations y and the random effects b considering unknown variance components is given by:

$$\begin{aligned} \log(L) = &-\tfrac{1}{2} \log|\sigma_\epsilon I| - \tfrac{1}{2}(y - X\beta - Zb)^T (\sigma_\epsilon^2 I)^{-1} (y - X\beta - Zb) \\ &- \tfrac{n}{2} \log(\sigma_b^2) - \tfrac{1}{2\sigma_b^2} b^T b. \end{aligned} \tag{7.3}$$

The estimates of β and b are the solution of:

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda I \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix}, \tag{7.4}$$

with $\lambda = \frac{\sigma_\epsilon^2}{\sigma_b^2}$. The random effects are estimated solving the following equation:

$$(Z^T Z + \lambda I)b = Z^T(y - X\widehat{\beta}). \tag{7.5}$$

The variance component are estimated optimizing the λ parameter through an EM-type algorithm. For a given value of the shrinkage parameter λ_0 the effective model dimension ED can be computed evaluating the trace of the hat matrix (Hastie and Tibshirani (1990) and chapter 3), then the noise variance follows from:

$$\sigma_\epsilon^2 = \frac{\|y - X\beta - Zb\|^2}{n - ED},$$

while the variance of the random effects is equal to:

$$\sigma_b^2 = \frac{b^T b}{ED}.$$

Using these values we compute a new λ parameter as the ratio of the two variance components leading to a new value of the model effective dimension. The variance components minimizing (7.3) are estimated iterating these steps until convergence as we shown in chapter 6 (Pawitan, 2001).

7.3   *Two stage estimate of mixed DE parameters*

Suppose we observe a set of n noisy measurements taken over k subjects with a nonlinear subject-specific trend appropriately described by the state function solving a certain differential equation with a known form but unknown parameters.

To clarify the problem settings we consider the example shown in figure 7.1. This figure depicts 100 measurements taken over 10 "subjects". The data have been simulated considering the following first order differential problem:

$$\frac{dy}{dt} + (\beta + b_j)y(t) = 0, \ t \in [0, 2]$$
$$y(0) = 2,$$

The DE parameters are the sum of a fixed (population shared) effect β and a random (subject-specific) effect $(b_j) \sim N(0, I\sigma_b^2)$ with $\sigma_b^2 = 0.01$. In the simulation a multivariate Gaussian noise with zero mean and a constant variance

equal to 0.07 has been added to the numerical solution of the differential model. The fixed effects have been considered equal to $\beta_1 = 1$.

Our aim is to estimate the unknown fixed and random DE parameters starting from the subject specific measurements. We propose a two-stage approach to get these estimates. We suggest to proceed as follows: 1) fit a P-spline smoother for each subject specific measurement set such that to separate the trend component from the random noise, 2) use the estimated penalized spline coefficients to approximate the state function and its derivatives. The unknown mixed DE and the variance components are estimated through a linear mixed model approach. These estimates can be conveniently obtained exploiting the relationship between LMM and penalized least squares discussed above.

For the example that we are considering the linear model to be estimated is:

$$y = X\beta + Zb + \epsilon, \ \ \epsilon \sim N(0, \sigma_\epsilon^2 I), \ b \sim N(0, \sigma_b^2 I),$$

with y a (1000×1) vector of $\widehat{y}'$ estimated over all the subjects, X a (1000×1) matrix containing $\widehat{y}$ estimated for each subject and Z a (1000×10) block diagonal matrix with each block representing the approximated subject-specific $\widehat{y}$:

$$Z = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & Z_m \end{bmatrix}.$$

The fixed and random effects are estimated solving:

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda I \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix}.$$

The estimates of the random coefficients depend on the value of the shrinkage parameter. In the mean time this parameter is used to estimate the variance of the random effect. These unknowns are estimated together optimizing the value of λ through an EM-algorithm in analogy with what we discussed in the previous section. On the other hand the error variance cannot be estimated in this step given that we are using a smooth version of the data to approximate the DE parameters. This variance component is obtained in
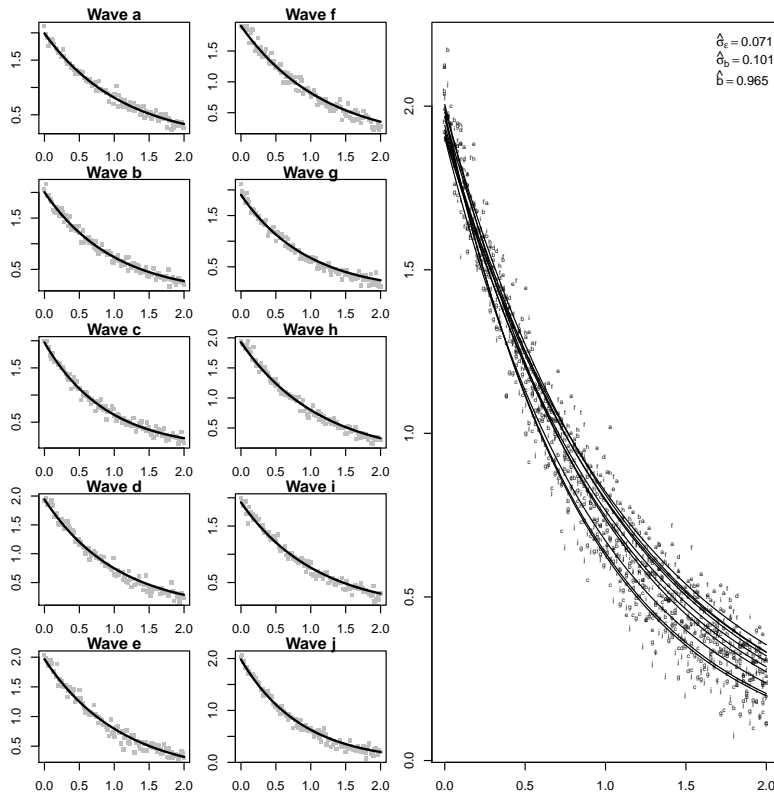
Figure 7.1: *Simulated data driven by a first order ODE with random period and damping parameters. The left panels show the subject specific profiles separately while the right panel shows the entire dataset (100 measurements over 10 subjects). The gray dots in the left panels show the subject-specific observations and the solid lines represent the data trend extracted penalizing a smoothing procedure by approximating the differential through a collocation scheme. The right panel shows the entire dataset and the subject-specific smoothing functions. In the legend of the right plot the estimates of the fixed effect and of the variance components are listed.*

a second moment when we smooth the raw data taking into account the estimated differential operator.

Indeed, once the fixed and random DE parameters have been estimated, we include the (linear) differential operator in a penalized smoothing framework in order to obtain subject-specific smooth curves (solid lines in figure 7.1):

$$\min_{\alpha} \|y - B\alpha\|^2 + \delta \|\check{V}(w)\alpha\|^2 \qquad (7.6)$$

where B is a B-spline matrix, α is a vector of spline coefficients to be estimated, δ is a smoothing parameter and $\check{V}(w)$ is a (linear) differential operator defined by a (linear) combination (with coefficients $w_j = \beta + b_j$) of basis spline matrices evaluated on an enlarged set of collocation points (analogously to what we discussed in chapter 6). The smoothing parameter δ regulates the balancing between the goodness of fit criterion and the adherence of the smoothing function to the numerical solution of the differential problem. On the other hand it represents the ratio between the noise variability and the variability of the penalty term $\|\check{V}(w)\alpha\|^2$ and so, optimizing it, we estimate σ_ϵ^2. The results are in table 7.1. The estimates are close to the true ones.

	True	Estimated	$\beta + b_j$	$\widehat{\beta} + \widehat{b}_j$
			0.945	0.930
β	1.000	0.986	0.971	0.960
			0.787	0.742
			1.026	0.993
			0.989	0.971
σ_ϵ	0.070	0.074	0.967	0.947
			0.912	0.889
			0.957	0.927
			1.106	1.081
σ_b	0.100	0.106	0.805	0.802

Table 7.1: *Estimates obtained for the example shown in figure 7.1.*

7.4 Simulations

We propose two simulation studies related to a first and a second order DE. We want to evaluate the influence of the sample size, the number of subjects and the variance of the noise component on the estimates.

Consider the ODE introduced in the previous section:

$$\frac{dy}{dt} + (\beta + b_j)y(t) = 0, \ t \in [0, 2]$$
$$y(0) = 2,$$

having the analytic solution $y^* = 2\exp((\beta + b_j)t)$. We simulate subject-specific measurements and add a noise with $\sigma_\epsilon = 0.07$. The random effects are drawn from a Gaussian distribution with zero mean and standard deviation $\sigma_b = 0.1$. We take into account $n = \{20, 50, 100, 200\}$ measurements recorded over $m = \{10, 20\}$ and a fixed effect equal to $\beta = 1$. See figure 7.1 for a typical result.

Table 7.2 summarizes the results based on 100 replications. According to

			$n = 20$			$n = 50$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
$m = 10$	$\beta = 1$	-0.0667	0.0298	0.0730	-0.0582	0.0303	0.0655
	$\sigma_b = 0.1$	-0.0035	0.0040	0.0053	-0.0018	0.0042	0.0046
	$\sigma_\epsilon = 0.07$	0.0005	0.0017	0.0018	0.0006	0.0013	0.0014
			$n = 100$			$n = 200$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
	$\beta = 1$	-0.0589	0.0292	0.0656	-0.0356	0.0318	0.0477
	$\sigma_b = 0.1$	-0.0042	0.0039	0.0057	0.0005	0.0044	0.0044
	$\sigma_\epsilon = 0.07$	0.0002	0.0025	0.0028	0.0001	0.0006	0.0006
			$n = 20$			$n = 50$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
$m = 20$	$\beta = 1$	-0.0711	0.0201	0.0738	-0.0637	0.0198	0.0666
	$\sigma_b = 0.1$	-0.0025	0.0029	0.0038	-0.0018	0.0028	0.0033
	$\sigma_\epsilon = 0.07$	0.0005	0.0019	0.0020	0.0005	0.0011	0.0012
			$n = 100$			$n = 200$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
	$\beta = 1$	-0.0630	0.0238	0.0673	-0.0349	0.0201	0.0402
	$\sigma_b = 0.1$	-0.0036	0.0031	0.0047	-0.0004	0.0029	0.0029
	$\sigma_\epsilon = 0.07$	0.0004	0.0027	0.0028	0.0002	0.0005	0.0005

Table 7.2: *Bias, standard deviations and RMSE of the estimated fixed effect and variance components obtained over 100 replicates of an experiment based on a first order DE with random parameters. The data have been generated considering different sample sizes and different numbers of subjects. These results have been obtained considering B splines built on 20 knots for each sample size. The bases included in the penalty terms have been evaluated over a set of 500 collocation points.*

the results shown in table 7.2 the quality of the estimates seems to improve

if we consider a larger sample size per each subject and, for a given sample size, if we consider a larger number of subjects. The fixed effect seems to be slightly underestimated. The bias of the estimated variance components appears really small and tends to decrease when the sample size or the number of subjects considered increases. The associated standard deviations and root mean square errors show the same behavior.

In order to evaluate the impact of the variability of the noise component on the final estimates we consider the results summarized in table 7.3. These results have been obtained considering $n = 50$ observations simulated for $m = \{10, 20\}$ subjects and adding a Gaussian noise with $\sigma_\epsilon \in \{0.07, 0.1, 0.15\}$ to the solution of the DE. As we expected, the quality of the estimates tends to decrease increasing the variability of the simulated noise component even if this effect seems negligible. Also the variability of the estimates tends to be higher when the variability of the noise component increases but these increments are moderate.

True	$m = 10$ BIAS	STD	RMSE	$m = 20$ BIAS	STD	RMSE
$\beta = 1$	-0.0582	0.0303	0.0655	-0.0614	0.0287	0.0677
$\sigma_b = 0.1$	-0.0018	0.0042	0.0046	-0.0028	0.0041	0.0049
$\sigma_\epsilon = 0.07$	6.0e-04	0.0013	0.0014	9.0e-04	0.0018	0.0020
True	$m = 10$ BIAS	STD	RMSE	$m = 20$ BIAS	STD	RMSE
$\beta = 1$	-0.0637	0.0198	0.0666	-0.0672	0.0190	0.0698
$\sigma_b = 0.1$	-0.0018	0.0028	0.0033	-0.0023	0.0028	0.0036
$\sigma_\epsilon = 0.1$	5.0e-04	0.0011	0.0012	5.0e-04	0.0019	0.0020
True	$m = 10$ BIAS	STD	RMSE	$m = 20$ BIAS	STD	RMSE
$\beta = 1$	-0.0724	0.0340	0.0799	-0.0679	0.0207	0.0710
$\sigma_b = 0.1$	-0.0023	0.0036	0.0043	-5.0e-04	0.0027	0.0033
$\sigma_\epsilon = 0.15$	-3.0e-04	0.0042	0.0042	3.0e-04	0.0053	0.0053

Table 7.3: *Bias, standard deviations and RMSE of the estimated fixed effect and variance components obtained over 100 replicates of an experiment based on a first order DE with random parameters. The data have been generated considering different degrees of variability of the noise component and 50 observations measured for $m = \{10, 20\}$ subjects.*

A similar simulation study can be used to evaluate the performance of the proposed procedure dealing with a set of grouped data described by the

second order ODE:

$$\frac{d^2y}{dt^2} + (\beta_1 + b_{1,j})\frac{dy}{dt} + (\beta_2 + b_{2,j})y(t) = 0, \ t \in [0,5]$$
$$y(0) = 0, \ y(5) = 1,$$

where β_1 and β_2 are the fixed effects and b_1, b_2 are the random DE parameters. For the simulations we set $\beta_1 = 0.5$, $\beta_2 = 0.5$ and $\{b_{1,j}, b_{2,j}\} \sim MVN((0,0), \sigma_b = \text{diag}(0.1, 0.1))$ and solve each DE numerically through a B-spline collocation procedure. A noise $\epsilon \sim N(0, \sigma = 0.1)$ is added to the approximated solution. To investigate the performance of the proposed P-spline-based two stage-approach we generate different data sets considering various sample sizes and numbers of subject-specific dynamics. In analogy with what we did for the first order example above, we consider $n = \{20, 50, 100, 200\}$ measurements taken over $m = \{10, 20\}$ subjects.

A possible data configuration is shown in figure 7.2. Table 7.4 summarizes the results based on 100 replications. Taking constant the number of subjects considered in the analysis the quality of the estimates tends to increase when the number of measurements per subject increases. This can be evaluated looking at the behavior of bias, standard deviation and RMSE of the final estimates. On the other hand, if more subjects are included, for a given number of subject-specific measurements, we obtain better estimates of the variance components.

It is interesting to evaluate how the noise component influences the estimates. We simulate $n = 50$ observations measured over $m = \{10, 20\}$ subjects using consider the same values for the fixed effects, $\{b_1, b_2\} \sim MVN((0,0), \sigma_b = \text{diag}(0.1, 0.1))$ and repeat this simulations for $\epsilon \sim N(0, \sigma = \{0.1, 0.15, 0.2\})$. Table 7.5 summarizes the results obtained for 100 replications. The bias of the final estimates tends to increase if the noise components is more variable even if this effect is not so evident. The variability and the root mean squared errors of the estimates show a similar behavior. On the other hand, the estimated values seem consistent to the simulated ones for each level of σ_ϵ.

| | | $n = 20$ | | | | $n = 50$ | |
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
$m = 10$	$\beta_1 = 0.5$	-0.0185	0.0368	0.0410	-0.0205	0.0325	0.0383
	$\beta_2 = 0.5$	-0.0138	0.0341	0.0366	-0.0184	0.0295	0.0347
	$\sigma_{b_1} = 0.1$	-0.0016	0.0044	0.0047	-0.0013	0.0049	0.0050
	$\sigma_{b_2} = 0.1$	-0.0025	0.0037	0.0044	-0.0011	0.0059	0.0060
	$\sigma_\epsilon = 0.1$	0.0019	0.0045	0.0049	0.0003	0.0022	0.0022
		$n = 100$				$n = 200$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
	$\beta_1 = 0.5$	-0.0237	0.0335	0.0409	-0.0063	0.0341	0.0345
	$\beta_2 = 0.5$	-0.0230	0.0319	0.0392	-0.0025	0.0348	0.0347
	$\sigma_{b_1} = 0.1$	-0.0011	0.0048	0.0049	-0.0014	0.0043	0.0046
	$\sigma_{b_2} = 0.1$	-0.0013	0.0045	0.0047	-0.0012	0.0039	0.0041
	$\sigma_\epsilon = 0.1$	-0.0001	0.0014	0.0014	1.2e-05	0.0011	0.0011
		$n = 20$				$n = 50$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
$m = 20$	$\beta_1 = 0.5$	-0.0216	0.0225	0.0311	-0.0235	0.0218	0.0320
	$\beta_2 = 0.5$	-0.0063	0.0224	0.0232	-0.0131	0.0244	0.0276
	$\sigma_{b_1} = 0.1$	-0.0007	0.0034	0.0035	9.5e-05	0.0035	0.0035
	$\sigma_{b_2} = 0.1$	-0.0017	0.0031	0.0035	-0.0014	0.0029	0.0032
	$\sigma_\epsilon = 0.1$	0.0026	0.0081	0.0085	0.00010	0.0022	0.0022
		$n = 100$				$n = 200$	
	True	BIAS	STD	RMSE	BIAS	STD	RMSE
	$\beta_1 = 0.5$	-0.0214	0.0217	0.0305	-0.0088	0.0199	0.0217
	$\beta_2 = 0.5$	-0.0227	0.0202	0.0304	-0.0084	0.0218	0.0232
	$\sigma_{b_1} = 0.1$	-0.0009	0.0030	0.0031	-0.0007	0.0036	0.0036
	$\sigma_{b_2} = 0.1$	-0.0008	0.0026	0.0028	-5.2e-05	0.0036	0.0036
	$\sigma_\epsilon = 0.1$	-0.0002	0.0015	0.0015	-5.6e-05	0.0009	0.0009

Table 7.4: *Bias, standard deviations and RMSE of the estimated fixed effect and variance components obtained over 100 replicates of an experiment based on a second order DE with random parameters. The data have been generated considering different sample sizes and different number of subjects.*
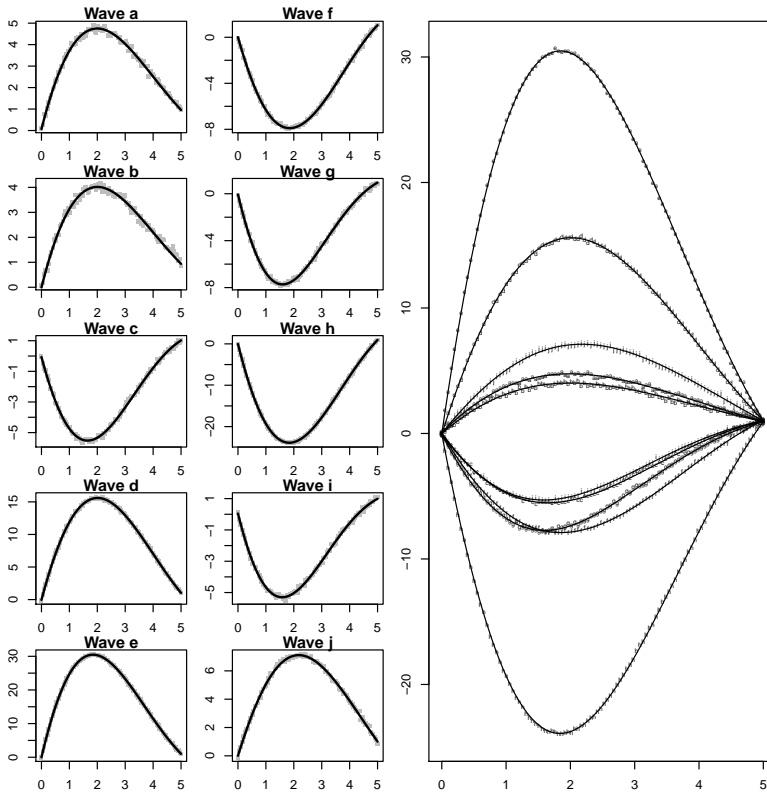
Figure 7.2: *Simulated data driven by a second order ODE with random period and damping parameters. The left panels show the subject specific profile separately while the right panel shows all the data together.*

7.5   *A real data example*

As real data example we consider the MRI data already discussed in chapter 6. The whole dataset counts 241 `.jpeg` MRI scans of the abdominal region forming a video of the stomach contractions during digestion. The lengths of cross-sections were measured at 10 positions along the stomach, at each point in time. This gives 10 time series of 300 measurements showing a rather oscillatory behavior. We take into account the dynamics described by the second order ODE:

$$\frac{d^2y}{dt^2} + (\beta_1 + b_{1,j})\frac{dy}{dt} + (\beta_2 + b_{2,j})y(t) = e_j,$$

where β_i and b_i indicate the fixed and random effects respectively and e_j represents a subject-specific driving force. As we mentioned in chapter 6, this equation describes a (damped) oscillatory behavior. Allowing for mixed DE

	$m = 10$			$m = 20$		
True	BIAS	STD	RMSE	BIAS	STD	RMSE
$\beta_1 = 0.5$	-0.0205	0.0325	0.0383	-0.0235	0.0218	0.0320
$\beta_2 = 0.5$	-0.0184	0.0295	0.0347	-0.0131	0.0244	0.0276
$\sigma_{b_1} = 0.1$	-0.0013	0.0049	0.0050	9.5e-05	0.0035	0.0035
$\sigma_{b_2} = 0.1$	-0.0011	0.0059	0.0060	-0.0014	0.0029	0.0032
$\sigma_\epsilon = 0.1$	0.0003	0.0022	0.0022	0.00010	0.0022	0.0022
	$m = 10$			$m = 20$		
True	BIAS	STD	RMSE	BIAS	STD	RMSE
$\beta_1 = 0.5$	-0.0280	0.0328	0.0430	-0.0243	0.0234	0.0336
$\beta_2 = 0.5$	-0.0368	0.0331	0.0494	-0.0354	0.0227	0.0420
$\sigma_{b_1} = 0.1$	-0.0015	0.0042	0.0045	-0.0008	0.0036	0.0037
$\sigma_{b_2} = 0.1$	-0.0011	0.0048	0.0050	9.1e-06	0.0041	0.0040
$\sigma_\epsilon = 0.15$	-0.0003	0.0045	0.0045	-0.0002	0.0042	0.0042
	$m = 10$			$m = 20$		
True	BIAS	STD	RMSE	BIAS	STD	RMSE
$\beta_1 = 0.5$	-0.0306	0.0328	0.0447	-0.0271	0.0212	0.0344
$\beta_2 = 0.5$	-0.0667	0.0406	0.0780	-0.0546	0.0290	0.0617
$\sigma_{b_1} = 0.1$	-0.0012	0.0048	0.0049	-0.0005	0.0034	0.0035
$\sigma_{b_2} = 0.1$	0.0037	0.0098	0.0104	0.0024	0.0058	0.0073
$\sigma_\epsilon = 0.2$	-0.0003	0.0092	0.0092	0.0003	0.0079	0.0079

Table 7.5: *Bias, standard deviations and RMSE of the estimated fixed effect and variance components obtained over 100 replicates of an experiment based on a second order DE with random parameters. The data have been generated considering different degrees of variability of the noise component and 50 observations measured over $m = \{10, 20\}$ subjects.*

parameters gives us the possibility to estimate "population-shared" parameters (i.e. the parameters that describe the average dynamics over the entire area we are considering) and "profile-specific" parameters summarizing the contractions over small parts of the considered region. In particular we are interested in estimating the population and the profile-specific periods and damping ratios (we have already introduced these indexes in chapter 6):

$$\widehat{\pi} = 2\pi \sqrt{\tfrac{1}{\beta_2}} \quad \widehat{\pi}_j = 2\pi \sqrt{\tfrac{1}{(\beta_2 + b_{2,j})}}$$

$$\widehat{\zeta} = \tfrac{\beta_2}{2\sqrt{\beta_1}} \qquad \widehat{\zeta}_j = \tfrac{(\beta_2 + b_{2,j})}{2\sqrt{(\beta_1 + b_{1,j})}}.$$

The fitted smoothers are shown in figure 7.3. These results have been obtained using 100 basis functions of fourth degree built on equally spaced knots. The bases used to build the differential operator penalty have been evaluated over a grid of 500 collocation points. Each panel of the plot shows
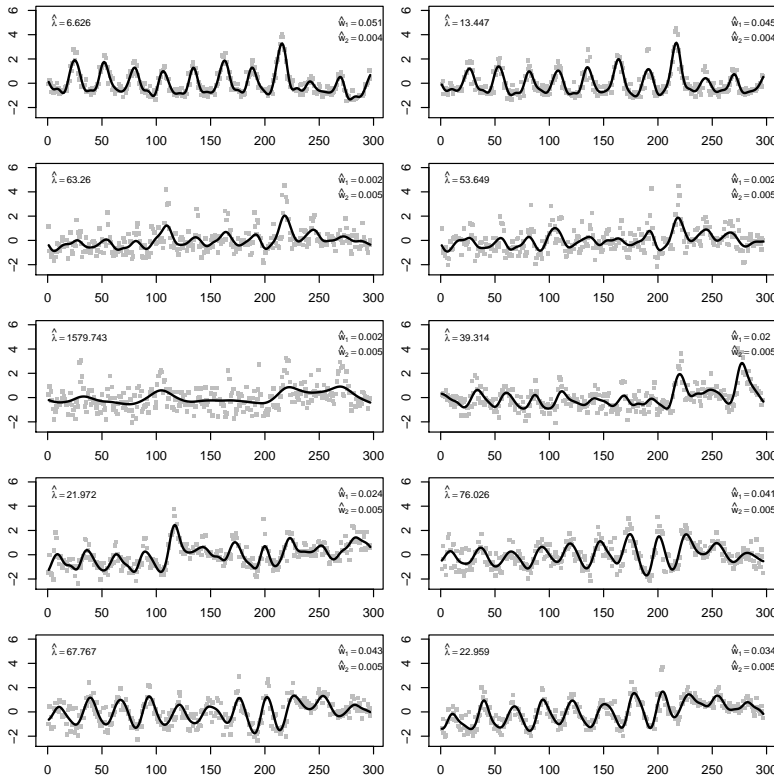
Figure 7.3: *Smoothing of 10 MRI stomach profiles considering random differential parameters. The final estimates (solid lines) have been obtained considering spline smoothers penalized by a damped oscillation DE.*

the estimated optimal smoothing parameter and the estimated DE parameters $w_1 = \widehat{\beta_1 + b_{1,j}}$ and $w_2 = \widehat{\beta_2 + b_{2,j}}$. These results are also summarized in table 7.6.

The estimated periods and damping ratios seem consistent with the observed data. The fixed periodic effect has been found approximately equal to 32.8 while a fixed damping effect equal to 0.012 has been estimated. The estimated variance components are all rather small. The profile damping factors are close to zero while the profile-specific periods tend to be higher for profiles with flat behaviors. The optimal smoothing parameters (δ) are higher for flat dynamics and smaller for more oscillatory ones testifying a rather small adhesion of the smoothing curves to the differential equation supposed to describe the data. This is reasonable if we look carefully at the data behavior. The estimated peaks appear asymmetric and this is not consistent with

	Wave	b_1	b_2	w_1	w_2	$\widehat{\delta}$	$\widehat{\pi}_j$	$\widehat{\zeta}_j$
$\widehat{\beta_1} = 0.0367$								
	1	0.0147	-2.0e-04	0.0514	0.0043	6.626	27.69	0.0096
$\widehat{\beta_2} = 0.0046$	2	0.0079	-0.0003	0.0446	0.0042	13.44	29.73	0.0100
	3	-0.0349	1.70e-6	0.0017	0.0046	63.26	152.1	0.0555
$\widehat{\pi} = 32.7986$	4	-0.0342	1.44e-6	0.0024	0.0046	53.65	126.6	0.0462
	5	-0.0347	-1.25e.6	0.0019	0.0046	1579.7	140.8	0.0513
$\widehat{\zeta} = 0.0120$	6	-0.0165	-6.0e-05	0.0202	0.0045	39.31	44.21	0.0159
	7	-0.0127	0.0004	0.0239	0.0049	21.97	40.59	0.0160
$\widehat{\sigma_\epsilon^2} = 0.2439$	8	0.0042	2.0e-05	0.0409	0.0046	76.02	31.03	0.0113
	9	0.0065	1.0e-05	0.0432	0.0046	67.76	30.21	0.0110
$\widehat{\sigma_{b_1}^2} = 0.0004$	10	-0.0031	6.0e-05	0.0336	0.0046	22.95	34.28	0.0127
$\widehat{\sigma_{b_2}^2} = 2.0e-6$								

Table 7.6: *Estimated period π and dumping ratio ζ for each profile. The table lists the estimated values of the DE parameters $\widehat{\beta_1 + b_1}$ and $\widehat{\beta_2 + b_2}$ as well as the optimal smoothing parameters $\widehat{\lambda}$.*

the linear second order differential law used to build the penalty term of the smoothing procedure. Probably it could be possible to obtain higher smoothing parameters and a more appropriate description of the observed dynamics using a nonlinear differential law.

7.6   *Conclusions*

In this chapter we have discussed a generalization of the framework introduced in chapter 6. A smoothing procedure driven by differential operators with mixed parameters and a two-stage estimation procedure have been introduced. In the first step the relationship between the raw data and the state function is summarized through a P-spline smoother. In the second step the estimated spline coefficients are used to reconstruct the differential operator linking the derivatives and the state function in such a way to consider the unknown differential parameters as the sum of fixed and random effects. The linear differential operator estimated in this last step is then introduced as a penalty term in a smoothing procedure.

Even if a more common mixed model estimation procedures (i.e. ML or REML) can easily be included in our procedure, we proposed to exploit the the link between penalized regression and mixed models. The formal equivalence between these regression techniques allows for a practical estimation

and interpretation of the variance of the random effects. Indeed we suggest to estimate this variance component optimizing the shrinkage parameter λ balancing between the residual sum of squares criterion and the distributional hypothesis formulated on the random parameter(s) (induced by a penalty term).

The variance of the noise component is estimated using the estimated subject-specific differential operators as penalty term in a smoothing procedure. This value is obtained optimizing the smoothing parameter (δ) balancing between the fidelity of the final estimates to the data and to a collocation scheme. This smoothing parameter has also another convenient interpretation clarifying the appropriateness of the hypothesized ODE in describing the data dynamics.

The performance of the proposed procedure has been evaluated considering both simulated and real data. A simulation studies based on first and second order differential problems have been discussed. Through simulations we have evaluated the influence of the data characteristics on the estimates obtained using the two stage procedure. We have found that, increasing the number of measurements available for each subject and/or the number of considered subjects, the bias, the standard deviation and the root mean squared errors of the estimated fixed effects and variance components tend to decrease. On the other we found that, even if a moderate number of observations and subjects is available, the quality of the estimates obtained using the proposed procedure is quite high.

We also discussed a real data example, stomach contractions during digestion. In this case we used a linear second order differential equation with mixed parameters to investigate the global and profile-specific contraction characteristics. The estimated DE parameters have been found consistent with the observed data behavior. On the other hand rather small optimal smoothing parameters have been found using an EM-type procedure. This is probably due to the fact that the considered linear DE is not appropriate to describe the asymmetric oscillations characterizing the observed data behavior.

One of the major limitation of our discussion lies in the fact that only linear differential problems have been considered. From a theoretical point of view the two stage procedure for the estimation of the unknown DE param-

eters can be easily adapted in order to deal with nonlinear mixed effect models. The definition of a nonlinear differential operator penalty can be obtained adopting some linearization strategy of the differential problem. Smoothing procedures penalized by nonlinear differential operators will represent one of the major topic of our future research.

DIFFERENTIAL PENALIZED SMOOTHING IN TWO DIMENSIONS

8

Partial differential equations (PDEs) are common tools to describe dynamic systems evolving in more than one dimension. In real data applications the parameters of suitable differential models are usually unknown and need to be estimated from the available measurements of a given phenomenon. In this chapter we propose a two-dimensional smoothing approach including a partial differential penalty for the analysis of two-dimensional data approximatively described by PDEs. We also propose a P-spline-based two-stage procedure for the estimation of the unknown DE parameters.

Keywords*: Partial differential equations, tensor product B-spline collocation, tensor product P-splines, penalized smoothing.*

8.1 Introduction

Partial differential equations (PDEs) are commonly used to describe dynamic systems evolving in more than one dimension. In many circumstances, it is difficult to have a precise knowledge of the differential law governing an observed dynamics. Fortunately many two-dimensional phenomena that are described, at least approximately, by partial differential equations can be measured. For this reason it is convenient to combine the observed information with the theoretical behavior of the phenomenon.

In this chapter we propose a differential operator smoothing procedure for the analysis of data which dynamic is well described by a partial differential equations. We follow the framework defined by Ramsay and Silverman (2005) based on the definition of a penalized smoothing procedure involving B-spline functions. We suggest to include a penalty term in the smoothing framework approximated by a tensor product B-spline collocation scheme.

As we already discussed for ODEs, it is not realistic to hypothesize a perfect knowledge of the differential law governing the observed dynamic and it becomes necessary to reconstruct this mathematical model using the available observations. We propose to estimate the unknown parameters defining the differential penalty term, using a two-stage approach based on tensor product P-splines. In a first step the two-dimensional data trend is separated from the noise of the observed measurements through a suitable non-isotropic smoother. Then the optimal spline coefficients estimated in the first step are used to approximate the functions defining the PDE and to estimate the best differential parameters through least squares. Once the differential operator has been reconstructed it is used as penalty in the smoothing procedure. The balance between the fidelity to the data and to the approximated solution of the PDE is tuned by a smoothing parameter.

This chapter is organized as follows: section 8.2 introduces the PDE-based penalized smoothing approach considering known differential operators, in section 8.3 we abandon the unrealistic hypothesis of knowledge of the differential law and introduce a two-stage procedure for the estimation of the unknown DE parameters while in section 8.4 the performance of the proposed procedure is evaluated through simulations.

8.2   PDE-based penalized smoothing

Suppose we observe a set of data (z) which dynamics is well described by a partial differential equation. What happens if our knowledge about the differential equation describing the observed dynamic is not precise or even not correct? In this case the solution of the differential problem would provide an inadequate description of the observed measurements. Unfortunately, this is not a remote eventuality given that small changes of the parameters or of the conditions defining the differential problem can lead to really different numerical solutions

It is possible to overcome this problem forcing the solution of the DE we have in mind to give an appropriate description of the data. This can be achieved using a smoothing approach penalized by a differential operator consistent with a collocation scheme. If we define as $\widehat{u}(x,y) = \sum_i^n \sum_j^m (B_{y_j} \otimes B_{x_i})c_{ij}$ the approximated two-dimensional state function

computed using a tensor product B-spline collocation procedure and, as $\breve{V}_{x,y}(w) = \sum_i^n \sum_j^m \left(\sum_{p=0}^k \sum_{q=0}^m w_{q,p}(\breve{B}_{y_j}^{(q)} \otimes \breve{B}_{x_i}^{(p)})c_{ij} \right)$, the B-spline approximation of the partial differential operator defined in consistency with a collocation scheme, our estimation problem is summarized by:

$$\min_c \|z - (B_y \otimes B_x)c\|^2 + \lambda\| \sum_{p,q=0}^m w_{q,p}(\breve{B}_y^{(q)} \otimes \breve{B}_x^{(p)})c\|^2, \tag{8.1}$$

where z is a vector of n observations in two-dimensions, λ is a smoothing parameter balancing between the residual sum of squares and the collocation approximation of the observed dynamics, c is a vector of unknown spline coefficients, w is a vector of DE parameters that we consider, for the moment, known (and m indicates the order of the differential operator that we consider to be linear in this chapter).

In (8.1) we used the breve accent to distinguish two grids of points on which the basis functions are evaluated in analogy with the notation used in chapter 6. The B-splines involved in the penalty term are computed over a dense grid of $N \times M$ collocation points. On the other hand the basis functions in $(B_y \otimes B_x)$ are evaluated at observed domain points. It is a convenient choice in those cases a moderate number of observations is available providing a more precise definition of the collocation solution of the hypothesized partial differential problem. This issue is particularly relevant in those circumstances in which we are interested in estimating the DE parameters from the raw data as we will see later.

To illustrate how to estimate the optimal spline coefficients satisfying (8.1) we propose a simple example. Figure 8.1 shows the two-dimensional cloud of 625 measurements (squared dots) and the smoothing surface describing the first order PDE diffusion equation introduced above. The estimation problem is summarized by:

$$\min_c \|z - (B_y \otimes B_x)c\|^2 + \lambda\|[(\breve{B}_y \otimes \breve{B}_x) + w_1(\breve{B}_y \otimes \breve{B}_x) + w_2(\breve{B}_y \otimes \breve{B}_x)]c\|^2.$$

The optimal spline coefficients are the solution of:

$$(B^T B + \lambda \breve{V}^T \breve{V})c = B^T z,$$

with $B = (B_{y_j} \otimes B_{x_i})$ and $\breve{V} = [2\breve{B}'_{y_j} \otimes \breve{B}_{x_i} + \breve{B}_{y_j} \otimes \breve{B}'_{x_i} + 3\breve{B}_{y_j} \otimes \breve{B}_{x_i}]$.
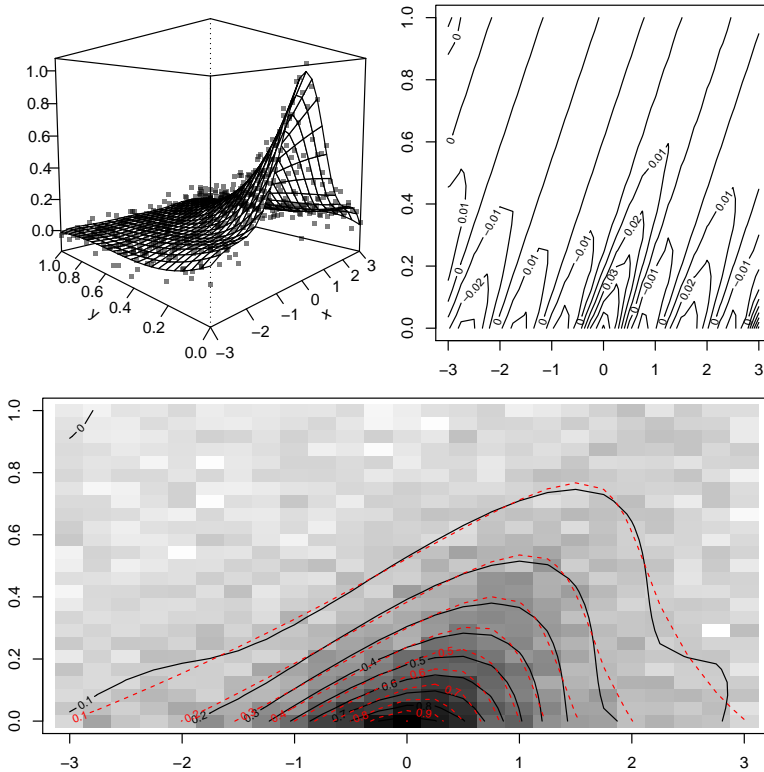
Figure 8.1: *Smoothing data driven by a PDE with known parameters. The upper left panel shows the raw observations and the smoothing surface while the upper right panel shows the residual surface between the smoother and the state function solving the PDE. The lower panel compares the contours of the smoothing function (black lines) with the contour of the solution of the PDE (red lines) while the panel background represents the observed data. The smoothing surface has been estimated using cubic B-splines built on 15 equally spaced knots defined in both x and y directions. The basis matrices used to define the differential penalties have been evaluated over a finer grid of 10^4 points.*
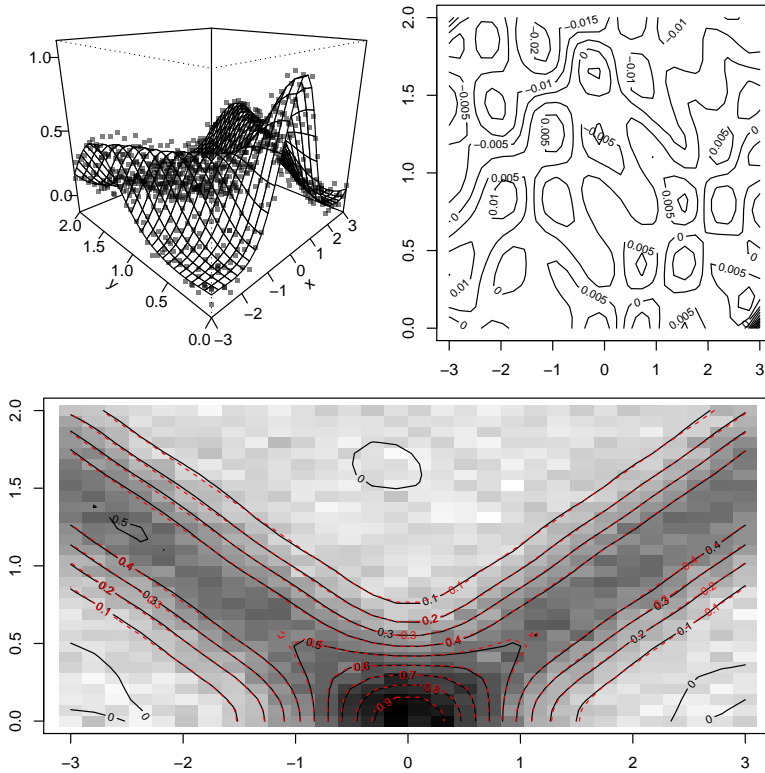
Figure 8.2: *Smoothing data driven by a second order PDE with known parameters. The upper left panel shows the raw observations and the smoothing surface while the upper right panel shows the residual surface between the smoother and the state function solving the PDE. The lower panel compares the contours of the smoothing function (black lines) with the contour of the solution of the PDE (red lines) while the panel background represents the observed data. The smoothing surface has been estimated using cubic B-splines built on 20 equally spaced knots defined in each direction. The basis matrices used to define the differential penalty have been evaluated over a grid of 10^4 points.*

A more complex example is described by the following second order Cauchy problem with mixed conditions:

$$u_{yy} - u_{xx} = 1, \text{ for } x \in \mathbb{R}, \ y \in \mathbb{R}$$
$$u(x, 0) = x^2, u_y(x, 0) = 1.$$

Figure 8.2 shows the raw data (gray dots in the upper right panel) and the estimated smoothing surface obtained for a cloud of 900 observations simulated adding a small noise to the numerical solution of the differential problem.

8.3   *Two-dimensional smoothing with unknown partial differential penalties*

In many applications we have only a vague idea about the differential law that could approximate a set of measurements describing a given phenomenon. A possibility consists in reconstructing the mathematical law we have in mind using the available observations.

This problem leads to the estimation of the unknown parameters defining the PDE. A possible approach consists in writing the analytic or numerical solution of the equation as explicit function of the unknown parameters estimated from the data. This generally leads to a nonlinear regression problem even if the PDE is linear. An alternative approach is to incorporate the partial differential equation as a penalty in a smoothing framework as suggested by Ramsay (2002) and Wood et al. (2008). This second approach is the one we follow here suggesting a tensor product P-spline-based two-stage procedure for the estimation of the unknown DE parameters.

We want to approximate the observed measurements in such a way the final estimates are consistent with the solution of the PDE. To reach this goal we need to estimate its unknown parameters. Our estimation problem can be summarized as follows:

$$\min_{w,c} \|z - (B_y \otimes B_x)c\|^2 + \lambda \|\check{V}c\|^2, \tag{8.2}$$

where $B_y \otimes B_x = B$ is a tensor product basis matrix, c is a vector of spline coefficients, λ is a smoothing parameter controlling the amount of penalty applied to the least squares criterion, $\check{V}$ is the B-spline approximation of a (linear) partial differential operator defined using a collocation scheme. This operator depends on the vector of weights w.

The unknown DE parameters can be estimated through a two-stage procedure. In analogy with what we discussed in chapter 6, we move from the consideration that the approximated state function implied in the PDE in the penalty term has to be consistent with the signal behind the observations if the differential model we have in mind is appropriate. This signal can be extracted from the noisy measurements using a tensor product P-spline. This smoother has to be able to separate the signal from the noise component of the raw observations. A possibility is to select the couple of optimal

smoothing parameter through an automatic procedure. The optimal spline coefficients defining the two-dimensional smoother provide a compact representation of the state function. Multiplying these coefficients by the basis matrices defining $\check{V}$ we obtain an approximation of the functions involved in the differential law. Once these functions have been approximated the parameter vector w can be estimated using least squares.

The smoothing parameter in (8.2) plays a crucial role. It regulates the relative importance of the residual sum of squares criterion over the collocation solution of the differential equation in determining the shape of the final fit. For $\lambda \to 0$ the smoothing surface tends to be rough given that the estimation procedure is driven exclusively by a residual sum of squares. On the other hand, for $\lambda \to \infty$ the estimated surface tends to mimic the numerical solution of the differential equation included in the penalty term.

Given the estimated differential parameters, it is possible to define the penalty term $\check{V}$ and estimate the optimal spline coefficients:

$$\widehat{c} = (B^T B + \lambda \check{V}^T \check{V})^{-1} B^T z \tag{8.3}$$

The optimal smoothing parameter λ can be selected using an automatic procedure. In the examples shown in this chapter we use an EM-type approach.

Figure 8.3 shows an example of the results obtained using the described approach. The top panels show the estimated smoother and the residuals between the fitting surface and the analytic solution of the differential problem while the lower panel shows a contour representation of the data and the smoothing function. The data have been generated taking into account the following PDE:

$$u_{yy} - 4u_{xx} = 0, \ \forall x \in \mathbb{R}, \ y \in \mathbb{R}^+$$
$$u(x,0) = \exp(-x^2), \ u_y(x,0) = 0.$$

The observations (squared dots in the left upper panel of figure 8.3) have been simulated adding a Gaussian noise to the approximated solution of this differential problem computed over $x_i \in [-3,3]$, $y_j \in [0,2]$. The DE parameter has been estimated equal to $\widehat{w} = -4.067$.
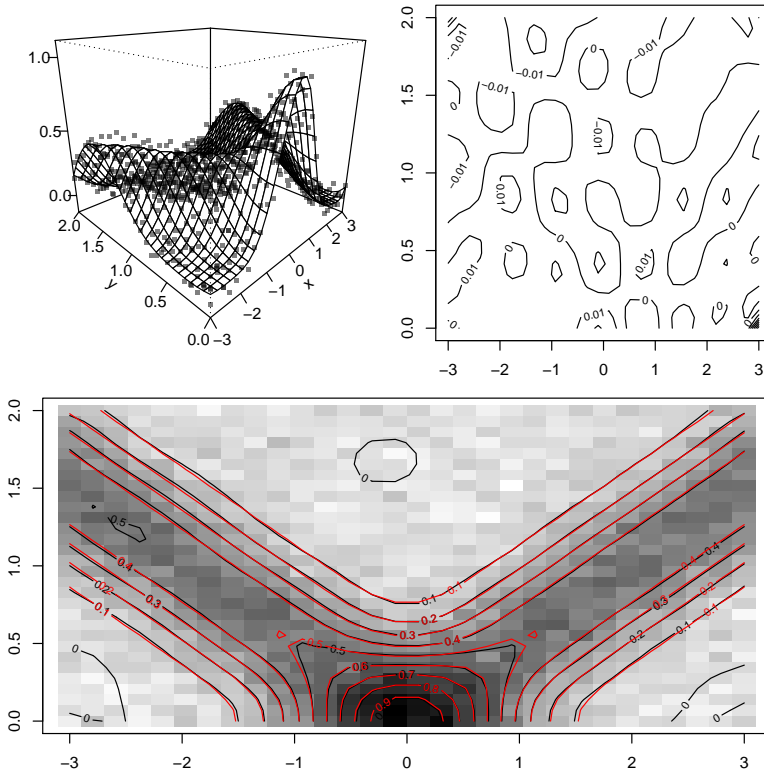
Figure 8.3: *Smoothing data driven by a PDE with unknown parameters. The smoothing surface has been estimated considering 20 forth order basis splines defined on equally spaced knots in each direction. The penalty bases have been evaluated over a grid of 10^4 points.*

8.4 Simulations

We are interested in evaluating how the sample size and the variability of the data noise influence the estimates obtained using the procedure described above. As first example we generate a set of data taking into account the solution of a first order PDE:

$$u_x + u_y = 2, \ x = y \in [-2, 2]$$
$$u(x, 0) = x^2.$$

We simulate 100 data clouds as follows: $z_{ij} = u^*(x_j, y_j) + \epsilon_{ij}(0, \sigma_\epsilon)$ with $i = j \in \{10, 15, 20, 25\}$, $\sigma_\epsilon \in \{0.5, 1, 1.5, 2\}$, and estimate the bias, the standard errors and the root mean square errors of the computed DE parameters. A possible data configuration obtained for $\sigma_\epsilon = 1$ is shown in figure 8.4. Table 8.1 lists the results for the described simulation study. From the results
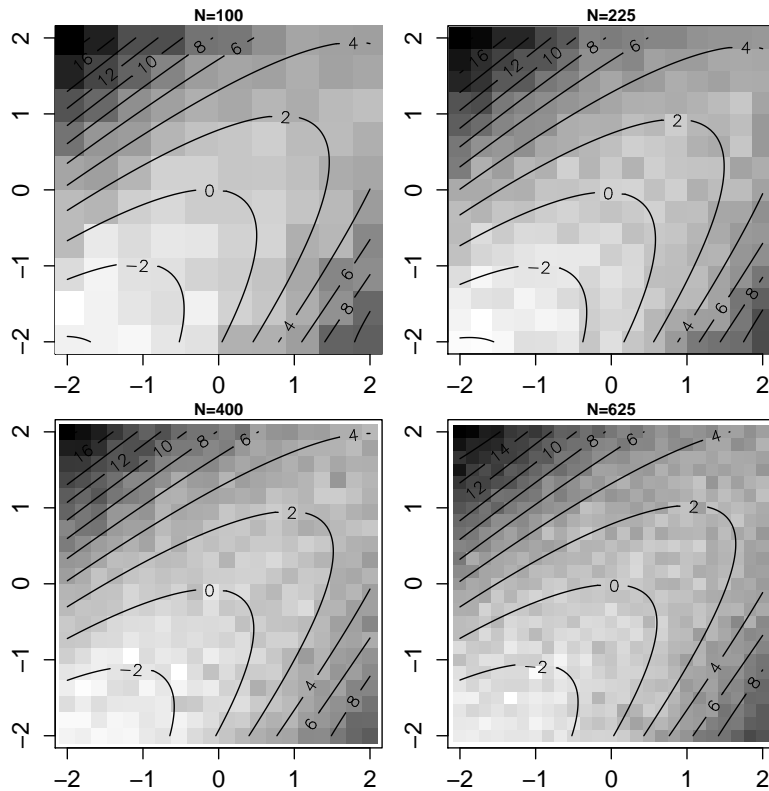
Figure 8.4: *Smoothing data driven by a first order PDE with unknown parameters. Each panel describes the smoothing surface obtained for different sample sizes. The estimates have been performed considering cubic B-splines built on 10 equally spaced knots selected on both x and y directions.*

it is possible to notice that the quality of the estimates tends to be lower if lower sample sizes are considered. Indeed, for larger sample sizes the bias, the standard deviation and the root mean square errors tend to be smaller taking constant the variability of the noise component. On the other hand, for a given sample size, the quality of the estimates tends to decrease if we consider a higher noise variability.

A similar simulation study can be conducted taking into account a more complex PDE:

$$u_{yy} - u_{xx} = 1, \ x = y \in [-2, 2]$$
$$u(x, 0) = x^2, \ u_y(x, 0) = 1.$$

We simulate 100 series of data: $z_{ij} = u^*(x_j, y_j) + \epsilon_{ij}(0, \sigma_\epsilon)$ with $i = j \in \{10, 15, 20, 25\}$, $\sigma_\epsilon \in \{0.5, 1, 1.5, 2\}$, and estimate the bias, the standard errors

True $w = 1$	$i = j = 10$			$i = j = 15$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma_\epsilon = 0.5$	0.0019	0.0232	0.0232	-0.0023	0.0170	0.0171
$\sigma_\epsilon = 1$	-0.0114	0.0568	0.0577	-0.0125	0.0394	0.0412
$\sigma_\epsilon = 1.5$	-0.0255	0.0669	0.0712	-0.0055	0.0475	0.0476
$\sigma_\epsilon = 2$	-0.0455	0.1039	0.1130	-0.0090	0.0755	0.0757

True $w = 1$	$i = j = 20$			$i = j = 25$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma_\epsilon = 0.5$	-2.0e-04	0.0138	0.0137	0.0018	0.0122	0.0123
$\sigma_\epsilon = 1$	0.0031	0.0307	0.0307	0.0040	0.0274	0.0276
$\sigma_\epsilon = 1.5$	-0.0073	0.0439	0.0443	-9.0e-04	0.0383	0.0381
$\sigma_\epsilon = 2$	-0.0172	0.0549	0.0573	0.0025	0.0545	0.0543

Table 8.1: *Bias, standard deviation and RMSE of the DE parameter estimated for a first order partial differential problem.*

and the root mean square errors of the computed DE parameters. A possible data configuration obtained considering $\sigma_\epsilon = 1$ is shown in figure 8.5. Table 8.2 summarizes the results obtained for the described experiment. Again the quality of the estimates seems to be higher for larger sample sizes and to decrease if we consider more variable noise components for each sample size.

8.5 Conclusions

In this chapter we have introduced a PDE-based penalized smoothing approach for the analysis of two-dimensional data which dynamics is approximately described by a partial differential law. Our approach involves a combination of tensor product basis spline functions for the approximation of the differential penalty in analogy with a tensor product B-spline collocation scheme.

In presence of unknown PDE parameters we adopt a two stage estimation procedure: 1) smooth the raw data using a P-spline smoother able to remove the noise, 2) estimate the DE parameters through least squares. This simplification separates the computation of the DE parameters from the data
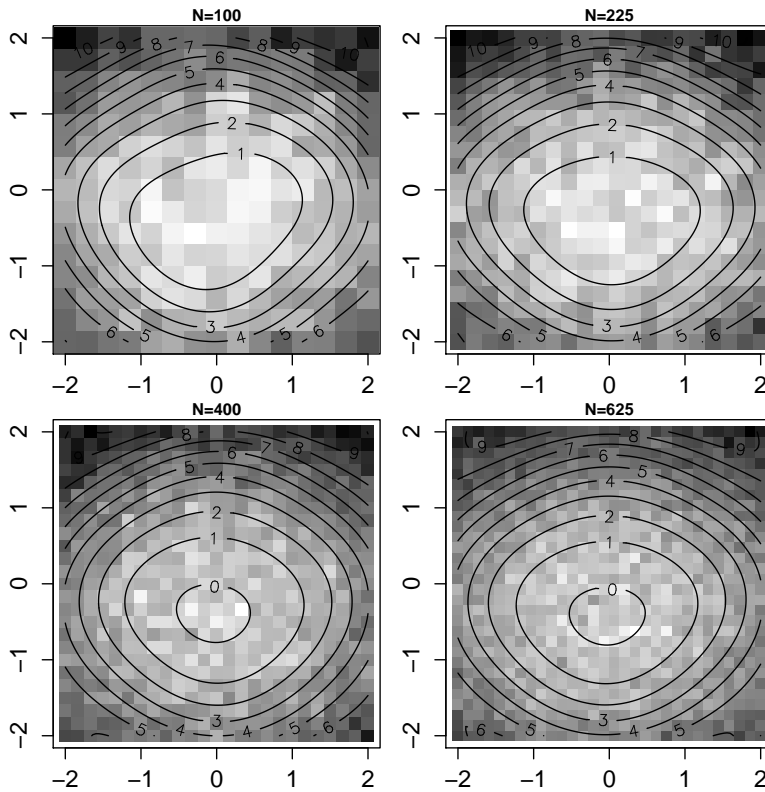
Figure 8.5: *Smoothing data driven by a second order PDE with unknown parameters. Each panel describes the smoothing surface obtained for different sample sizes. These results have been obtained considering 15 fourth order B-splines for each direction built on equally spaced knots.*

smoothing problem. The performance of this approach has been evaluated with simulated and real data.

Further generalizations of the discussed procedure are possible. An open question is the inclusion of nonlinear partial differential penalties with unknown parameters. The two stage procedure for the estimation of the unknown DE parameters can be easily adapted in order to consider non linear models. On the other hand a penalty term built considering a non linear differential operator can be included in the smoothing procedure only adopting some linearization strategy of the differential problem.

True $w = -1$	$i = j = 10$			$i = j = 15$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma_\epsilon = 0.5$	0.0863	0.0252	0.0899	0.067	0.021	0.0702
$\sigma_\epsilon = 1$	0.0894	0.0506	0.1026	0.0599	0.0451	0.0748
$\sigma_\epsilon = 1.5$	0.0680	0.0685	0.0963	0.0572	0.0668	0.0877
$\sigma_\epsilon = 2$	0.0576	0.1143	0.1275	0.0491	0.0873	0.0998

True $w = -1$	$i = j = 20$			$i = j = 25$		
	BIAS	STD	RMSE	BIAS	STD	RMSE
$\sigma_\epsilon = 0.5$	0.0529	0.0182	0.0559	0.0433	0.015	0.0458
$\sigma_\epsilon = 1$	0.0489	0.0361	0.0607	0.0401	0.031	0.0505
$\sigma_\epsilon = 1.5$	0.0412	0.0518	0.0660	0.0399	0.0407	0.0569
$\sigma_\epsilon = 2$	0.0480	0.0741	0.0880	0.0440	0.0632	0.0767

Table 8.2: *Bias, standard deviation and RMSE of the DE parameter estimated for a second order partial differential problem.*

9 GENERALIZATIONS OF COLLOCATION BASED PROCEDURES

> *In this chapter we briefly discuss some generalizations of the framework introduced in the previous chapters. The following sections present a sort of overview of the state of our art in analyzing more complex dynamics. In particular, we extend the collocation scheme for the solution of differential problems involving some kind of nonlinearity. We deal with differential problems involving inequality conditions, delay differential equations (DDEs) and nonlinear DEs. We also introduce a DDE-based penalized smoothing procedure.*
>
> ***Keywords**: Inequality conditions, asymmetric least squares, delay differential equations, nonlinear ODE, DDE-based penalized smoothing.*

9.1 Introduction

The B-spline collocation approach is flexible enough to allow for the solution of several classes of differential problems. In the previous chapters we focused on linear ordinary and differential equations. Here we concentrate on more complicated problems: ODE with inequality constraints, delay differential equations (DDE) and nonlinear ODEs.

Inequality constraints can be included in the collocation scheme with asymmetrically weighted penalties. In particular, in this chapter, we deal with constraints limiting the codomain of the state function and with monotonicity constraints.

Delay differential equations or difference differential equations (DDEs) have been introduced in the previous century in the field of automatic control. Any system with a feedback control involves time delays. A time delay arises because a finite time is required to sense information and then to react to it. In this chapter we evaluate the appropriateness of the B-spline colloca-

tion scheme for solving this kind of differential equations. We also introduce a DDE-based smoothing procedure to analyze data described by delayed dynamics.

Furthermore, the B-spline collocation scheme can also be applied to nonlinear differential equations. In particular we focus on ordinary differential equations. In order to solve a nonlinear problem through collocation some iterative linearization of the differential operator is needed. In this chapter, we suggest to use a Newton-type linearization method.

This chapter is organized as follows: in section 9.2 we introduce a B-spline collocation scheme for the solution of ordinary and partial DEs with unconventional conditions through weighted least squares. In section 9.5 the theoretical aspects about are briefly introduced and a simple problem is solved. In section 9.4 we deal with nonlinear ODEs. In section 9.5 we introduce a DDE-based penalized smoothing procedure for the analysis of data driven by delay dynamics.

9.2   *B-spline collocation solution of differential equations with unconventional conditions*

Suppose we are interested in obtaining a numerical solution for a given differential equation with a set of inequality constraints. An example is the second order harmonic oscillation problem described by:

$$y'' + 9y = 0, \ t \in [0, 5]$$
$$y(1) = 0, \ y(5) = 1,$$
$$y(t) \geq -1.$$

This problem does not have an analytic solution and the inequality constraint makes it nonlinear. A possible strategy is to face the problem without the last constraint and impose that the inequality holds in a least square sense. In other words it is possible to impose this constraint to the state function using asymmetrically penalized least squares. The penalty controls for violations

of the inequality condition. The estimation problem is:

$$\sum_j c_j B_j''(t_i, p) + 9 \sum_j c_j B_j(t_i, p) = e_i$$
$$\sum_j c_j B_j(t_i = 1) = 0$$
$$\sum_j c_j B_j(t_i = 5) = 1$$
$$\sum_i r_i (\sum_j c_j B_j(t_i) + 1)^2.$$

where e_i is is the approximation error evaluated on the ith collocation point and the last condition have been imposed using asymmetrically weighted least squares. The optimal spline coefficients follow from:

$$\begin{bmatrix} V^T V + B^T R_k B & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} c \\ l \end{bmatrix} = \begin{bmatrix} V^T(0 + \text{diag}(R_k)) \\ b \end{bmatrix}, \qquad (9.1)$$

where $V_{ij} = B_j''(t_i, p) + 9\ B_j(t_i, p)$ is the approximated differential operator, $K = [B(t_i = 0)\ B(t_i = 5)]^T$ and R_k is a weighting matrix that, in our example, assigns weight 0 where $Bc \geq -1$ and weight k otherwise. In our formulation k represents an asymmetric constraining factor. This penalty term only works where the inequality constraint is violated. A direct solution for c does not exist, but a simple iteration scheme is possible. For a given vector Bc it is trivial to compute the weights looking at those estimated values violating the inequality condition. Then the asymmetrically penalized least squares problem is solved again and the weights are updated. The iterative procedure goes on until convergence (usually, after few iterations no more estimated values violate the condition). In figure 9.1 the results for the example already introduced are shown. The estimates have been obtained using 50 equally spaced cubic splines. The asymmetric weights (k) have been set equal to 10^4.

Asymmetrically weighted least squares can also be applied to restrict the codomain of the state function (or one of its derivatives) when solving partial differential equations. Suppose we know that the state function is equal to zero over a certain region. This irregularity of the solution codomain can be viewed as an additional constraint. As an example suppose to be interested in solving the following problem:

$$u_x + u_y = 2, \ x \in [-10, 10], \ y \in [-10, 10]$$
$$u(x, 0) = x^2,$$
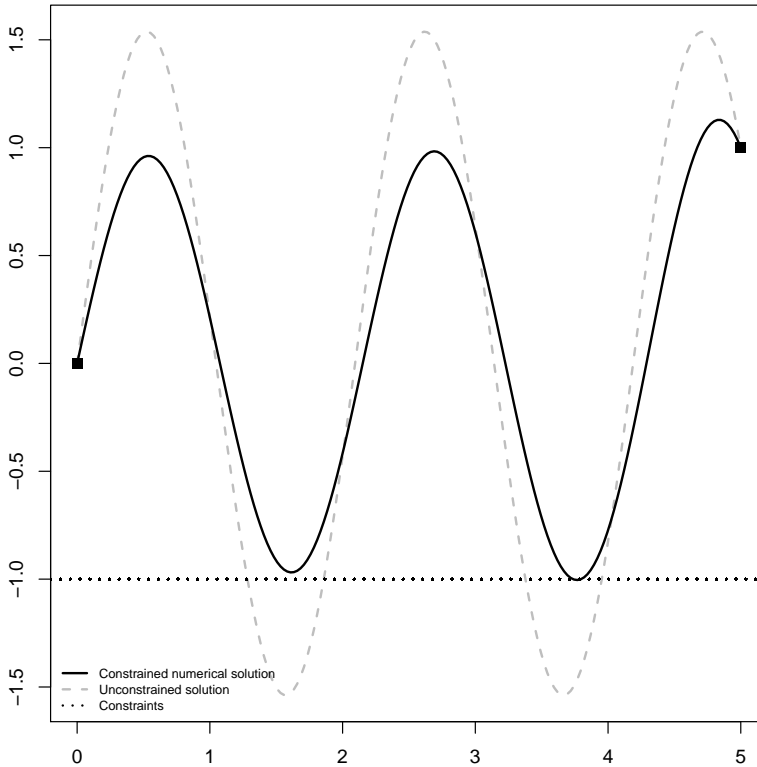$$u(x, y) = 0, \text{ for } (x, y) \in [-10, -2].$$

Figure 9.1: *Numerical solution of a second order BVP with inequality constraints. The gray dashed line indicates the unconstrained analytic solution, the dotted line indicates the constraint, the full line shows the constrained numerical solution together with the boundary conditions (squared points).*

From the definition of the problem it appears that the codomain of the state function has a jump at the boundaries delimited by $-10 \leq x \leq -2$ and $-10 \leq y \leq -2$. The approximate solution can be computed solving:

$$\sum_i \sum_j (B_{y_j} \otimes B'_{x_i})c_{ij} + \sum_i \sum_j (B'_{y_j} \otimes B_{x_i})c_{ij} - 2 = e_{ij}$$

$$\sum_i \sum_j (B_y(y_j = 0) \otimes B_{x_i})c_{ij} = \frac{1}{x_i^2}$$

$$\sum_i \sum_j (B_y(y_j \in [-10, -2]) \otimes B_x(x_i \in [-10, -2]))c_{ij} = 0.$$

where e_i is is the approximation error. This problem can be solved using a system of equations analogous to (9.2) defining a weighting matrix R that assumes value 0 if $(x, y) \neq [-10, -2]$ and one otherwise. The results are
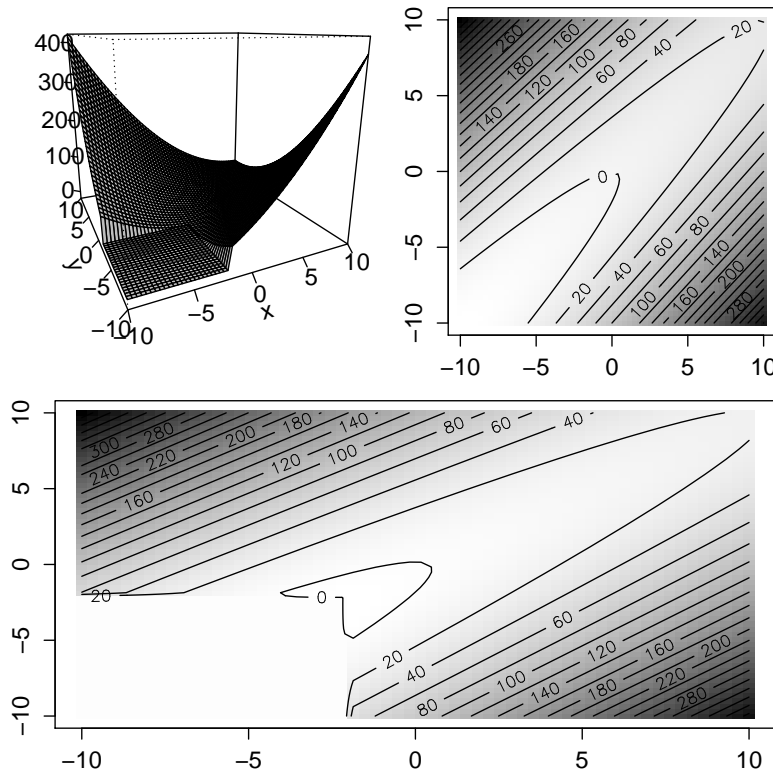
Figure 9.2: *Numerical solution of a first order PDE with a "jump" in the codomain. The upper panels show the numerical solution surface obtained using the irregular domain and the contours of the solution computed over the regular domain. The lower panel shows the contour of the solution computed over the irregular domain. These results obtained using 10 cubic basis spline functions defined on equidistant knots for both x and y.*

presented in figure 9.2.

A more complex problem can also be faced using a similar approach. Suppose to have an inequality condition restricting the codomain of the state function. In particular we can, for example, be interested in forcing the state function to be positive:

$$u_x + u_y = 2, \ x \in [-2, 2], \ y \in [-2, 2]$$
$$u(x, 0) = x^2,$$
$$u(x, y) > 0.$$

The unconstrained solution is $u^*(x, y) = 2y + (x - y)^2$. We again can approximate the inequality constrained state function using asymmetrically weighted penalized least squares. In analogy with what we saw for the ODE
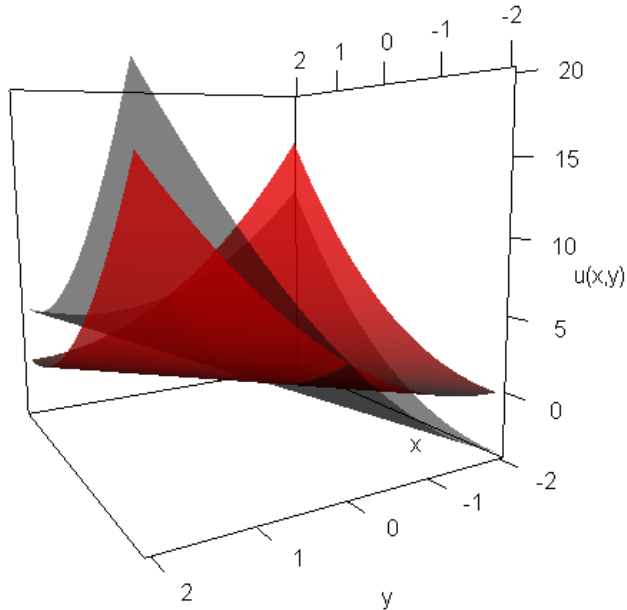
Figure 9.3: *Numerical solution of a first order PDE with a positivity constraint. The estimates have been obtained using 15 equally spaced cubic splines defined on the x and y directions. The asymmetric weights (k) have been set equal to 10^4. This figure shows the approximated state function (red surface) and the unconstrained analytic solution (gray surface).*

example above, the optimal spline coefficients follow from:

$$
\begin{bmatrix} V^T V + B^T R_k B & K^T \\ K & 0 \end{bmatrix} \begin{bmatrix} c \\ l \end{bmatrix} = \begin{bmatrix} V^T(0 + \text{diag}(R_k)) \\ x^2 \end{bmatrix},
$$

where $V = [B'_{y_j} \otimes B_{x_i} + B_{y_j} \otimes B'_{x_i} - 2]$ is the approximated differential operator and $B = B_y \otimes B_x$, R_k is a weighting matrix that, in our example, assigns weight 0 to those estimates for which $Bc \geq 0$ and weight k otherwise. Again, k represents an asymmetric constraining factor. In figure 9.3 the results for this example are depicted.

9.3 *B-spline collocation solution of delay differential equations*

In this section we introduce differential problems the delay differential equations. Here we provide a really brief introduction of the topic and refer to specialized references for a more detailed description (Erneux, 2009).

Delay differential equations or difference differential equations (DDEs) have been introduced in the previous century in the field of automatic control. Any system involving a feedback control involves time delays. A time delay arises because a finite time is required to sense information and then react to it. DDEs have properties that distinguish them completely form the conventional ordinary or partial differential equations.

First of all a time-dependent solution of a DDE is not determined by its initial state at a given moment but, instead, the solution profile on an interval with length equal to the delay h (or time lag) has to be provided. Usually this interval is antecedent to the initial point and for this reason the solution profile on it is often named "history". Analytic solutions can be found only in really simple cases. The method of steps is an elementary procedure that can be used to solve some DDEs analytically. This method is based on the solution of successive DEs on subintervals of the domain defined taking steps of length h.

To clarify this aspect we show an example:

$$\frac{dy}{dt} = -y(t - h)$$
$$y(t) = 1, \text{ for } -h \leq t \leq 0,$$

with $h = 1$ and $y_0 = 1$. We first solve this problem for the interval $t \in [0, 1[$ obtaining $\frac{dy}{dt} = -1$ which we solve using $y(0) = 1$. The solution is $y = 1 - t$ for $t \in [0, 1[$. Now we take one step ahead taking $t \in [1, 2[$ and solving $\frac{dy}{dt} = -1 + (t - 1)$ with initial condition computed at $t = 1$ that is $y(1) = 0$. The is $y = -(t - 1) + 0.5(t - 1)^2$. We can then move to the next interval $[2, 3[$, and so on. This procedure can, in principle, be continued as far as desired although the calculations quickly become tedious and complex for non-straightforward equations.

A collocation scheme, involving B-splines or other kind of bases, represents a convenient numerical procedure to solve DDEs. The approximation
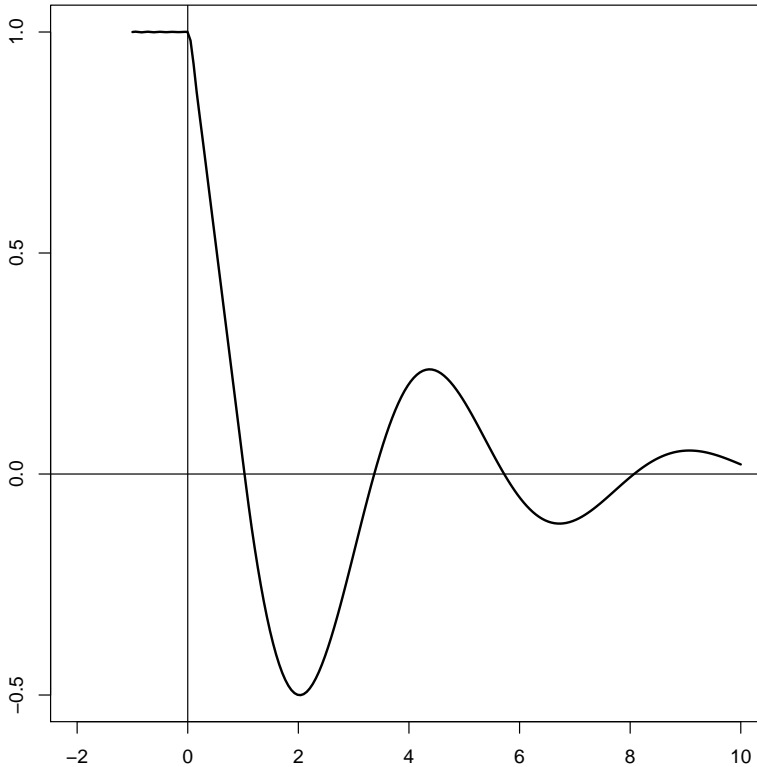
Figure 9.4: *Numerical solution of a first order DDE. These results have been obtained using* 100 *quadratic basis spline functions defined on equally spaced knots and on a domain spanned by 200 time points.*

of the delayed functions requires only the computation of shifted bases and the same set of coefficients.

Now we show how to solve this problem with a B-spline collocation scheme on the domain $t \in [0, 10]$. We have to solve:

$$\left[\begin{array}{cc} V^T V & K \\ K & 0 \end{array} \right] \left[\begin{array}{c} c \\ l \end{array} \right] = \left[\begin{array}{c} V^T 0 \\ b \end{array} \right],$$
(9.2)

with $V_{ij} = B'_j(t_i) + B_j(t_i - k)$, $K = B_j(t_i \in [-1, 0[)$ and $b = y(t \in [-k, 0[)$. The solution is depicted in figure 9.4. It is clear that the state function solving this first order DDE has an oscillatory behavior and a discontinuity in its first derivative for $t \in [-1, 0[$. Oscillatory behavior and discontinuous state functions are typical for this kind of problems.

9.4 *Nonlinear ODEs*

The B-spline collocation scheme can also be used to solve nonlinear ODEs. In this section we develop a linearized algorithm based on Taylor series. An example is the following second order ODE:

$$0.01y'' + y^2 = 1, \ t \in [0, 1]$$
$$y'(0) = y(1) = 1.$$

Suppose we start with an initial guess ($i = 0$) of the state function, $y_i = t^2 - 1$ a parabola (making $y'' = 0$). Interpolating the values of y_i we estimate an initial vector of spline coefficients c_i. We use these coefficients to approximate $\hat{y}_{i+1} = \sum_j B_j(t)c_{j,i}$. This initial approximation of the state function is used to define the linearized differential operator $V = w_{i+1}B(t) + 0.01B''(t)$ with $w_{i+1} = 2\hat{y}_{i+1}$ (the first derivative of the nonlinear term in the DE). The next optimal spline coefficients c_{i+1} are computed solving the following problem using least squares:

$$\sum_j w_{i+1}B_j(t)c_{j,i+1} + \sum 0.01B''(t)c_{j,i+1} = \hat{y}_i^2 - 1$$
$$\sum_j B_j'(t = 0)c_{j,i+1} = 0$$
$$\sum_j B_j(t = 1)c_{j,i+1} = 0,$$

that leads to the following system:

$$\begin{bmatrix} B'(t = 0) \\ w_{i+1}B(t) + 0.01B''(t) \\ B(t = 1) \end{bmatrix} c_{i+1} = \begin{bmatrix} 0 \\ \hat{y}_i^2 - 1 \\ 0 \end{bmatrix}. \tag{9.3}$$

Once the new spline coefficients have been computed we can start again the procedure. The iterative procedure stops when some convergence criterion is reached (typically when t $\max(|\hat{y}_i - \hat{y}_{i+1}|)$ is lower than a tolerance level). Figure 9.5 shows the approximated solution obtained for this BVP. The algorithm converges in $n = 8$ steps with a maximum absolute difference between the state function approximated in the previous (seventh) iteration less then 10^{-9}.
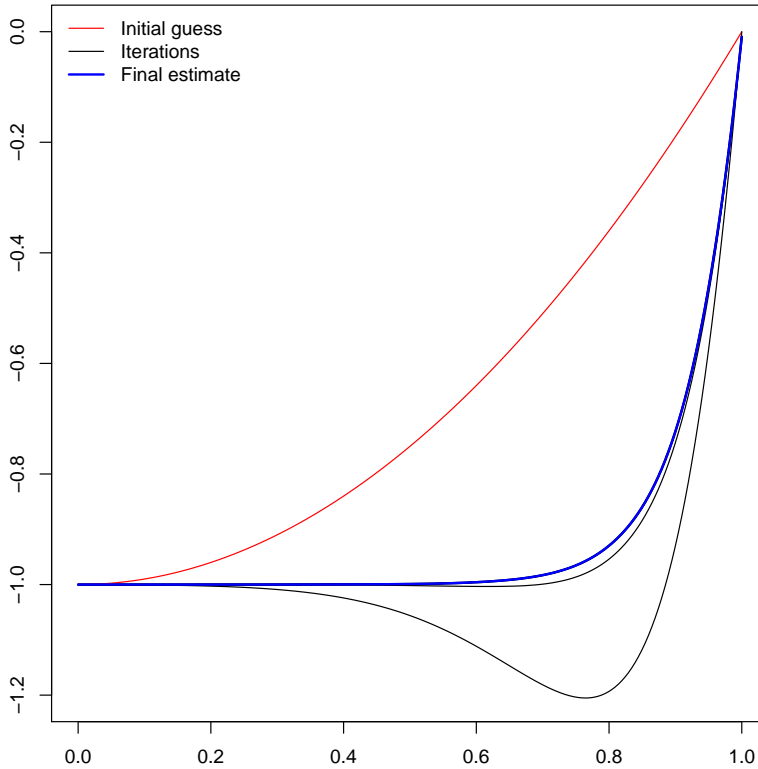
Figure 9.5: *Numerical solution of a second order nonlinear BVP. These results have been obtained using cubic basis spline functions built on 50 equally spaced knots selected over 200 collocation points.*

Figure 9.6 shows the solution of a logistic differential equation.

$$y^{'}(t) - y(t) + y(t)^2 = 0,\ t \in [-6, 6]$$
$$y(0) = 0.5.$$

In the first panel of this figure we compere the approximated state function (solid green line) with the analytic solution $y^* = \frac{1}{1+\exp(-t)}$ (squared dots) showing the successive steps computed using a Newton's approach (dashed black lines).

9.5   *Smoothing with delay differential penalties*

In this section we introduce a DE-based penalized smoothing approach for the analysis of delayed dynamic systems. The estimation problem is analogous to what we already discussed in the previous chapters of this work.
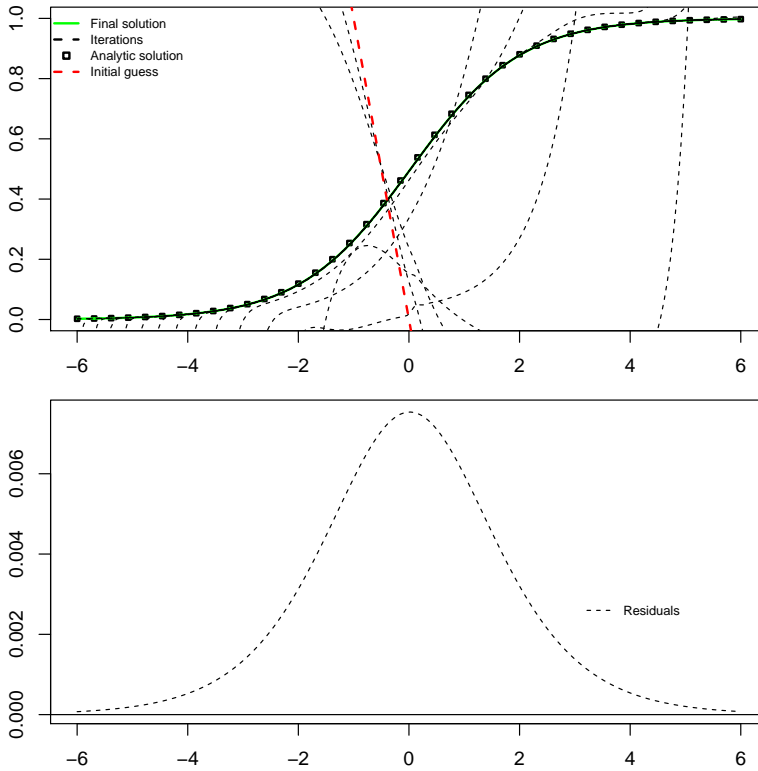
Figure 9.6: *Numerical solution of the logistic differential equation. These results have been obtained using an initial guess $y_0 = -t$ (dashed red line) and using cubic B-spline functions built on 100 equally spaced knots selected over 200 t points. The algorithm converges after 19 iterations reaching a maximum absolute difference with the estimation in the previous iteration less than 10^{-9}. The lower panel of the figure shows the residuals between the approximated and the analytic state function solving this nonlinear DE.*

As an example suppose to observe a set of data which dynamics is appropriately described by the delay differential equation introduced in section . A possible way to describe the observe dynamics is to find a smooth function that efficiently reproduces the data trend and that is consistent with the solution of the DDE we have in mind. Consider the data represented by the gray dots in figure 9.7 (upper panel). We solve the following minimization problem:

$$\min_{c,w} \|y - \sum_{j}^{n} B_j c_j\|^2 + \lambda \| \sum_{j}^{n} \breve{B}_j^{'}(t) + w\breve{B}_j(t-1)\|^2, \qquad (9.4)$$

In the formulation above we used the breve accent to distinguish two

sets of points on which the basis functions are evaluated. The B-splines involved in the penalty term are evaluated on an "enlarged" set of N collocation points. On the other hand the basis functions in B are evaluated at observed domain points. This definition of the penalty term is convenient in those cases where a moderate number of observations is available.

The parameter w is estimated using two-stage approach. In the first stage we smooth the observations using a P-spline with an appropriate amount of penalty. Then, the optimal penalized coefficients can be used to approximate the functions included in the differential penalty of (9.4). Once these approximated functions are obtained the optimal DE parameter can be computed using least squares.

It is possible to notice that the DDE-based penalized smoothing estimates are really close to the collocation approximation of the state function (see also lower panel of figure 9.7). The optimal DE parameter has been found approximately equal to 1 (i.e. the value that we used to simulate the data). A valuable feature is the reproduction of the discontinuity at $t = 0$.

9.6   *Conclusions*

In this chapter we have introduced some topics that will be part of our future research.

Our discussion started with possible generalizations of the B-spline collocation scheme. First of all we have shown how to approximate the state function solving differential problems with nonconventional differential conditions. These conditions also have application in DE-based data smoothing. We could be interested, for example, in monotone smoothing applications or in smoothing data observed over irregular domains.

Then we introduced a collocation scheme for the solution of delayed differential equations which are common in many scientific fields studying dynamic systems with feedbacks.

Furthermore we introduced a DDE-based penalized spline approach for the analysis of observed phenomenon driven by delayed evolutions. In case delayed dynamics, possible discontinuities in the state function call for a flexible smoothing approaches.

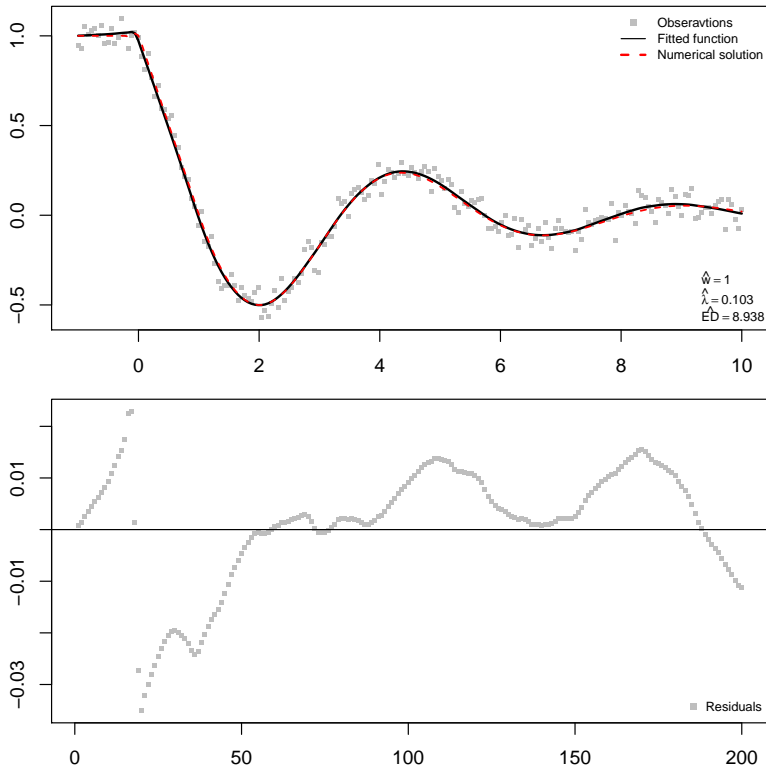Finally, the collocation scheme has been adopted for the solution of non-

Figure 9.7: *Numerical solution of a first order DDE. The data have been simulated from the numerical solution of the delay differential equation discussed in section 9.5 computed for 200 time values and adding a Gaussian noise with null mean and standard deviation equal to 0.1. The lower panel shows the simulated data (gray dots), the estimated smoothing function (black line) and the collocation solution (dashed red line). In the legend of the first plot the estimated DE parameter, the optimal λ parameter and the related model effective dimension are listed. The lower panel of the figure depicts the differences between the smoothing function and the numerical solution od the DDE. These estimates have been obtained using quadratic basis splines built on a generous number of knots (200). The B-splines matrices involved in the penalty term have been evaluated over 500 collocation points. The red dashed line depicts the collocation solution of the DDE.*

linear ODEs using an iterative linearization scheme.

10

CONCLUDING COMMENTS

The aim of this thesis is to introduce advanced methods for of complex smoothing problems. We mainly treat two aspects: the selection of the optimal weight of the roughness penalty and the definition of differential operator penalty terms. Here we review shortly the topics treated in the previous chapters and introduce the directions of our future research.

With respect to the first topic, in chapter 3, we introduce the L-curve smoothing parameter selection procedure. It selects the smoothing parameter through a direct comparison between the goodness of fit and the smoothness of the estimates. The optimal smoothing parameter can be selected by locating the point of maximum curvature, i.e. the corner, of the L-curve as was originally argued in the works of Hansen (Hansen, 1992; Hansen and O'Leary, 1993). We also propose an alternative selection procedure based on the minimization of the Euclidean distance between the adjacent points of the curve: the V-curve criterion.

In chapter 4, the L-curve procedure is generalized to select the optimal amount of penalty in two-dimensional non-isotropic tensor product P-spline smoothing. A natural two dimensional generalization of the L-curve is represented by the L-surface, defined by a set of 1D L-curve computed in the x and y directions. Each L-curve has a convex region and the points defining them become closer to each other in the proximity of the corner. The Euclidean distance between adjacent points on a single L-curve is described by a V-shaped curve and the single dimensional optimal smoothing parameter is located at the minimum of the V-curve.

In the next chapter we concentrate on smoothing procedures that use differential operator penalties. In particular we move from the introduction of a B-spline collocation procedure for ordinary and partial differential equations. As practical application we show how to solve the Black and Scholes

differential equation for pricing European options using a B-spline colloca-
tion procedure.

In chapter 5, the B-spline collocation scheme an be adopted to define
smoothing procedures penalized by ordinary and partial differential opera-
tors. In our approach the penalty term involves a combination of basis spline
functions for the approximation of the differential operator in analogy with a
B-spline collocation scheme. Penalizing the smoothing procedure by a linear
differential operator consistency between the estimated smoothing function
and the state function approximating the solution of the DE-based penalty.
Our aim can be viewed in terms of generalized differential conditions. We
look for a solution of the DE such to optimally describe the observed data
in a least squares sense. On the other hand conventional initial/boundary
conditions as well as integral conditions can be included in the estimation
procedure through Lagrange multipliers.

A smoothing parameter regulates the fidelity of the final estimates to the
approximated state function. This parameter, if selected using an automatic
procedure, can be interpreted as a data-based measure of the appropriate-
ness of the differential operator penalty. In particular we suggested to select
this parameter through an EM-type algorithm exploiting the mixed model
interpretation of penalized least squares.

In presence of unknown DE parameters we adopt a two stage estima-
tion procedure: 1) smooth the raw data using a P-spline smoother able to
remove the noise, 2) estimate the DE parameters through least squares. This
simplification separates the computation of the DE parameters from the data
smoothing problem. The performance of this approach has been evaluated
with simulated and real data.

In many applications, the data are grouped with respect to some factor.
This leads to dynamic systems described by DEs with similar parameters. In
chapter 7, our DE-based penalized smoothing approach has been generalized
to take into account mixed DE parameters in one dimensional applications.
In order to estimate the mixed effect model defining the DE hypothesized to
govern the observed data we combine the two-stage P-spline-based proce-
dure with a ridge regression framework. In this way we exploit the link be-
tween penalized regression and mixed models (Pawitan, 2001). This allows a
practical estimation and interpretation of the variance components. In mixed

models two variance parameters occur: one for the observation noise and one related to the random effects. Both can be estimated from the marginal likelihood, using an EM algorithm.

In the last chapter we introduce some topics that will be part of our future research. We consider possible generalizations of the B-spline collocation scheme. We show how to approximate the state function solving differential problems with inequality and shape constraints. An example is forcing the numerical solution to be nonnegative or convex. Then we generalize the B-spline collocation framework for the solution of nonlinear ODEs and delay differential equations (DDEs) introducing also a DDE-based smoothing procedure based on the same framework illustrated for the smoothing of dynamic systems data described by ODEs and PDEs.

10.1 *Further research*

Our further research will develop in two directions and can be summarized as follows:

- The first challenge we will face is related to the definition of a formal statistical framework for the L-curve smoothing parameter selection procedure. Indeed, deeper studies on the non-statistical foundations and interpretation of this selection procedure are needed. First of all it is of primary interest to give a formal explanation of its robustness to correlated noise. On the other hand we believe that a generalization of the L-curve (for one and two dimensional applications) for smoothing problems with non Gaussian data and penalty terms defined by vector norms different from the Euclidean one. Our future research will also concentrate on an L-curve criterion for adaptive smoothing problems and on L-curve suitable for expectile smoothing problems.

- A second line of research will be related to the B-spline collocation scheme. In particular we believe that the identification of the optimal collocation points can represent an important issue in many applications, especially dealing with observed data. Indeed, the accuracy of the approximation and the computational cost required by the collocation scheme depend also on this aspect. In this work we adopt a rather intuitive strategy locating the collocation points on a dense grid cover-

ing of the domain. The accuracy in terms of error analysis of our choice was not investigated but it will be part of our future research.

- We will also focus on nonlinear DE-based penalized smoothing procedure. The requirement for linear differential penalties represents a big limitation for the applicability of the procedures presented in this thesis. A quasi-linearization Newton-type approach looks promising. The introduction of a nonlinear DE-based penalized smoothing approach could represent an opportunity for a more appropriate description of the MRI-based stomach contractions analysis proposed in chapters 6 and 7 of this thesis.

- In chapter 9 we proposed a DDE-based smoothing procedure showing a two-stage P-spline based procedure for the estimation of the unknown DDE parameters. We implicitly supposed to know the size of the delay h. In man application h is unknown and has to be estimated from the data. The study of an appropriate procedure to estimate the delay parameter will also be part of our further research.

- In this thesis, we supposed to observe data with exogenous source of noise. We did not deal with those circumstances in which the model errors are part of the dynamics. In this case modeling errors are introduced by systematic noise characterizing the differential model. This consideration leads to the definition of stochastic differential based penalties in the smoothing settings. The differential operator penalizing the smoothing procedure would represent the system noise in these cases.

- Finally, an obvious subject would be to combine the L-curve procedure and differential operator penalties in a smoothing setting.

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 can be useful to understand why, for well behaved curves, we can simplify the selection procedure avoiding the computation of the curvature. In particular it is possible to show that the L-curve has a negative first derivative and that it has a convex region if some condition on the first derivative of its components hold. We also show that if a convex area exists the point of maximum curvature is well approximated minimizing the Euclidean distance between adjacent points on the curve.

In order 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. In this case $\widehat{z}_\lambda = (I + \lambda D^T D)^{-1} y$ represents the vector of regularized fitted values (from this moment 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:

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

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

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

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

and, 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:

$$\frac{d\omega}{d\lambda}\frac{d\lambda}{d\theta} = -\lambda$$
$$\frac{d\theta}{d\omega} = -\frac{1}{\lambda}$$
$$\frac{d^2\theta}{d\omega^2} = \frac{d}{d\omega}\left(-\frac{1}{\lambda}\right)$$
$$= \left(\frac{1}{\lambda^2}\right)\frac{d\lambda}{d\omega}$$
$$= \left(\lambda^2 \frac{d\omega}{d\lambda}\right)^{-1}$$
$$= \left(2\lambda^3 z^T P^T H P z\right)^{-1}. \tag{A.4}$$

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:

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

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

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

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:

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

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:

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

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:

$$\left(\frac{d\phi}{d\psi}\right)^2\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] > -\frac{d\phi}{d\psi}$$

$$\frac{d\phi}{d\psi}\left[-\frac{1}{\lambda}\frac{d\lambda}{d\phi} - 1\right] < -1$$

$$-\frac{1}{\lambda}\frac{d\lambda}{d\phi} > -\frac{d\psi}{d\phi} + 1$$

$$-\frac{1}{\lambda}\frac{d\lambda}{d\phi} > \frac{-d\psi + d\phi}{d\phi}$$

$$\frac{1}{\lambda}d\lambda > d\psi - d\phi. \tag{A.12}$$

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 (considering $\log(\lambda) = \ell$):

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

This condition can be verified numerically. Consider as examples two cases showed above. In particular we take a case in which a clear corner is present and another in which the convex region of the L-curve is not pronounced. The first case is given by the example in figure 3. Figure A.1 shows the curvature function related with this example. The smaller segments under this curve show the positive curvature points found using (A.13). The second panel of figure A.1 plots the numerator of the curvature function and the vertical lines indicate the points of positive curvature found applying the criterion in (A.13). On the other hand figure A.2 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 parameterizations, the numerator of (A.5) has 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 that area and that if square them we get smaller values.

It is also possible to notice that the denominator of (A.5) is strictly related to our simplified selection criterion based on the Euclidean distance between adjacent points on the curve. Extending the previous reasoning to this quantity it appears clear why this minimization procedure leads to a maximum curvature point for well-behaved L-curves.
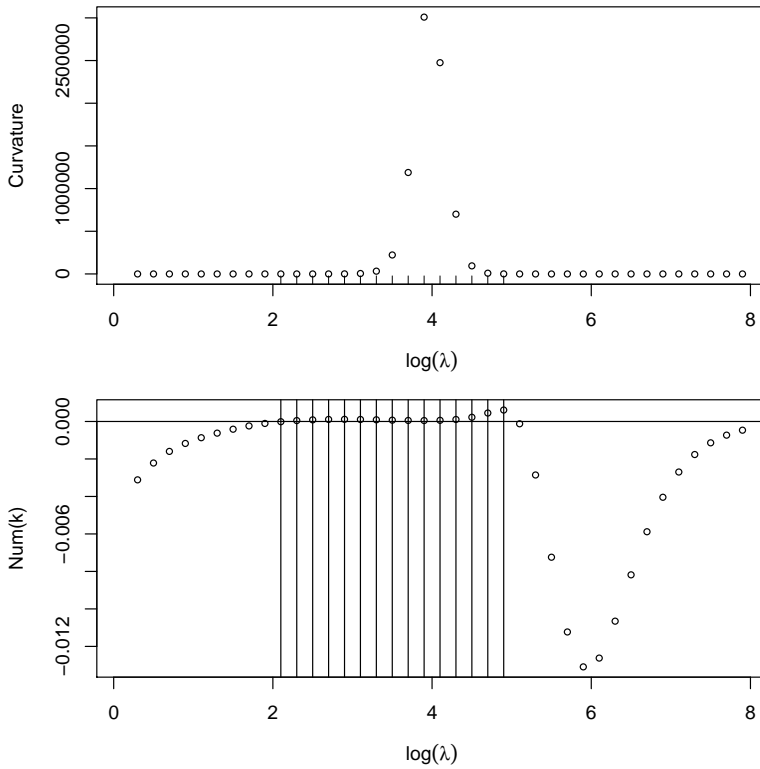
Figure A.1: *Curvature function and numerator of the curvature function for the example in figure 2. In the second panel the horizontal line indicates the zero abscissa level.*
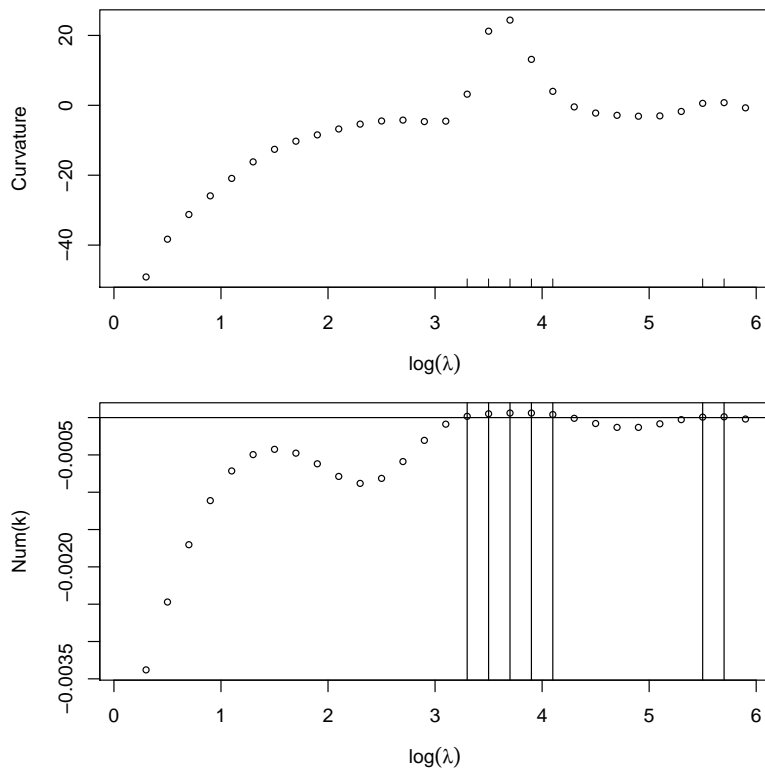
Figure A.2: *Curvature function and numerator of the curvature function for the mean sea level example. In the second panel the horizontal line indicates the zero abscissa level.*

BIBLIOGRAPHY

Ascher, U. M. and Petzold, L. R. (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1st edition.

Belge, M., Kilmer, M., and Miller, E. L. (1998). Simultaneous multiple regularization parameter selection by means of the l-hypersurface with applications to linear inverse problems posed in the wavelet transform domain.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Botella, O. (2002). On a collocation B-spline method for the solution of the Navier–Stokes equations. *Computers and Fluids*, 31(4-7):397–420.

Cao, J. and Ramsay, J. O. (2007). Parameter cascades and profiling in functional data analysis. *Comput. Stat.*, 22(3):335–351.

Carmack, P. S., Spence, J. S., and Schucany, W. R. (2012). Generalised correlated cross-validation. *Journal of Nonparametric Statistics*, 24(2):269–282.

Coddington, E. A. and Levinson, N. (1984). *Theory of Ordinary Differential Equations*. Robert E. Krieger, Malabar, Florida.

Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263.

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

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag.

de Boor, C. and Swartz, B. (1973). Collocation at Gaussian Points. *SIAM Journal on Numerical Analysis*, 10(4).

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Oxford University Press.

Eilers, P., Currie, I., and Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76.

Eilers, P. H. (2003). A perfect smoother. *Analytical chemistry*, 75(14):3631–3636.

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

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66:159–174.

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

Erneux, T. (2009). *Applied Delay Differential Equations*. Surveys and tutorials in the applied mathematical sciences. Springer-Verlag New York.

Evans, L. C. (1998). *Partial Differential Equations (Graduate Studies in Mathematics, V. 19) GSM/19*. American Mathematical Society.

Golub, G. H. and Ortega, J. M. (1992). *Scientific Computing and Differential Equations*. Academic Press, New York and London.

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

Hansen, P. C. (2000). 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.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.

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

Johns, C., Nychka, D., T.G.F., K., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, 98:796–806.

Kauermann, G., Krivobokova, T., and Semmler, W. (2011). Filtering time series with penalized splines. *Studies in Nonlinear Dynamics & Econometrics*, 15(2):2.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102:1328–1337.

Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

Pandit, S. M. and Wu, S. M. (1993). *Time series and system analysis with applications*. Krieger.

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition.

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):307–319.

Ruppert, D., W. M. C. R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics (No. 12).

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. John Wiley and Sons, New York.

Shahrak, N. M., S. A. O. (2013). Robust psd determination of micro and meso-pore adsorbents via novel modified u curve method. *Chemical Engineering Research and Design*, 1(91):51–62.

Stock, J. and Watson, M. (2003). *Introduction to econometrics*. The Addison-Wesley series in economics. Addison-Wesley, Boston, Mass. [u.a.], internat. ed edition.

Vogel, C. R. (1996). Non-convergence of the l-curve regularization parameter selection method. *Inverse Problems*, 12(4):535.

Wahba, G. (1990). *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.

Wang, L., Cao, J., Ramsay, J. O., Burger, D. M., Laporte, C. J. L., and Rock-stroh, J. K. (2012). Estimating mixed-effects differential equation models. *Statistics and Computing*, pages 1–11.

Welham, S. J., Cullis, B. R., Kenward, M. G., and Thompson, R. (2006). The analysis of longitudinal data using mixed model l-splines. *Biometrics*, 62(2):392–401.

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

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

Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap film smoothing. *Journal Of The Royal Statistical Society Series B*, 70(5):931–955.