NTNU
Norwegian University of
Science and Technology

Stochastic Models for Smoothing Splines

A Bayesian Approach

Kristoffer Herland Hellton

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Håvard Rue, MATH
Co-supervisor: Daniel Simpson, EXT

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem description

In this thesis, stochastic models for (smoothing) splines are discussed. In particular, we compare mainstream models for splines and those that can be constructed using finite element representation of the solution of stochastic differential equations.

Assignment given: 14. January 2011
Supervisor: Håvard Rue, MATH
Co-supervisor: Daniel P. Simpson, MATH

Many difficulties which nature throws in our way,
may be smoothed away by the exercise of intelligence.
Titus Livius
(Roman historian, 59 BC - 17 AD)

Preface

This thesis completes my Master of Science and a five year study program, *Industrial mathematics*, at the Department of Mathematical Sciences, Norwegian University of Science and Technology and is the end result of twenty weeks of hard work, the spring 2011, written under the course code *TMA4905, Statistics, Master Thesis*.

After finishing my Specialization Project, the autumn 2010, in extreme value methods, I wanted to write a more theoretical thesis within statistics and approached Professor Håvard Rue with this aim. He responded he had the perfect topic for me, which resulted in this thesis and I thank my supervisor Prof. Rue for all help and this wonderfully interesting and challenging opportunity. I also especially want to thank my co-supervisor Dr. Daniel P. Simpson, for all his advice, support and encouragement.

Finally, I would like to give my warmest thanks to my parents for always being there for me.

Kristoffer Herland Hellton
Trondheim, June 10, 2011

Abstract

Flexible data regression is an important tool for capturing complicated trends in data. One approach is penalized smoothing splines, where there are several mainstream methods. A weakness is, however, the quantification of uncertainty. We will in thesis present two mainstream smoothing spline methods, P-splines and O'Sullivan splines, and the RW2 model; a Bayesian hierarchical model based on a latent field. The Bayesian prior is specified by a stochastic Poisson equation, and spline estimates are approximated along a finite element Galerkin approach. We evaluate the three methods using integrated nested Laplace approximations (INLA) for a full Bayesian analysis, supplying credible bands. The methods give fairly similar results and we investigate the theoretical motivates behind the methods. As an extension of the Bayesian models, the smoothing parameter is incorporated in the latent field. This gives an adaptive smoothing method, which better estimates jumps and quick curvature changes. Further, the close relationship between O'Sullivan splines smoothing splines is discussed, revealing O'Sullivan splines to be a finite element Petrov-Galerkin approximation of smoothing splines. The main results are the possibility of credible bands, the extension to adaptive smoothing and the finite element understanding of O'Sullivan splines.

Contents

List of Figures

Chapter 1

Introduction

The core of modern science is perhaps the collection and analysis of data. Within this analysis, one of the most important problems is to find relationships between measurements, specifically, how a set of underlying variables influence a response. Different questions arise: How does an increase in the oil prize affect consumer buying power, does higher acidity in water samples result in fewer fish, and can the height and girth of a tree be used to predict the final volume of timber? All these questions ask for a relationship between variables, the classical problem of regression. The mathematical model incorporates the idea of a relation $f(\cdot)$ between x and y and a measurement error ϵ,

$$y = f(x) + \epsilon,$$

where the assumption of random errors $\epsilon \sim N(0, \sigma^2)$ brings the whole problem into the field of statistics. Historically, the development started with Carl Friedrich Gauss' least squares method[1] for an assumed relationship or a parametric function, where the simplest one is linear. The term regression was later introduced by Galton around 1880 and a new turn was taken with the non-parametric approach developed by Nadaraya and Watson in the 1960's.

The typical parametric regression assume a specific function $f(x)$ parametrized in terms of a small set of parameters, such as $\alpha + \beta x$ or $e^{\alpha x}$. This is, however, a restrictive assumption, which is not always appropriate. If we instead assume the underlying function to be smooth and approximate $f(x)$ by capturing the patterns in the data, we arrive at the classical smoothing problem. This procedure stands in contrast to parametric curve fitting and instead focuses on important information and takes away noise and fine-scale changes. Many different smoothing methods

[1]Gauss is credited with developing the least squares in 1795 as a tool in astronomy, but Legendre was the first to publish the method in 1805 , when the French government defined the metre.

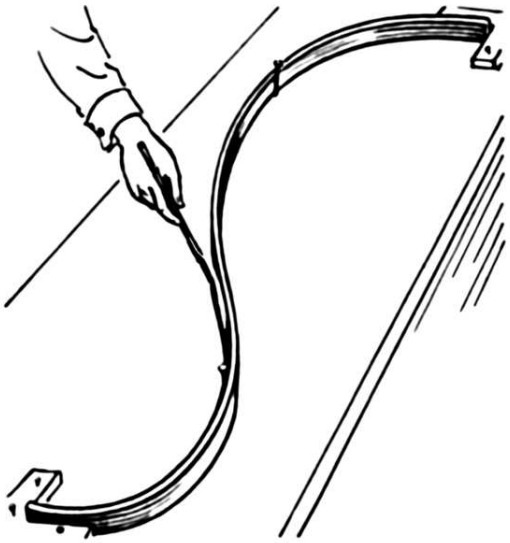


Figure 1.1: A flexible drawing device; the spline.

based on different ideas have been developed, such as kernel smoothers, Kalman filters, wavelets and smoothing splines. In this thesis, we will examine the use of smoothing splines from several different perspectives.

1.1 Smoothing splines

The idea behind smoothing splines is to combine measures of the smoothness of a function and how well it fits the data. For the goodness of fit for the function, we use the residual sum of squares as a criterion. Non-smoothness, like noise and rapid changes, can be suppress by minimizing the penalty $\int f''(x)^2 dx$, the integrated second derivative. Together these two criteria formulate the smoothing spline technique as minimization problem

$$S_\lambda(f) = \sum_{i=1}^{n} \Big(y_i - f(x_i) \Big)^2 + \lambda \int f''(x)^2 dx, \tag{1.1}$$

where the estimated solution is $\hat{f}_\lambda(x) = \arg\min S_\lambda(f)$. It was shown by Schoenberg (1964), that the minimizer of the expression (1.1) is a natural cubic spline with knots at the data points. The origin of the term, spline, comes from a drawing device seen in Figure 1.1, a flexible piece of wood or metal, used by shipbuilders to draw smooth curves. Markers fixed the device at certain points and the flexible material would bend to minimize the internal energy, outlining a smooth curve. A spline is, therefore, a function minimizing an energy.

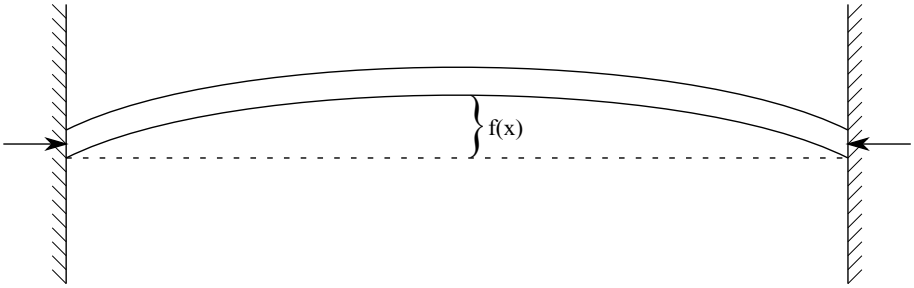


Figure 1.2: A bending rod deviating from neutral position.

The fact that the penalty, $\int f''(x)^2 dx$, represents a bending energy is motivated by linear elasticity theory. This is a field within structural mechanics concerned with bending of materials. Jones (2006) derive the internal energy of a bending rod by relating the deflection $f(x)$ from neutral position, as in Figure 1.2, to the bending moment M and strain energy U. The rotational energy, $E_{rot} = \frac{1}{2}M\theta$, will relate to the deflection $f(x)$ for a small part of the rod. The relationship between the radius of rotation R and the moment M is $M = \frac{c}{R}$, where c is determined by the Young's modulus and moment of inertia, the physical qualities of the beam. With the relation $Rd\theta = ds$, the energy is obtain

$$dE = \frac{1}{2}Md\theta = \frac{c}{2R}d\theta = \frac{c}{2}\frac{d\theta}{ds}d\theta. \tag{1.2}$$

The angle of the bending moment relates to the deflection, $\frac{df(x)}{dx} = \tan\theta$ and the derivative for small angles, $\theta << 1$, give the approximation

$$\frac{d^2f(x)}{dx^2} = \frac{1}{\cos^2\theta}\frac{d\theta}{dx} \approx \frac{d\theta}{dx}. \tag{1.3}$$

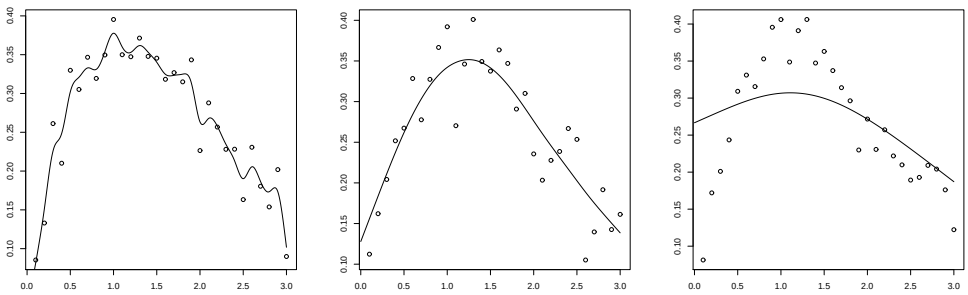
For small angles and deflections, we use the approximations $ds \approx dx$ and $\frac{d\theta}{ds} \approx \frac{d\theta}{dx}$, which yield

$$dE \approx \frac{c}{2}\frac{d\theta}{dx}d\theta = \frac{c}{2}\left(\frac{d^2f(x)}{dx^2}\right)^2 dx. \tag{1.4}$$

The total bending energy of the rod is then proportional to

$$E(x) \propto \int \left(\frac{d^2f(x)}{dx^2}\right)^2 dx. \tag{1.5}$$

Under the idealized physical conditions used in linear elasticity theory, the bending energy is proportional to the integrated square curvature. This underlines the fact that the penalty in (1.1) is an energy. The true bending energy for the physical



(a) (b) (c)

Figure 1.3: Examples of smoothing: a) Under-smoothing. b) Proper smoothing. c)Over-smoothing.

device, however, will be non-linear, but we still derive our penalty from the physical motivation.

The spline estimate is the trade-off between interpolating the data and minimizing the bending energy of the function. A key point is the smoothing parameter λ, which controls this trade-off between goodness-of-fit and roughness, as seen in Figure 1.3. If the smoothing decreases, $\lambda \to 0$, then noise is seen as patterns, as seen in Figure 1.3a), and if $\lambda \to \infty$, then all structures are smoothed out, leaving only a linear regression estimate, as seen in Figure 1.3c).

B-splines

The estimate $\hat{f}_\lambda(x)$ is a cubic spline, meaning the function is a piecewise cubic polynomial. It is divided into intervals associated with a knot sequence

$$a = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_{k-1} \leq \tau_k = b. \tag{1.6}$$

The spline can be represented in terms of a basis, which give the term B-spline. The function $S(x)$ is specified by a coefficient or weight vector $\mathbf{w} = [w_1, \ldots, w_n]^T$ and the basis matrix $B(x) = [B_{1,d}(x), \ldots, B_{n,d}(x)]^T$, where $B_{i,d}(x)$ are i basis functions. Figure 1.4 shows a set of uniform basis functions. The polynomial spline is given by

$$S(x) = \sum_{j=1}^{n} B_j(x) w_j. \tag{1.7}$$

A B-spline on an interval $[a, b]$ has order $m = d + 1$, where d is the degree of the polynomials, and number of internal knots k. An augmented knot sequence τ_i is defined by placing m equal boundary knots on the end points and such a knot sequence can support a basis of order l up to $l \leq m$

$$a = \tau_1 = \cdots = \tau_m < \tau_{m+1} < \cdots < \tau_{k+m} < \tau_{k+m+1} = \cdots = \tau_{K+2m} = b. \tag{1.8}$$

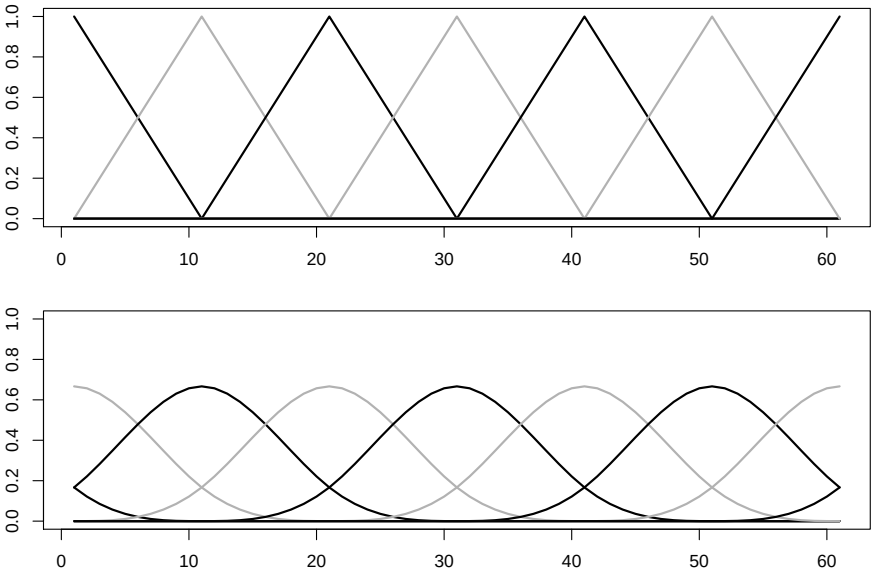


Figure 1.4: Uniform basis splines with degree one and three, in the upper and lower panel, respectively.

Each duplication of the boundary knots results in the loss of one continuous derivative. The number of basis function B_i supported by the augmented sequence is $k + m$. This can be shown by counting the parameters needed to be specified. We will have $k + 1$ regions multiplied with m function parameters per region, but we must subtract the parameters specified by the continuity constraints given by k internal knots multiplied by $m - 1$ constraints on derivatives per knot. So the number of parameters and basis functions are

$$(k + 1) \cdot m - k \cdot (m - 1) = k + m. \tag{1.9}$$

The basis is found by the Cox-de Boor recursion formula for $i = 1, \ldots, k+2m-1$:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if} \quad \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

$$B_{i,m}(t) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x). \tag{1.11}$$

If the knot sequence is uniform, the basis functions become shifted copies of another.

1.2 Penalized splines: a Bayesian approach

In this section, we will introduce the main focus in this thesis, the Bayesian formulation of smoothing splines. Mainstream spline methods use B-splines to represent $f(x)$ in (1.1) and approximate differently the smoothing penalty to achieve feasible calculations. However, the problem can be formulated quite differently, as Wahba (1978) showed. An equivalent smoothing spline formulation with a exact solution involves a Bayesian hierarchical model with a Gaussian process prior, given as

$$f(x) = \theta_1 + \theta_2 x + b^{-\frac{1}{2}} F(x), \tag{1.12}$$

$$y_i = f(x_i) + \epsilon, \tag{1.13}$$

where $x \in [0,1]$, $i = 1, \ldots, n$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The coefficients θ_1 and θ_2 can be fixed and unknown or random, $b > 0$ is a precision parameter and

$$F(x) = \int_0^1 (x-t)_+ dW(t), \tag{1.14}$$

where $(\cdot)_+ = \max(0, \cdot)$. $F(x)$ is an one-fold integrated Wiener process and the solution of the stochastic differential equation (SDE)

$$\frac{d^2 f(x)}{dx^2} = \frac{dW(x)}{dx}. \tag{1.15}$$

The motivations behind this stochastic process is in fact Taylor's theorem. If $f(x)$ is a function on $[0,1]$ with two continuous derivatives and $f''(x) \in \mathcal{L}^2[0,1]$, then

$$f(x) = \{f(0) + f'(0) \cdot x\} + \left\{ \int_0^1 (x-t)_+ f''(t) dt \right\}. \tag{1.16}$$

We see that the function $f(x)$ can be decomposed in two parts, where the second part represent all functions with the condition $f(0) = f'(0) = 0$. Wahba (1990) showed, with great mathematical rigor, that these functions span the same space as all possible sample paths form the process $F(x)$. When $f(x)$ solves the SDE (1.15), the second part of (1.16) is equivalent to the stochastic process $F(x)$

$$\int_0^1 (x-t)_+ \frac{d^2 f(t)}{dt^2} dt = \int_0^1 (x-t)_+ \frac{dW(t)}{dt} dt \tag{1.17}$$

$$= \int_0^1 (x-t)_+ dW(t), \tag{1.18}$$

underlining that $F(x)$ is a prior for $f(x)$.

We define $\hat{F}(x)$ as the minimum variance, unbiased linear estimate of $F(x)$, when given

$$\hat{F}(x) = \sum_{j=1}^{n} \beta_j(x) y_j \tag{1.19}$$

and it minimizes the variance $E(\hat{F}(x) - F(x))^2$ with respect to θ, $E(\hat{F}(x)|\theta) = E(F(x)|\theta)$. Wahba (1978) showed, that if $\hat{F}(x)$ is given by a set of responses y_i, and $f_\lambda(x)$ is the minimizer of

$$\sum_{i=0}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(u))^2 du, \tag{1.20}$$

then

$$\hat{F}(x) = f_\lambda(x), \quad \lambda = \sigma^2 b. \tag{1.21}$$

This means the smoothing spline solution can be found by solving a SDE and a Bayesian model. It is not immediately obvious that this is a practical observation, but it will show its usefulness in Section 2.3.

Bayesian inference

With the latent field specified by Wahba (1978), the smoothing problem can be reformulated a Bayesian way. The starting point of Bayesian statistics is the desire to incorporate known information and therefore the parameters are assigned densities $\pi(\boldsymbol{\theta})$. These distributions are dependent on some hyperparameters, which are assumed to be known. A Bayesian hierarchical model introduces a latent variable, in addition to the observations and hyperparameter,

Observations:	$\mathbf{y}\|\mathbf{f} \sim \pi(\mathbf{y}\|\mathbf{f})$,	(1.22)
Latent variable:	$\mathbf{f}\|\boldsymbol{\theta} \sim \pi(\mathbf{f}\|\boldsymbol{\theta})$,	(1.23)
Parameters:	$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.	(1.24)

The distribution of latent variable and parameters $\pi(\mathbf{f}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ represent $f(x)$ and θ before the observations are done, and they are therefore called priors. But we want to find the distributions after observing $\mathbf{y}$, the posterior distributions. Therefore we use Bayes' theorem

$$\pi(a|b) = \frac{\pi(b|a)\pi(a)}{\pi(b)}, \tag{1.25}$$

where $\pi(b) = \int \pi(b|a)\pi(a)da$. Since $\pi(b)$ is independent of a, we write

$$\pi(a|b) \propto \pi(b|a)\pi(a), \tag{1.26}$$

linking the posterior, $\pi(a|b)$, to the prior distribution $\pi(a)$.

The posterior distribution of $\mathbf{f}$ and $\boldsymbol{\theta}$ is then given

$$\pi(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{f})\pi(\mathbf{f}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.27}$$

The observational distribution, $\pi(\mathbf{y}|\mathbf{f})$, is the probability density of the observed $\mathbf{y}$ given $\mathbf{f}$ and for classical regression $y = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$ gives,

$$\pi(\mathbf{y}|\mathbf{f}) = \mathcal{N}(f(x), \sigma^2). \tag{1.28}$$

We wish to have a prior distribution for the latent variable $f(x)$. Therefore, we involve the concept of stochastic processes, which makes it possible to describe the statistical distribution of a function. A sample path of the process represent a possible function.

The Wiener process

In order to compute Bayesian smoothing splines, we will use some basic results from the theory of stochastic processes. A stochastic process $\{F(x)|x \in X\}$ is a family of random variables $F(x)$ on the index set $x \in X$ defined on the probability space $(\Omega, \mathcal{F}, P)$. It is function, such that $F(x_0, \cdot)$ for a fixed x_0 is a random variable and $F(\cdot, \omega_0)$ for a fixed $\omega_0 \in \omega$ is a sample path. The Gaussian process is defined in the following way:

Definition 1.2.1. *(Lindgren, 2010) A stochastic process $\{F(x), x \in \mathbb{R}\}$ is a Gaussian process if every linear combination $S = \sum_k a_k F(x_k)$ for real a_k and $x_k \in \mathbb{R}$ has a Gaussian distribution.*

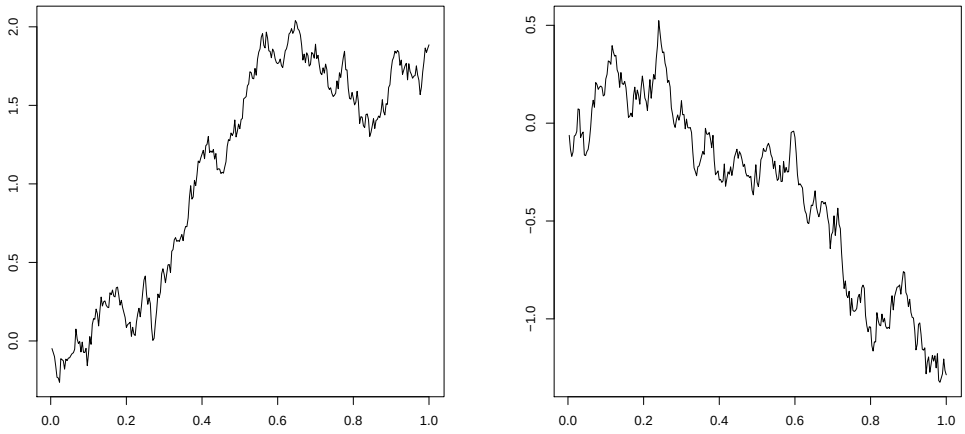
It follows, if they exist, that the derivative and integral of a Gaussian process are also Gaussian. The derivative is the limit of $\frac{F(x+h)-F(x)}{h}$ as $h \to 0$.There are several important Gaussian processes with different properties and characteristics, such as the Wiener process, as seen in Figure 1.5.

Definition 1.2.2. *(Lindgren, 2010) The Wiener process $\{W(x), 0 \leq x\}$ is a Gaussian process with $W(0) = 0$ giving $\mathbb{E}(W(x)) = 0$ and the variance of the increment $W(x+h) - W(x)$ for $h > 0$ is proportional to h,*

$$\mathrm{Var}(W(x+h) - W(x)) = h\sigma^2. \tag{1.29}$$

The Wiener process is characterized by the following properties:

- Independent increments: $W(t) - W(s)$ is independent of $\{W(\tau)\}_{\tau \leq s}$ for $0 \leq s \leq t$.
- Stationarity: The distribution of $W(t) - W(s)$ is independent of s.
- Continuity: $W(t)$ is almost surely continuous seen as a function of t.
- $W(t)$ is Gaussian, $W(t) \sim \mathcal{N}(0, t)$, and the covariance is $\mathbb{E}\, W(t)W(s) = \min(s, t)$.



(a) (b)

Figure 1.5: Sample paths from a standard Wiener process.

Gaussian white noise

Another important concept is the "derivative" of a Wiener process, Gaussian white noise, as seen in Figure 1.6. The sample paths of a Wiener process are continuous functions, but they are not differentiable. However, we can define the generalized derivative, $W'(x)$, using an integration-by-parts style formula with a smooth function $g(x)$

$$g(x)W(x) = \int_0^x g(t)W'(t)dt + \int_0^x g'(t)W(t)zdt. \tag{1.30}$$

Since $W(x)$ is a Gaussian process, the derivative $W'(x)$ will be Gaussian with expectation and generalized covariance

$$\mathbb{E}\, W'(x) = 0, \tag{1.31}$$
$$\mathbb{E}(W'(x)W'(x')) = \delta(x - x'). \tag{1.32}$$

The solution of the stochastic differential equation presented in this thesis is an integrated Wiener process, defined by integrals of the form

$$G_i = \int g_i(t)W'(t)dt. \tag{1.33}$$

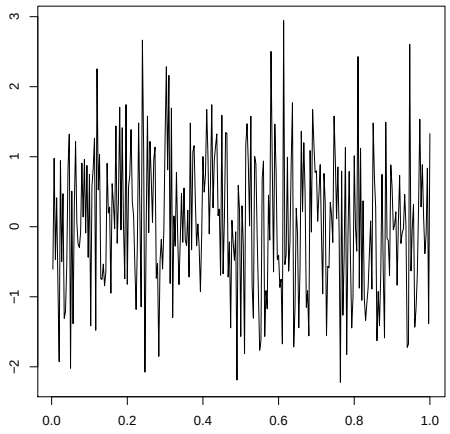


Figure 1.6: Gaussian white noise, the derivative of a Wiener process.

The process G_i will, by construction, have a Gaussian distribution with zero-mean and covariance defined by

$$\begin{aligned}\mathrm{Cov}\,(G_i, G_j) &= \mathbb{E}\left(\int g_i(t)W'(t)dt \int g_j(t)W'(t)dt\right) &(1.34)\\ &= \int\int g_i(t)g_j(t')\,\mathbb{E}(W'(t)W'(t'))dtdt', &(1.35)\\ &= \int g_i(t)g_j(t)dt. &(1.36)\end{aligned}$$

The process G_i, defined by the function g_i, has a Gaussian distribution with expectation and covariance

$$\begin{aligned}\mathbb{E}\,G_i(x) &= 0 &(1.37)\\ \mathrm{Cov}(G_i, G_j) &= \int g_i(t)g_j(t)dt. &(1.38)\end{aligned}$$

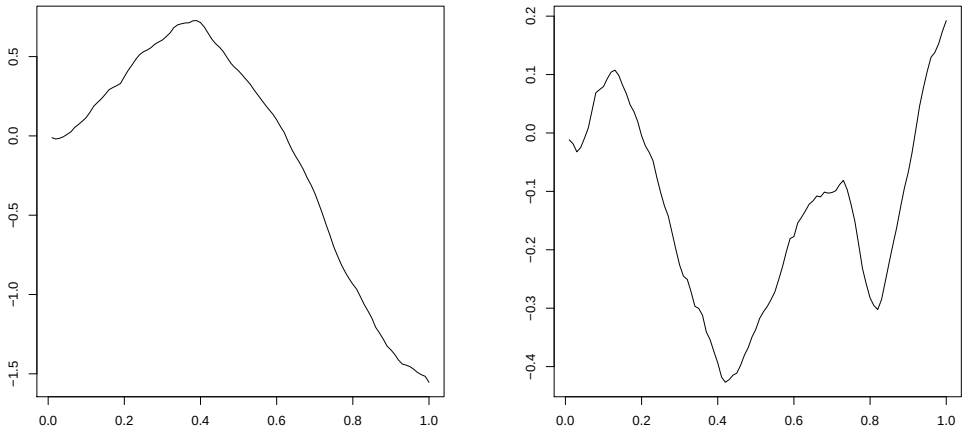
Discrete processes: The second order random walk

Random walk is an example of a discrete stochastic process, and it is equivalent to the discrete observations of a Wiener process. A random walk is defined as

$$z_t = z_{t-1} + \epsilon_t, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1.39}$$

When $z_0 = 0$, we find by recursion the following

$$z_t = \sum_{i=1}^{t} \epsilon_i, \tag{1.40}$$



(a) (b)

Figure 1.7: Sample paths from a integrated Wiener process, a continuous second-order random walk.

which give $\mathbb{E}\, z_t = 0$ and $\operatorname{Var} z_t = t\sigma^2$, as the Wiener process. The error ϵ_i corresponds to Gaussian white noise. In parallel, we introduce the process corresponding to discrete observations of the latent field $F(x)$, the integrated Wiener process, given by

$$F(x) = \int_0^1 (x-t)_+ W'(t)dt, \tag{1.41}$$

as seen in Figure 1.7.

This discrete process is the second-order random walk and we construct the model by independent, identically distributed second-order increments, following the notation from Rue and Held (2005)

$$\Delta^2 f_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, \ldots, n-2. \tag{1.42}$$

This can be rewritten as

$$f_t = 2f_{t-2} - f_{t-1} + \epsilon_t, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1.43}$$

It is easy to show that this process is the cumulative sum of a random walk, just as the random walk is cumulative sum of Gaussian variables,

$$f_t = \sum_{i=1}^{t} z_i = \sum_{i=1}^{t-1} z_i + z_t \tag{1.44}$$

$$= f_{t-1} + z_{t-1} + \epsilon_t = f_{t-1} + f_{t-1} - f_{t-2} + \epsilon_t \tag{1.45}$$

$$= 2f_{t-2} - f_{t-1} + \epsilon_t. \tag{1.46}$$

The second order random walk "integrates" a random walk, corresponding to it being the discrete observations of a integrated Wiener process $F(x)$. When we in addition know that $F(x)$ is the solution of

$$\frac{d^2 f(x)}{dx^2} = \frac{dW(x)}{dx}, \tag{1.47}$$

calling f_t of second-order is quite natural.

Rue and Held (2005) define the second-order walk in connection to Gaussian Markov random fields (GMRF), which make another aspect of the result (1.13) become clear. They define the observational distribution density in a Bayesian setting with increment precision $\tau = 1/\sigma^2$:

$$\pi(\mathbf{x}|\kappa) \propto \kappa^{(n-2)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-2}(x_{i-2} - 2x_{i-1} + x_i)\right), \tag{1.48}$$

$$= \kappa^{(n-2)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=2}^{n-1}(\Delta^2 x_i)\right), \tag{1.49}$$

$$= \kappa^{(n-2)/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right), \tag{1.50}$$

where the precision matrix is given as

$$\mathbf{Q} = \kappa \begin{bmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{bmatrix}. \tag{1.51}$$

The rank of $\mathbf{Q}$ is $n-2$, which makes it improper with rank deficiency two and an example of an intrinsic GRMF of second-order. This means the density is invariant to the addition of any line $\alpha+\beta x$ to $\mathbf{x}$, which is the property highlighted by Wahba in (1.13) by $\theta_1 + \theta_2 x$.

1.3 Integrated nested Laplace approximation (INLA)

Rue and Martino (2009) have developed the `R` computer package INLA for approximate Bayesian inference using integrated nested Laplace approximations. This

handles general Gaussian hierarchical models, such as the specified Bayesian formulation of smoothing splines. INLA compute the credible bands from the posterior densities, making it possible to quantify the uncertainty in the spline estimate.

The general Gaussian hierarchical model have a hyperparameter θ with prior $\pi(\theta)$, a latent variable $\mathbf{f}$ with density $\pi(\mathbf{f}|\theta)$ and an observed response $\mathbf{y}$ with likelihood $\pi(\mathbf{y}|\mathbf{f})$. The posterior is then given

$$\pi(\mathbf{f}, \theta|\mathbf{y}) \propto \pi(\theta)\pi(\mathbf{f}|\theta)\pi(\mathbf{y}|\mathbf{f}), \tag{1.52}$$

and we want to find the posterior marginals $\pi(f_i|\mathbf{y})$ and $\pi(\theta_i|\mathbf{y})$. This can by done by using INLA, which compute the approximated marginals directly without using MCMC-methods. The classical Laplace approximation of an integral uses the Taylor expansion of $\ln f(x) = g(x)$ around x_0,

$$\ln f(x) = g(x_0) + g'(x_0) \cdot (x - x_0) + \frac{1}{2} g''(x_0) \cdot (x - x_0)^2 + \dots \tag{1.53}$$

If x_0 is taken to be the mode of $\ln f(x)$, where $f(x)$ is a unimodal density function, the first derivative must be zero, $\frac{\partial \ln f(x)}{\partial x}\Big|_{x=x_0} = 0$. Then the approximation is a Gaussian density. The posterior marginal of interest can be written as

$$\pi(f_i|\mathbf{y}) = \int \pi(f_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta, \tag{1.54}$$

$$\pi(\theta_i|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta, \tag{1.55}$$

and the approximations done by INLA construct nested approximation with $\tilde{\pi}$ as an approximated density

$$\tilde{\pi}(f_i|\mathbf{y}) = \int \tilde{\pi}(f_i|\theta, \mathbf{y})\tilde{\pi}(\theta|\mathbf{y})d\theta, \tag{1.56}$$

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \int \tilde{\pi}(\theta|\mathbf{y})d\theta. \tag{1.57}$$

An approximation of $\pi(f_i|\mathbf{y})$ is computed by approximating $\pi(f_i|\theta, \mathbf{y})$ and $\pi(\theta|\mathbf{y})$ and using numerical integration to integrate out the parameter θ. This nested approach make the Laplace approximations very accurate. The $\tilde{\pi}(\theta_i|\mathbf{y})$ is approximated by

$$\tilde{\pi}(\theta_i|\mathbf{y}) \propto \frac{\pi(\mathbf{f}, \theta, \mathbf{y})}{\pi_G(\mathbf{f}|\theta, \mathbf{y})}\Bigg|_{f=f^*(\theta)}, \tag{1.58}$$

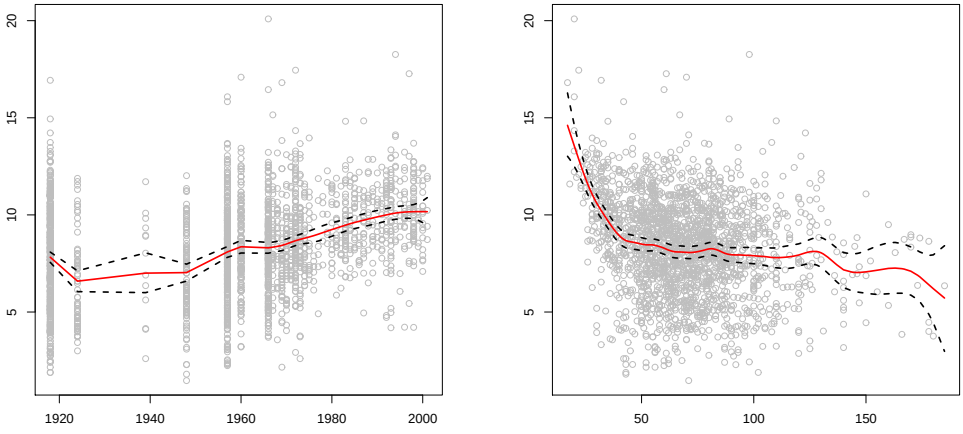
where $\pi_G(\mathbf{f}|\theta, \mathbf{y})$ is the Gaussian approximation and $\mathbf{f}^*(\theta)$ is the mode.

The Munich rental guide

INLA is implemented in `R` and we illustrate with a classical example, the Munich rental guide. According to German law, flat owners can increase the rent based on the average of comparable flats. Therefore, the public provides a rental guide with the average rent per square meter given by several housing variable, such as the flat size and building year. The data provided come from two thousand flats in Munich and we estimate a smooth trend, using the following `R` code with the `rw2` option:

```
library(INLA)
data(Munich)
x <-  Munich$year; y <-  Munich$rent;
data <- data.frame(y=y,x=x)
formula <- y~f(x,model="rw2",prior="loggamma",param=c(1,0.01))
result <- inla(formula,data=data,family="gaussian")
```

The prior on the smoothing parameter is Gamma(α, β), where $\alpha = 1$ and $\beta = 10^{-2}$, as specified in (2.10). The plotted results in Figure 1.8 display the smooth trend between the variables, giving flat owners the opportunity to set rent based on the estimated curve. The dashed lines show the 95 % credible bands, expressing how precise the estimate is considered to be.



(a) (b)

Figure 1.8: The Munich rent guide: a) Building year against rent per square meter. b) Floor size against rent per square meter. The dashed line show the 95% credible bands.

1.4 Outline of thesis

In Chapter 1, we have introduced the concept of smoothing splines as an energy minimizer and the Bayesian formulation in terms of a SDE. Chapter 2 will explore three different approximative smoothing splines, two frequentist methods, P-splines and O'Sullivan splines and a Bayesian approach, the RW2 model. Graphical examples made using INLA, will showcase the comparison between the methods and differences and advantages are discussed. Chapter 3 investigate the relationship between O'Sullivan splines and the RW2 model, and shows that the O'Sullivan splines are a finite element approximation to smoothing splines. The extension of to Bayesian adaptive smoothing, is explored in Chapter 4 and in Chapter 5 the different results are discussed.

Chapter 2

Stochastic models for splines

In this chapter, we will present two different penalized spline methods used in regression, given as

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{2.1}$$

These two mainstream methods give computationally feasible versions of the smoothing splines, formulated as

$$S_\lambda(f) = \sum_{i=1}^{n} \Big(y_i - f(x_i) \Big)^2 + \lambda \int f''(x)^2 dx, \tag{2.2}$$

where the function $f(x)$ and the penalty are approximated in different ways. Both methods, P-splines and O'Sullivan splines, use a basis spline representation with the knot sequence of τ_i with k internal knots, order m and $l = m + k$, for $f(x)$

$$\hat{f}(x) = \sum_{j=1}^{l} w_j B_j(x), \tag{2.3}$$

where w_j is the coefficient of the jth basis function $B_j(x)$. The mainstream methods are frequentist and formulate spline smoothing as a minimization problem of a specified objective function, where one part represent the goodness-of-fit and the other quantify the roughness of the estimate. The function is parametrized in terms of a set of basis functions, reducing the problem to finding the weights $\mathbf{w}$. The penalty, essentially on $\mathbf{w}$, must be approximated, making the problem a system of linear equations, where the minimization is done along the lines of the ordinary normal equations. The objective function is given in matrix notation

$$S_\lambda(\mathbf{w}) = (\mathbf{y} - \mathbf{B}\mathbf{w})^T(\mathbf{y} - \mathbf{B}\mathbf{w}) + \lambda\, \mathbf{w}^T \mathbf{\Omega}\, \mathbf{w}, \tag{2.4}$$

where $\hat{\mathbf{w}}_\lambda = \arg\min_w S(\mathbf{w})$ and $\mathbf{\Omega}$ is a matrix representing the approximation of the penalty. P-splines approximate the penalty by higher-order differences on the coefficients $\mathbf{w}$, while the O'Sullivan splines use an approximation along the lines of a finite element method.

The Bayesian approach

The second half of this chapter will be use to explore a method based on the results of Wahba (1978). Mainstream penalized splines methods focus on the minimization of an objective function, which can be reinterpreted as a Bayesian model. A field prior, described by a stochastic differential equation, specify the statistical distribution of $\mathbf{w}$ and $f(x)$. This gives a hierarchical Bayesian model and we reformulate the two frequentist methods to evaluate them within the Bayesian framework.

The Bayesian hierarchical model has the following distributions:

$$\mathbf{y}|\mathbf{w} \sim \pi(\mathbf{y}|\mathbf{w}) = \mathcal{N}\Big(\mathbf{Bw}, \sigma^2 I\Big) \tag{2.5}$$

$$\mathbf{w}|b \sim \pi(\mathbf{w}|b) = \mathcal{N}\Big(0, (b\mathbf{\Omega})^{-1}\Big) \tag{2.6}$$

$$b \sim \pi(b) = \text{Gamma}(\alpha, \beta), \tag{2.7}$$

giving the posterior distribution

$$\begin{aligned}\pi(\mathbf{w}, \lambda|\mathbf{y}) &\propto \pi(\mathbf{y}|\mathbf{w})\pi(\mathbf{w}|b)\pi(b) \\ &\propto \pi(b)|\mathbf{\Omega}|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Bw})^T(\mathbf{y}-\mathbf{Bw}) - \frac{b}{2}\,\mathbf{w}^T\mathbf{\Omega}\,\mathbf{w}\right\}. \end{aligned} \tag{2.8}$$

The solution of $\hat{\mathbf{w}}$ is found to be the maximum *a posteriori* probability (MAP) estimate. This shows equivalence between the Bayesian model and the objective function, when $\lambda = \sigma^2 b$ and we fix b to be constant,

$$\begin{aligned}\hat{\mathbf{w}}_\lambda &= \arg\min_w \left\{(\mathbf{y}-\mathbf{Bw})^T(\mathbf{y}-\mathbf{Bw}) + \lambda \mathbf{w}^T\mathbf{\Omega}\,\mathbf{w}\right\}, \\ &= \arg\max_w \left\{-\frac{1}{2}(\mathbf{y}-\mathbf{Bw})^T(\mathbf{y}-\mathbf{Bw}) - \frac{\sigma^2 b}{2}\,\mathbf{w}^T\mathbf{\Omega}\,\mathbf{w}\right\}, \\ &= \arg\max_w \left\{e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Bw})^T(\mathbf{y}-\mathbf{Bw}) - \frac{b}{2}\mathbf{w}^T\mathbf{\Omega}\,\mathbf{w}}\right\}, \\ &= \hat{\mathbf{w}}_{MAP}. \end{aligned} \tag{2.9}$$

The MAP estimates and the posterior means can be found by using the integrated nested Laplace approximations (INLA) package in `R`. One important aspect is the choice of smoothing parameter. The usual frequentist approach is to use a general cross-validation scheme, based on the removal of a single data point and optimizing for the prediction power.

All calculations carried out in INLA use a Gamma(α, β) prior on the smoothing parameter b

$$\pi(b) = \frac{\beta^\alpha}{\Gamma(\alpha)} b^{\alpha-1} e^{-\beta b}, \tag{2.10}$$

with mean $\frac{\alpha}{\beta}$, variance $\frac{\alpha}{\beta^2}$ and mode $\frac{\alpha-1}{\beta}$. The internal structure of INLA reparametrize b is as $\theta = \log b$, therefore the option is specified by `prior=loggamma`, for further details see (Rue and Martino, 2009). For simplicity, we will only look at examples where the data points x_i are placed on knots τ_i with irregular and regular data. Additionally, we place one knot between each data point, resulting in twice as many knots as data.

2.1 Difference penalty

In the papers *Flexible smoothing with B-splines and penalties* (Eilers and Marx, 1996) and *Spline, knots and penalties* (Eilers and Marx, 2005) present P-splines, which use B-splines with uniform knots and difference penalties. Eilers and Marx (2005) compare the method with the use of truncated power functions with knots based on quantiles of the data and a ridge penalty, presented in (Ruppert et al., 2003) and conclude that uniform B-splines are to be preferred. Their goal is to highlight the differences between penalized B-splines and penalized truncated power functions and to make a plea for equidistant knots.

P-splines are constructed on a basis of quadratic or cubic splines using equally-spaced knots with coefficients $\mathbf{w}$ and basis matrix $\mathbf{B}$ computed on x,

$$y_i = \sum_{j=1}^{l} B_j(x_i) w_j = \mathbf{Bw}. \tag{2.11}$$

The penalty matrix $\mathbf{\Omega}$ from (2.4) is the kth differences of the B-spline coefficients corresponding to minimizing of the objective function

$$S(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{B}(x_i)\mathbf{w})^2 + \lambda \sum_{j=k+1}^{l} (\Delta^k w_j)^2 \tag{2.12}$$

$$= (\mathbf{y} - \mathbf{Bw})^T (\mathbf{y} - \mathbf{Bw}) + \lambda\, \mathbf{w}^T (\mathbf{D}_k^T \mathbf{D}_k) \mathbf{w} \tag{2.13}$$

where Δ^k is the kth difference. The first and second difference is

$$\Delta w_i = w_i - w_{i-1}, \tag{2.14}$$
$$\Delta^2 w_i = w_i - 2w_{i-1} + w_{i-2}. \tag{2.15}$$

The vector of differences Δ^k is denoted by the matrix $\mathbf{D}_k$, giving the penalty

matrix $\Omega = \mathbf{D}_k^T \mathbf{D}_k$. For $k = 2$ the penalty becomes

$$\mathbf{\Omega} = \begin{bmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{bmatrix}. \tag{2.16}$$

The minimizer of the objective function $\hat{\mathbf{w}}$ is given by

$$(\mathbf{B}^T\mathbf{B} + \lambda \mathbf{D}_k^T \mathbf{D}_k)\hat{\mathbf{w}} = \mathbf{B}^T\mathbf{y}, \tag{2.17}$$

and the resulting estimate becomes

$$\hat{f}(x) = \sum_{j=1}^{l} \hat{\mathbf{w}} B_j(x) \tag{2.18}$$

$$= \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda \mathbf{D}_k^T \mathbf{D}_k)^{-1}\mathbf{B}^T\mathbf{y}. \tag{2.19}$$

The parameter λ tunes the smoothness and $\lambda \to 0$ the reduce the system to a linear regression on the basis functions $B_j(x)$. When the smoothing increases, $\lambda \to \infty$, the data is fitted to a linear ($k = 2$) or quadratic ($k = 3$) curve. The penalty λ insures that the matrix $\mathbf{B}^T\mathbf{B} + \lambda \mathbf{D}^T\mathbf{D}$ is not singular in (2.17), but is also understood as the precision of the prior distribution of the coefficients $\pi(\mathbf{w}|b)$.

The differencing penalty of the weights $\mathbf{w}$ is difficult to interpret, but can be seen as an discrete approximation of dth derivative with equidistant knots. We can see this for a special case, applying the Cox-de Boor formula for derivatives of B-splines as done in (Eilers and Marx, 1996), yielding

$$\sum_i w_i B_i'(x; m) = -\frac{1}{h} \sum_i \Delta w_{i+1} B_i(x; m-1). \tag{2.20}$$

With cubic B-splines, the penalty becomes

$$\begin{aligned} h^2 \int_a^b f''(x)^2 dx &= h^2 \int_a^b \left\{ \sum_i w_i B_i'(x,3) \right\}^2 dx = \int_a^b \left\{ \sum_i \Delta^2 w_i B_i(x,1) \right\}^2 dx \\ &= \int_a^b \left\{ \sum_i \sum_j \Delta^2 w_i \Delta^2 w_j B_i(x,1) B_j(x,1) \right\} dx \\ &= \int_a^b \left[\sum_i (\Delta^2 w_i B_i(x,1))^2 + 2 \sum_i \Delta^2 w_i \Delta^2 w_{i-1} B_i(x,1) B_{i-1}(x,1) \right] dx \\ &= c_1 \sum_i (\Delta^2 w_i)^2 + c_2 \sum_i \Delta^2 w_i \Delta^2 w_{i-1}, \end{aligned} \tag{2.21}$$

where c_1 and c_2 for equidistant knots are constant

$$c_1 = \int_a^b B_i(x,1)^2 dx, \quad c_2 = \int_a^b B_j(x,1) B_{j-1}(x,1) dx. \tag{2.22}$$

The first term of (2.21) is equivalent to the difference penalty with $d = 2$, meaning it approximates the integrated penalty without cross products, the effect from overlapping, neighboring basis functions. Eilers and Marx (2005) shows that B-splines with equally spaced knots can be constructed using truncated power functions $f_{j,d}(x) = (x - t_j)_+^d$

$$B_{j,d}(x) = (-1)^{d+1} \Delta^{d+1} \frac{f_{j,d}(x)}{h^d d!}, \tag{2.23}$$

where h is the distance between knots. This is a simplification of the Cox-de Boor algorithm due to the equal spacing. They claim that quantile-bases knots "underestimates the power of the penalty" (Eilers and Marx, 2005). If the model make sense continuously, it should be independent, within reason, of the knot sequence. There is also a problem with the waste of calculation power. Uniform knots will place structure, where there is possibly no need for it.

P-splines are based on equidistant knots and a discrete difference penalty, which falls easily within the common framework of statistics. The greatest advantages of the method is that it is easily programmed. To obtain the penalty matrix, we need no calculations or any knowledge of the basis functions. The free choice of the differencing order, $\mathbf{D}_d$, also gives the possibility of a higher order smoothing penalty and increased flexibility. The choice $d = 2$ is however the most natural, being the equivalent of the second derivative.

2.2 O'Sullivan penalty

The paper *On semiparametric regression with O'Sullivan penalized splines* (Wand and Ormerod, 2008) introduce penalized splines using quantile-based, non-uniform

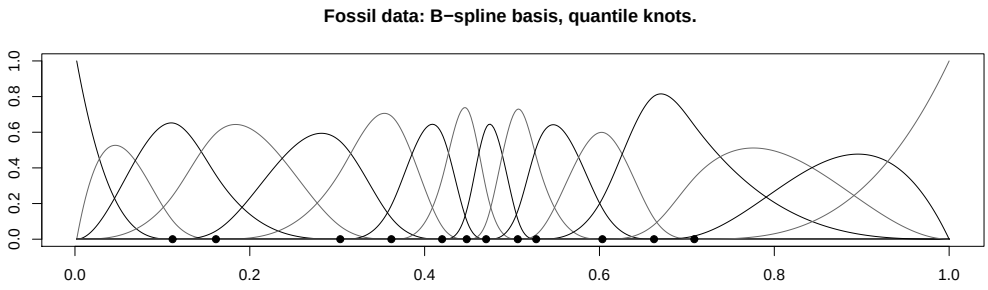


Figure 2.1: Example of 16 basis functions for fossil data with 12 quantile-based knots, indicated by circle.

knots. The penalty is based on the integral of B-spline basis functions, as introduced by O'Sullivan, and the spline estimate satisfies natural boundary conditions. The cubic B-spline basis functions $B_1, \ldots, B_{K+4}$ are defined by a knot sequence given as (1.8). The knots κ_i are chosen to be the $\frac{i}{K+1}$th sample quantile of the unique data points x_i, where

$$K = \begin{cases} n & n < 50 \\ 100 & n = 200 \\ 140 & n = 800 \\ 200 + (n - 3200)^{1/5} & n > 3200. \end{cases} \tag{2.24}$$

Figure 2.1 shows the basis functions with 16 knots for a dataset of fossil shells, presented in Section 2.5. The design matrix $\mathbf{B}$ has entries $B_{ij} = B_i(x_j)$ and the $(K+4) \times (K+4)$ penalty matrix $\mathbf{\Omega}$ is defined by

$$\mathbf{\Omega}_{ij} = \int_a^b B_i''(x) B_j''(x) dx. \tag{2.25}$$

The estimate $\hat{f}(x; \lambda)$ is the minimizer of

$$S(\mathbf{w}) = (\mathbf{y} - \mathbf{B}\mathbf{w})^T (\mathbf{y} - \mathbf{B}\mathbf{w}) + \lambda \, \mathbf{w}^T \mathbf{\Omega} \, \mathbf{w}, \tag{2.26}$$

giving

$$\hat{f}(x; \lambda) = \mathbf{B}\hat{\mathbf{w}}, \quad \text{where} \quad \hat{\mathbf{w}} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{\Omega})^{-1}\mathbf{B}^T\mathbf{y}, \tag{2.27}$$

where $\mathbf{B} = [B_1, \ldots, B_{K+4}]$ and $\lambda > 0$ is the smoothing parameter.

Wand and Ormerod (2008) comment that the cubic smoothing spline arises in the special case $k = n$ and $\tau_{i+4} = x_i$ for $1 \leq k \leq n$, provided that the x_i are distinct.

The estimate $\hat{f}(x;\lambda)$ satisfies the natural boundary condition, a constraint on the derivatives, meaning that

$$\hat{f}''(a;\lambda) = \hat{f}''(b;\lambda) = \hat{f}'''(a;\lambda) = \hat{f}'''(b;\lambda) = 0, \tag{2.28}$$

which implies that $\hat{f}(x;\lambda)$ is approximately linear over the intervals $[a, \kappa_5]$ and $[\kappa_{K+4}, b]$. The linearity is exact if $\kappa_5 = \min(x_i)$ and $\kappa_{K+4} = \max(x_i)$.

The penalty matrix $\mathbf{\Omega}$ can be computed in R by the second derivative design matrix $\mathbf{B}''$ and is given for a cubic basis

$$\mathbf{\Omega} = (\tilde{\mathbf{B}}'')^T \operatorname{diag}(\mathbf{c})\tilde{\mathbf{B}}'', \tag{2.29}$$

where $\tilde{\mathbf{B}}''$ is the $3(K+7) \times (K+1)$ matrix with entries $B_j''(\tilde{x}_i)$, $\tilde{x}_i$ is from the vector

$$\tilde{\mathbf{x}} = \left(\kappa_1, \frac{\kappa_1+\kappa_2}{2}, \kappa_2, \kappa_2, \frac{\kappa_2+\kappa_2}{2}, \ldots, \kappa_{K+7}, \frac{\kappa_{K+7}+\kappa_{K+8}}{2}, \kappa_{K+8}\right) \tag{2.30}$$

and $\mathbf{c}$ is the $3(K+7) \times 1$ vector

$$\mathbf{c} = \left(\frac{1}{6}\Delta\kappa_1, \frac{4}{6}\Delta\kappa_1, \frac{1}{6}\Delta\kappa_1, \ldots, \frac{1}{6}\Delta\kappa_{K+7}, \frac{4}{6}\Delta\kappa_{K+7}, \frac{1}{6}\Delta\kappa_{K+7}\right), \tag{2.31}$$

where $\Delta\kappa_i = \kappa_{i+1} - \kappa_i$. This result is given by applying Simpson's rule over each of the inter-knot difference using the second derivative design matrix to calculate the integrals defining $\mathbf{\Omega}_{ij}$. Since each function $B_i''B_j''$ is piecewise-quadratic, Simpson's rule will calculate the integral exactly.

The O'Sullivan splines use quantile-based knots and the differences between this and equidistant knots will be minor in most situations. The O'Sullivan splines are compared to P-splines by using equally spaced knots, evaluating only the penalty matrices. The differences are relatively small, but give noticeable different results at the boundaries. An empirical study on 18 homoscedastic regression settings with 200 samples estimated the closeness between the estimate and the smoothing spline. In all 72 cases the O'Sullivan-splines were closer to the smoothing splines than P-splines (Wand and Ormerod, 2008). The difference in the two methods is clear at the boundaries, where P-splines deviates from the natural boundary conditions of the smoothing splines. This comes from the discrete penalty approximation near the boundary. The greatest advantages with O'Sullivan splines is the direct use of the B-spline basis functions to calculate the penalty matrix, giving a better approximation to the smoothing spline penalty. The function $f(x)$ is approximated by a weighted set of basis functions, as a parallel to finite element methods.

The calculations are slightly more complicated than for the P-spline difference penalty, but still easily done. The method has added flexibility due to the use of

quantile-based knots, that assign more basis functions to areas with higher density of data. This insures that computer power is not wasted on areas with little data and high uncertainty. Strategically placed knots, can reduced the number of basis functions. Another important feature of O'Sullivan splines are the natural boundary property shared with the smoothing spline. This gives a better behaviour than P-splines near the boundaries.

2.3 Second-order random walk model

After examining the mainstream, frequentist methods based on minimization, we introduce the method based on a Bayesian latent field. Wahba (1978) showed that the following Bayesian prior, gives the same exact solution as smoothing splines:

$$f(x) = \theta_1 + \theta_2 x + b^{-\frac{1}{2}} F(x), \quad x \in [0,1], \tag{2.32}$$

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{2.33}$$

where θ_i can be fixed or random and $F(x)$ is the following stochastic process

$$F(x) = \int_0^1 (x-t)_+ \, dW(t). \tag{2.34}$$

Lindgren and Rue (2008) introduced the second-order random walk (RW2) model with irregular location, using a Galerkin approximation to $F(x)$, as the solution of

$$\frac{d^2 f(x)}{dx^2} = \frac{dW(x)}{dx}. \tag{2.35}$$

With a sequence of fixed, irregular locations $s_1 < s_2 < \cdots < s_n$ and the observations at these locations y_i, we set the model as the discrete observations of $F(x)$. The process is an integrated Wiener process and hence Gaussian. We want to find the statistical properties, the expectation and precision matrix of this stochastic field, which describe the prior of $f(x)$.

For this, we use a Galerkin approach, which is discussed in detail in Chapter 3, to solve the stochastic differential diffusion equation in (2.35). The weak solution of the SDE, denoted in terms of the inner-product $\langle f, g \rangle = \int f(t)g(t)dt$, is defined by the identity

$$\langle \phi(x), y''(x) \rangle = \langle \phi(x), W'(x) \rangle \,, \tag{2.36}$$

which must hold for all appropriate test functions $\phi(x)$. The approximation $\tilde{f}(x)$ to the SDE is constructed as a linear combinations of a set of basis function $\{\psi_i(x)\}, i = 1, \ldots, n$ for some subset of all possible solutions,

$$\tilde{f}(x) = \sum_{i=1}^{n} \psi_i(x) w_i, \tag{2.37}$$

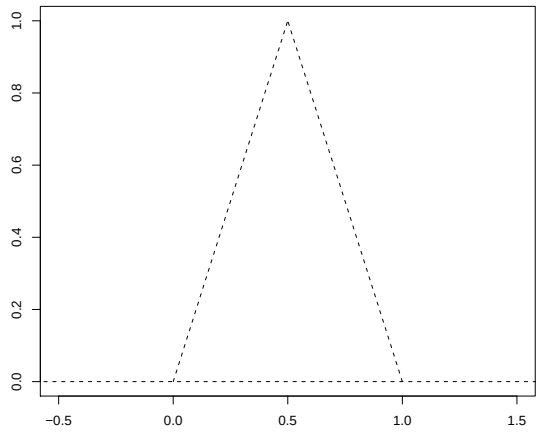


Figure 2.2: A standard piecewise linear Galerkin basis function.

such that the joint distribution of the approximation equals the joint distribution of the true solution

$$\left\langle \psi_i(x), \tilde{f}''(x) \right\rangle \stackrel{d}{=} \langle \psi_i(x), f''(x) \rangle \,, \tag{2.38}$$

where the right-hand side is described in terms of the Gaussian white noise $\langle \psi_i(x), W'(x) \rangle$. The problem is reduced to finding the distribution of the weights $\mathbf{w} = [w_1, \ldots, w_n]^T$.

If $s_1 < s_2 < \cdots < s_n$ is the location sequence and $d_i = s_{i+1} - s_i$ for $i = 1, \ldots, n-1$, the standard set of basis functions is piecewise linear, as seen in Figure 2.2,

$$\psi_i(t) = \begin{cases} 0, & x < s_{i-1}, \\ \frac{x - s_{i-1}}{d_{i-1}}, & s_{i-1} \leq x < s_i, \\ 1 - \frac{x - s_i}{d_i}, & s_i \leq x < s_{i+1}, \\ 0, & s_{i+1} \leq x \end{cases} . \tag{2.39}$$

Now, the Galerkin approximation is expanded in terms of the basis, resulting in

$$\left[\left\langle \psi_i(x), \tilde{f}''(x) \right\rangle \right]_{i=1,\ldots,n} = \left[\sum_j w_j \langle \psi_i(x), \psi''(x) \rangle \right]_{i=1,\ldots,n} = \mathbf{H}\mathbf{w}, \tag{2.40}$$

The integral giving the elements of matrix $\mathbf{H}$ can be simplified by integration-by-parts, using the natural boundary condition $f'(a) = f'(b) = 0$, to yield

$$\mathbf{H}_{ij} = \int_a^b \psi_i(x)\psi''(x)dx = -\int_a^b \psi_i'(x)\psi'(x)dx. \tag{2.41}$$

Therefore, $\mathbf{H}$, is given as a tridiagonal matrix with elements

$$H_{i,i-1} = \frac{1}{d_{i-1}}, \quad H_{i,i} = -\left(\frac{1}{d_{i-1}} + \frac{1}{d_i}\right), \quad H_{i,i+1} = \frac{1}{d_i}, \quad 2 \leq i \leq n-1. \tag{2.42}$$

The first and last row in $\mathbf{H}$ are zero.

Now the right-hand side of (2.38) is given by the statistical properties of Gaussian white noise $W'(x)$ defined in (1.38). Therefore $\langle \psi_i, \frac{dW(t)}{dt} \rangle \quad i = 1, \ldots, n$ will have a Gaussian distribution with expectation 0 and covariance matrix $\mathbf{B}_{ij} = [\langle \psi_i, \psi_j \rangle]$, which give

$$B_{i,i-1} = \frac{d_{i-1}}{6}, \quad B_{i,i} = \frac{d_{i-1} + d_i}{3}, \quad B_{i,i+1} = \frac{d_i}{6} \tag{2.43}$$

with modification at the boundaries. The requirement that $\left[\left\langle \psi_i(x), \tilde{f}''(x) \right\rangle\right]_{i=1,\ldots,n}$ should have the same distribution as $[\langle \psi_i(x), f''(x) \rangle]_{i=1,\ldots,n}$ is fulfilled by the random vector $\mathbf{w}$ with the dense precision matrix

$$\mathbf{Q} = \mathbf{H}^T \mathbf{B}^{-1} \mathbf{H}. \tag{2.44}$$

This makes the Galerkin model computationally expensive, but it is possible to approximate $\mathbf{B}$ with a diagonal matrix $\mathbf{A}$ to obtain a sparse precision matrix. The matrix $\mathbf{A}$ is constructed by approximating the integrated functions of $\mathbf{B}$ with constants, giving the non-zero elements

$$A_{i,i} = \frac{d_{i-1} + d_i}{2}, \quad A_{11} = \frac{d_1}{2}, \quad A_{nn} = \frac{d_n}{2}. \tag{2.45}$$

This effect can be interpreted as uncorrelated noise and numerical evaluations (Lindgren and Rue, 2008) have shown that this approximation does not change the solution. Together,this gives a two-banded diagonal precision matrix $\mathbf{Q}$

$$Q_{i,i-2} = \frac{2}{d_{i-2} d_{i-1} (d_{i-2} + d_{i-1})} \tag{2.46}$$

$$Q_{i,i-1} = -\frac{2}{d_{i-1}^2} \left(\frac{1}{d_{i-2}} + \frac{1}{d_i}\right) \tag{2.47}$$

$$Q_{i,i} = \frac{2}{d_{i-1}^2 (d_{i-2} + d_{i-1})} + \frac{2}{d_{i-1} d_i} \left(\frac{1}{d_{i-1}} + \frac{1}{d_i}\right) + \frac{2}{d_i^2 (d_i + d_{i+1})} \tag{2.48}$$

At the end points $d_{-1} = d_0 = d_n = d_{n+1} = \infty$ resulting in some alterations in the corners

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} \frac{2}{d_1^2 (d_1 + d_2)} & \frac{-2}{d_1^2 d_2} \\ \frac{-2}{d_1^2 d_2} & \frac{2}{d_1 d_2} \left(\frac{1}{d_1} + \frac{1}{d_2}\right) + \frac{2}{d_2^2 (d_2 + d_3)} \end{bmatrix}. \tag{2.49}$$

It can be verified that the matrix $\mathbf{Q}$ has rank $n-2$. In the special case of equally space points s_i, the $d_i = 1$ for all i and the $\mathbf{Q}$-matrix reduces to the precision matrix for the second-order random walk in (1.51).

2.4 Comparisons

We now apply the methods on two different problems, a synthetic, regular spaced example and a irregular fossil shell dataset. The analysis use the INLA package in `R` and we will place the data points x_i on a knots τ_i, due to the structure of the `generic0`-function in INLA. Between two data points we place one knot, giving twice as many knots as data points. When the data is regularly spaced, the penalty matrix for the P-splines and RW2-model will be the same, supply the same weights $\hat{\mathbf{w}}$. The estimated function $\hat{f}$ will however be different, since RW2 use linear basis functions to display the solution and P-splines use cubic B-splines. For irregular data P-splines can not be use, since the knots are place on data points and P-splines require uniform knots.

Figure 2.3 shows the synthetic data, which is regular spaced based on the curve $f(x) = \sin \pi x$ with normally distributed errors $\epsilon \sim \mathcal{N}(0, 0.1^2)$. We see that the three estimates, the red, yellow and blue curve, are very similar and the true function, the black line, is always within the credible bands. The blue curve is the O'Sullivan spline, the red curve the RW2 model and the yellow curve P-splines. All three methods give good approximations to the true function, the black curve. The prior on the smoothing parameter used in INLA is Gamma$(10, 10)$ and estimated smoothing is $\lambda = 0.337$ for O'Sullivan splines and $\lambda = 0.323$ for the RW2 model and P-splines.

For the irregular spaced data we use a data set, first use by Chaudhuri and Marron (1999) provided by Bralower et al. (1997). The data contains the ratios of strontium isotopes and age, dated by biostratigraphic methods, from 106 samples of fossil shells. The two estimated curves are very similar and the credible bands narrows and widen depending on the density of data. For around 95 to 98 million year of age there is no data, making the curves deviate slightly from each other due to a small difference in the smoothing parameter. For the O'Sullivan penalty the smoothing becomes 28.15, while the RW2 model use 29.12. The prior used in both cases is Gamma $(z1, 10^{-4})$. For both the regular and irregular data, the estimates seem to give good results and the credible bands quantify the uncertainty in a good way.

Figure 2.5 show a sine function $\sin(\pi x), x \in [0, 2]$ with missing data between 0.8 and 1.2. The blue curve is the O'Sullivan spline, the red curve the RW2 model and the yellow curve P-splines. The true function is the black curve. We see that the uncertainty increase, where there is no data, but because of the smoothing, the credible bands give a good impression of the underlying function.

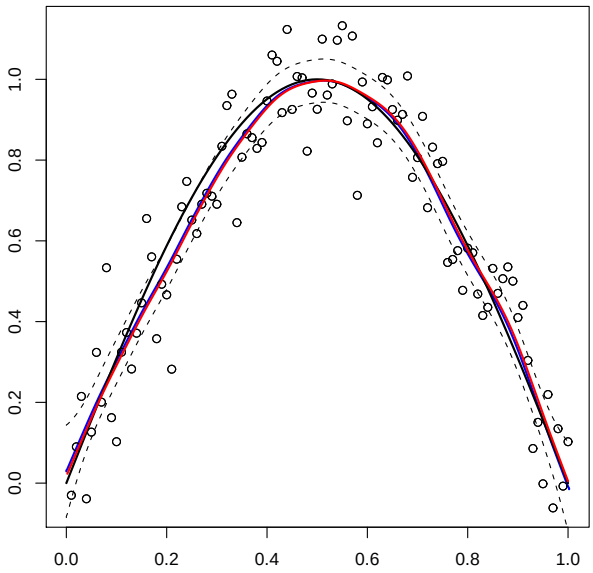


Figure 2.3: Regular spaced synthetic data, where the black curve is the true function $\sin(\pi x)$, the blue curve is the O'Sullivan spline, the red curve the RW2 model and the yellow curve P-splines.The prior on λ is Gamma$(10, 10)$ and the estimated values are $\lambda_O = 0.337$ and $\lambda_{P,RW2} = 0.323$. The dashed lines are 95% credible bands.

2.5 Conclusion

There are several advantages with the RW2-model over the ordinary spline methods previously described. The most important is the formulation in terms of a Bayesian field adding great flexibility and extendability, in terms of other latent fields and increased dimensions. One possibility is adaptive smoothing, where the smoothing is a variable $\lambda(x)$ incorporated in the field, which should better adapt to jumps and rapid curvature changes.

The Bayesian formulation and framework give a more natural explanation for the smoothing parameter. In this the case the smoothing is equivalent to the precision of Gaussian prior, a direct result of available information. In the frequentist view the the minimization of the penalty is somewhat unfounded, but in the Bayesian case it is a natural consequence of specified model. The a high level of smoothing means directly a precision for the prior, meaning the prior should be less effected by the observation. Very low smoothing, indicating low precision, makes the observation highly influential on the estimate. In terms of methodology, there is also a well established Bayesian framework for credible bands and prediction, which is very useful in quantifying uncertainty in the data and utilizing the

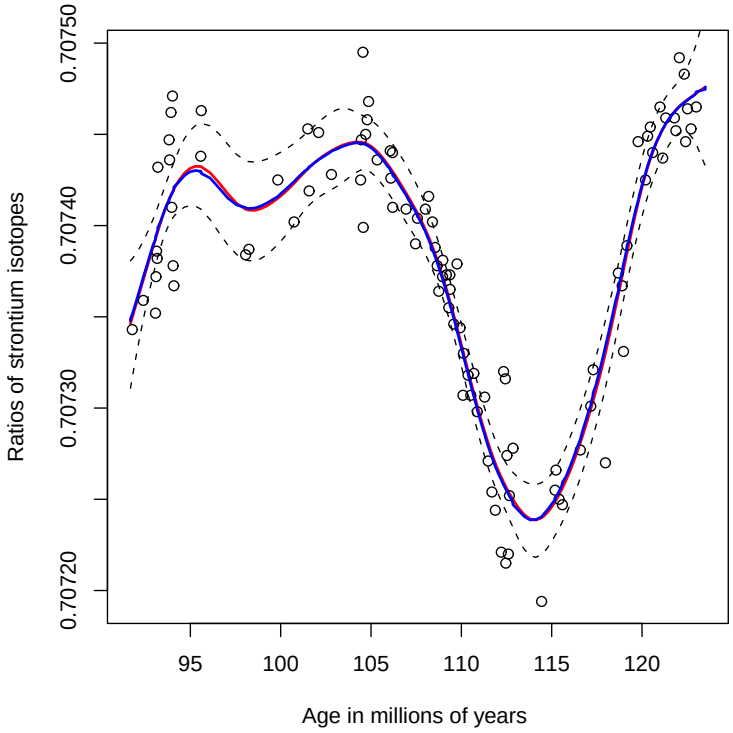


Figure 2.4: Data from 106 fossil shells. The red curve is estimated with a O'Sullivan penalty and the blue curve with a RW2 penalty. The dashed lines are 95% credible bands.

smoothing analysis.

Since we in addition specify a prior distribution for the hyperparameter λ, the analysis is less dependent of initial parameter choice. The prior allows the data in certain degree to seek out the proper smoothing. The specifications needed on the prior is enough precision to avoid instabilities in the numerical methods and at the same time have variation for the optimal smoothing level to be found.

In this section we have presented two mainstream smoothing methods and reformulated them as a Bayesian hierarchical model. In comparison we presented an approach based on a stochastic differential equation prior as given by (Wahba, 1978). Both the mainstream and SDE methods are evaluated within the INLA framework giving easy access to credible bands and estimates. With the Gaussian hierarchical model an exact solution can be found by a numerical optimization routine. Graphical evaluation showed that the methods are very similar and yield the same results, even though the theoretical motivations are different. This leads to a discussion of practicality against theoretical foundation. There are different advantages and problems, which will be discussed later.

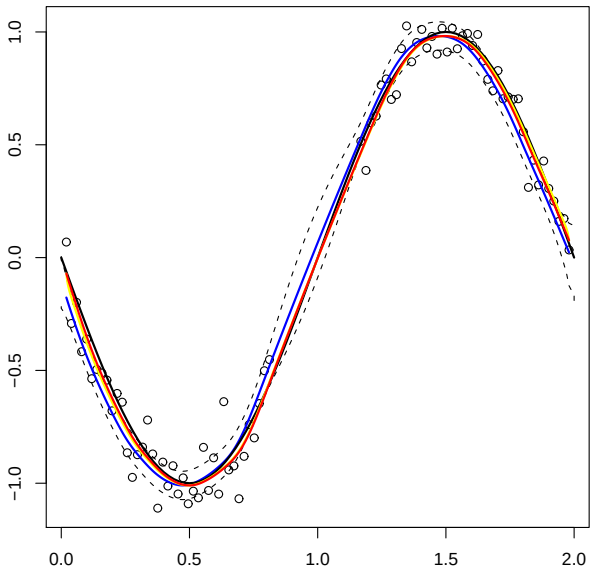


Figure 2.5: Sine functions $\sin \pi x$ on $[0, 2]$, with missing data between 0.8 and 1.2. The smoothing prior is Gamma$(\alpha, \beta)$with $\alpha = 10^1, \beta = 10^2$. The blue curve is the O'Sullivan spline, the red curve the RW2 model and the yellow curve P-splines.

Chapter 3

O'Sullivan splines and Galerkin methods

In Chapter 2, we have presented different methods for approximating exact smoothing splines. One method involved Galerkin approximations to solve a stochastic differential equation, while another used the smoothing penalty introduced by O'Sullivan (1986). Wand and Ormerod (2008) noted that the O'Sullivan spline estimate behaves almost identically to the exact smoothing splines solution found by solving (1.1). This is in stark contrast to the P-splines, which behave incorrectly at the boundaries. The comparison in Wand (2000) found that in all cases O'Sullivan splines were closer to smoothing splines than P-splines. Especially, the comparison showed O'Sullivan splines to closely mimic the natural boundary behavior of smoothing splines.

The question becomes: Why do O'Sullivan splines behave like smoothing splines? In this chapter, we will investigate this correspondence in light of the SDE model, which converges to exact smoothing splines as the number of basis functions increases. Our aim is to bridge the gap between O'Sullivan splines and the SDE model, and there is, in fact, a close connection between these two approaches.

3.1 Galerkin method

Finite element Galerkin methods are the most widely use approximation method for ordinary and partial differential equations. The key idea behind the Galerkin method is to solve the PDE exactly for a finite-dimensional space, replacing the infinite solution and test space. This stands in contrast to finite difference methods, which approximates the equation on a set of discrete locations. The Galerkin method is named after the Soviet mathematician and engineer Boris Galerkin,

who published it in 1915. He also made several contributions within industrial construction and helped build dams and hydroelectric plants in the Soviet Union.

The Bayesian formulation of smoothing splines uses the field prior specified by the stochastic differential equation on the domain $[a, b]$

$$\begin{cases} f''(x) &= W'(x), \\ f'(a) &= f'(b) = 0, \end{cases} \tag{3.1}$$

where $W(x)$ is the Wiener process. This equation is the same as a one-dimensional Poisson equation with a stochastic source term.

To find a solution, we start by looking at all possible test functions $\psi(x)$, which will span the test space V. A suitable test space could be all piecewise linear functions on an interval. Now we multiply both sides of equation (3.1) with the test functions and integrate, giving

$$\int f''(x)\psi(x)dx = \int W'(x)\psi(x)dx, \quad \forall \psi \in V. \tag{3.2}$$

We find the solution of $f(x)$, which satisfies this identity and this is called the *weak* solution. This solves (3.1) only with respect to a test function and it is therefore not required to hold absolutely.The weak formulation is the same as equating the inner products, $\langle f, g\rangle = \int f(x)g(x)dx$, of the right and left-hand side with a test function

$$\langle f''(x), \psi(x)\rangle = \langle W'(x), \psi(x)\rangle \,. \tag{3.3}$$

Unfortunately, we cannot test (3.2) for infinitely many set of test functions, so, in order to construct a practical solution method, we restrict ourself to a finite dimensional test space $V_h \subset V$. The usual choice of basis is the piecewise linear functions (2.39) given by a location sequence $s_1, s_2, \ldots, s_n$. The h refers to the mesh size, which the basis functions depend on. For $\psi_i(x)$ the h_i will be the difference between two s_i, $h_i = s_i - s_{i-1}$. This is a measure of how the finite number basis functions covers the original space. It is important that V_h approximates V well and when the mesh size decreases, the span of the approximated space should increases. The limit, $h \to 0$, should be that V_h spans the whole of V, satisfying the approximability property.

The collection of all possible function $f(x)$, we call the solution space W. But to be able to find an approximation to $f(x)$, we must choose a finite number of basis function for the solution space as well,

$$\tilde{f}(x) = \sum_{j=1}^{n} w_j \phi_j(x). \tag{3.4}$$

This reduces (3.2) to finding the weight vector, $\mathbf{w} = [w_1, \ldots, w_n]^T$. In this way, we search for functions in the finite-dimensional space W_h.

If we choose the test functions to be equal to the solution basis functions,

$$\phi_i(x) = \psi_i(x), \quad i = 1, \ldots, n, \tag{3.5}$$

we obtain the standard Galerkin method. Here the approximated spaces V_h and W_h are the same. If, on the other hand, we choose different bases for the two spaces, we obtain a Petrov-Galerkin method. Ern and Guermond (2004) show further definitions and conditions for all the important mathematical properties like conformity, consistency, orthogonality and well-posedness. They also comment that the approximability property may seem unnecessary to verify and practitioners generally do not bother to check it, but there are a number of situations, where seemingly sensible methods fail.

With the standard Galerkin method, $\phi_i = \psi_i$, the approximated function becomes

$$\tilde{f}(x) = \sum_{j=1}^{n} w_j \psi_j(x), \tag{3.6}$$

meaning the second derivative does not exist, since the basis functions are piecewise linear functions. To get around this, we modify the left-hand side using integration-by-parts, the Neumann boundary conditions $\tilde{f}'(a) = \tilde{f}'(b) = 0$ and the symmetry, $W'(x) = -W'(x)$ of the Wiener process, obtaining

$$\int_{[a,b]} \tilde{f}'(x)\psi_i'(x)dx = \int_{[a,b]} W'(x)\psi_i dx \quad i = 1, \ldots, n. \tag{3.7}$$

Since this is a stochastic differential equation, it is the distribution of both sides that must be equal for each collection of test functions $\{\psi_i(x)\}_{i=1,\ldots,n}$. We start with the left-hand side of 3.7 by inserting, $\tilde{f}'(x) = \sum_j w_j \psi_j'(x)$,

$$\left[\int \sum_j w_j \psi_j'(x)\psi_i'(x)dx\right]_{i=1,\ldots,n} = \left[\sum_j w_j \int \psi_j'(x)\psi_i'(x)dx\right]_{i=1,\ldots,n} = \mathbf{H}\mathbf{w}, \tag{3.8}$$

giving the elements $\mathbf{H}_{ij} = \int \psi_j'(x)\psi_i'(x)dx$. The right-hand side of 3.7 is Gaussian distributed with zero-mean and covariance matrix B with

$$\mathbf{B}_{ij} = \int \psi_i(x)\psi_j(x)dx. \tag{3.9}$$

In (Lindgren and Rue, 2008) it is shown that $\int_{[a,b]} \tilde{f}'(x)\psi_i'(x)dx$ has the same distribution as $\int_{[a,b]} f(x)\psi_i(x)dx$ for a Gaussian distributed $\mathbf{w}$ with zero-mean and dense precision matrix $\mathbf{Q} = \mathbf{H}^T\mathbf{B}^{-1}\mathbf{H}$.

3.2 Petrov-Galerkin approach

As mentioned earlier, we can also choose different bases for the test and solution space and this is called a Petrov-Galerkin method. We start with same approximation

$$\tilde{f}(x) = \sum_{j=1}^{n} w_j \phi_j(x), \tag{3.10}$$

and weak formulation of the SDE

$$\int f''(x)\psi(x)dx = \int W'(x)\psi(x)dx, \quad \forall \phi_i \in V. \tag{3.11}$$

Following the general method, we obtain the elements of $\mathbf{H}$ and $\mathbf{B}$

$$\mathbf{H}_{ij} = \int \phi_i''(x)\psi_j(x)dx, \quad \mathbf{B}_{ij} = \int \psi_i(x)\psi_j(x)dx. \tag{3.12}$$

The idea is to make a clever choice for $\phi_i(x)$ and $\psi_i(x)$. Firstly, the solution basis functions can be chosen to be cubic B-splines $\phi_j(x) = B_j(x)$ as in the earlier cases

$$\tilde{f}(x) = \sum_{j=1}^{n} w_j B_j(x) \tag{3.13}$$

The Galerkin basis functions (2.39) are $B_{j,1}(x)$, B-splines basis functions with degree one. If we make the clever choice of test functions to be the second derivative of cubic B-splines,

$$\psi_i(x) = B_i''(x), \tag{3.14}$$

which are piecewise linear, we achieve the following

$$\mathbf{H}_{ij} = \int B_{i,d}''(x)B_{j,d}''(x)dx, \quad \mathbf{B}_{ij} = \int B_{i,d}''(x)B_{j,d}''(x)dx. \tag{3.15}$$

The matrices are equal! This reduces the precision matrix

$$\mathbf{Q} = \mathbf{H}^T\mathbf{B}^{-1}\mathbf{H} = \mathbf{H}, \tag{3.16}$$

since $\mathbf{H}$ also is symmetric, and gives

$$\mathbf{Q}_{ij} = \int B_i''(x)B_j''(x)dx = \mathbf{\Omega}_{ij}, \tag{3.17}$$

which is exactly the precision matrix obtain with the O'Sullivan penalty approach. This means that O'Sullivan spline smoothing can be interpreted as a Petrov-Galerkin method with solution space basis function, $\phi_i(x) = B_i(x)$, and the second derivative as test functions, $\psi(x)_i = B_i''(x)$. The solution space consists of quadratic B-splines and the test space is spanned by piecewise linear functions, as for a standard Galerkin method, which makes it possible to interpret O'Sullivan splines in terms of smoothing splines. Furthermore, the methods used in (Lindgren et al., 2011) can be used to show O'Sullivan splines converge to the true smoothing spline solution as the number of basis functions approach infinity.

3.3 Conclusion

Smoothing splines has a Bayesian formulation in terms of a stochastic differential equation and a finite element approximation of the SDE model will converge to the exact solution. We have, in this chapter, shown that a specific Petrov-Galerkin method leads to a O'Sullivan penalty and this realization bridges the gap between O'Sullivan splines and smoothing splines. O'Sullivan splines are a non-standard finite element approximation to smoothing splines. The Petrov-Galerkin method is constructed with a solution space spanned by cubic B-splines and test functions as the second derivative of these B-splines.

This makes a very good argument for the O'Sullivan spline and establishes a proper mathematical interpretation for the method. The O'Sullivan solution will converge to smoothing splines, when the number of basis functions increases. This explains the similar behavior of O'Sullivan splines and smoothing splines, noted by Wand and Ormerod (2008). The shared properties, as the natural boundary condition, has a clear explanation.

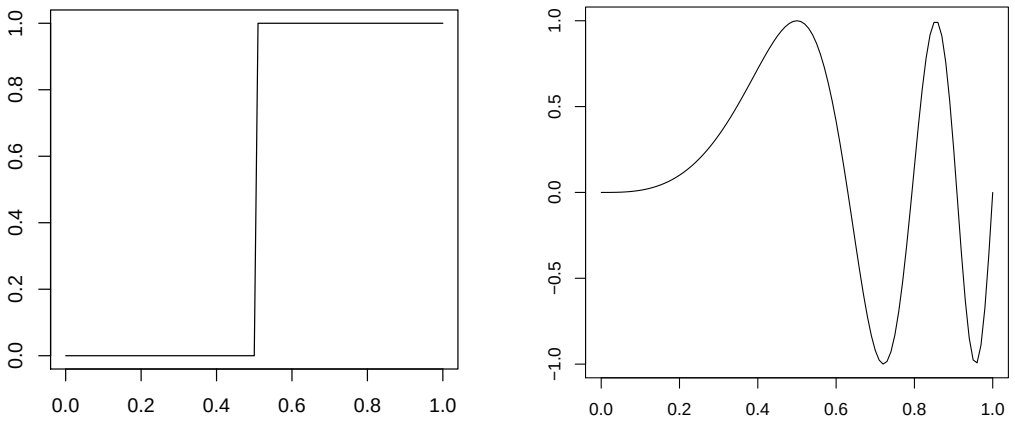
When the O'Sullivan splines can be seen as finite element solution of smoothing splines, the interpretation of the penalty in (1.1), raises some questions. The penalty is given by the integrated square second derivative, but for the standard Galerkin methods the solution is given by linear functions with no second derivatives. The frequentist penalty, therefore, can not be evaluated for the Bayesian RW2 model, but for the Petrov-Galerkin solution, the second derivatives exist, giving the penalty. This suggests the O'Sullivan splines are well motivated from a theoretical point of view. But why is this not done as the standard method, when considering finite element methods? The answer lies in what happens in higher dimensions. When the finite element approach is extended to two dimensions, the task of finding proper basis functions becomes difficult. If the basis functions were to have higher-order derivatives, the degrees of freedom would just be too large, yielding massive calculations with quite dense matrices (Brenner and Scott, 2008). The only feasible approach in two or more dimensions is to use piecewise linear tent functions, which, in the frequentist case, were explored by Roberts and Hegland (2004).

The main result is, however, the strong relationship between smoothing splines and the O'Sullivan spline, through the Petrov-Galerkin approximation. The O'Sullivan solution will converges to the exact smoothing spline as the number of basis functions approaches infinity. Wand and Ormerod (2008) calls O'Sullivan splines a direct generalization of smoothing splines, but it is better understood as a optimal approximation, in the sense that it is the best approximation over a mathematically consistent subspace.

Chapter 4

Adaptive smoothing

A natural extension of spline smoothing is a varying smoothing parameter $\lambda(x)$. The motivations lies in that ordinary methods can perform badly, when trying to estimate functions with jumps, peaks or quickly changing curvatures. We will look at two difficult examples of this sort, a simple step function and a sine function with increasing frequency. Adaptive smoothing can been done in many different ways.



(a) (b)

Figure 4.1: Examples of difficult functions, a) Step function, $I_{\{x>0.5\}}(x)$ b) Increasing frequency sine, $\sin(4\pi x^3)$.

For instance, Abramovich and Steinberg (1996) introduced a reproducing kernel Hilbert space representation following Wahba's approach, with a variable smoothing based on a roughness function $\psi(x)^2$ proportional to the derivative. Pintore et al. (2006) proposed a piecewise constant model for the smoothing. Ruppert and Carroll (2000) used the P-splines of Eilers and Marx (1996) with varying smoothing on the difference penalties of the weights. Other methods are based on wavelet

shrinkage, multiple hyperparameter interpolation (Mackay and Takeuchi, 1998) or regression splines with adaptive knot selection and non-stationary Gaussian process regression models (Paciorek and Schervish, 2006).

4.1 Method

Our approach is to specify an extended Bayesian hierarchical model and let the smoothing parameter be incorporated into the latent field, in the following way

$$\text{Observations:} \qquad y_i = f(x_i) + \epsilon_i, \quad x \in [0,1] \tag{4.1}$$

$$\text{Latent field:} \qquad \frac{d^2}{dx^2}\left(b(x)^{\frac{1}{2}} f(x)\right) = W'(x), \quad \lambda(x) > . \tag{4.2}$$

The prior on $b(x)$ is either a secondary field or a parametric form with hyperparameters. The function $b(x)$ can be seen as a instantaneous variance or local scaling, which compress and stretch the function. A small $b(x)$ compresses the scale giving quick oscillations, while a high value stretch $f(x)$, decreasing the roughness. The smoothing parameter must be positive to make sense, so the natural reformulation is $b(x) = e^{2\nu(x)}$, where $\nu(x) \in \mathbb{R}$. This connects $b(x)$ directly to the precision b from (1.13), which derive from

$$\int \left[b^{\frac{1}{2}} f''(x)\right]^2 dx. \tag{4.3}$$

The smoothing is given $\lambda(x) = \sigma^2 b(x)$. This is further extendable to the field

$$\frac{d}{dx}\left(a(x)\frac{d}{dx}\left(b(x)f(x)\right)\right) = W'(x), \quad a(x), b(x) > 0, \tag{4.4}$$

where $a(x)$ and $b(x)$ are function that stabilize the variance and range and both (4.4) and (4.2) are non-stationary models.

We explore the model with $a(x) = 1$ and $b(x) = e^{2\nu(x)}$, such that the function space of possible $f(x)$ has a distribution specified by the field $\frac{d^2}{dx^2}\left(e^{\nu(x)} f(x)\right) = W'(x)$. The smoothing function $\nu(t)$ is represented as a weighted sum of basis functions, $\xi_i(x)$, giving the following general hierarchical model:

$$\text{1.) Observations:} \qquad y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{4.5}$$

$$\text{2.) Prior field:} \qquad \frac{d^2}{dx^2}\left(e^{\nu(x)} f(x)\right) = W'(x), \quad \lambda > 0, \tag{4.6}$$

$$\nu(x, \boldsymbol{\beta}) = \sum_i \beta_i \xi_i(x) \tag{4.7}$$

$$\text{3.) Hyperparameter} \qquad \boldsymbol{\beta} \sim N(0, \Sigma_\beta). \tag{4.8}$$

To find the resulting precision matrix given by the field, we use the Galerkin approach. For all test functions ψ it must hold

$$\left[\left\langle \psi_i, \frac{d^2}{dx^2}\left(e^{\nu(x)} f(x)\right)\right\rangle\right]_{i=1,\dots,n} = \left[\langle \psi_i, W'(x)\rangle\right]_{i=1,\dots,n}. \tag{4.9}$$

Then, for the left-hand side, the Galerkin approximation is given

$$\left[\left\langle \psi_i, \frac{d^2}{dx^2}\left(e^{\nu(x)} \tilde{f}(x)\right)\right\rangle\right]_{i=1,\dots,n} = \left[\sum_j w_j \left\langle \psi_i, \frac{d^2}{dx^2}\left(e^{\nu(x)}\psi_j\right)\right\rangle\right]_{i=1,\dots,n} = \mathbf{H}\mathbf{w}. \tag{4.10}$$

Each element in the matrix $\mathbf{H}$ is given by the integral over two basis function,

$$\mathbf{H}_{ij} = \left\langle \psi_i(x), \frac{d^2}{dx^2}\left(e^{\nu(x)}\psi_j(x)\right)\right\rangle = -\left\langle \frac{d\psi_i}{dx}, \frac{d}{dx}\left(e^{\nu(x)}\psi_j(x)\right)\right\rangle, \tag{4.11}$$

which is simplified with integration-by-parts, giving

$$\mathbf{H}_{ij} = -\int_\Omega \frac{d\psi_i}{dx}\frac{d}{dx}(b(x)\psi_j(x))\,dx, \quad b(x) = e^{\nu(x)}, \tag{4.12}$$

where

$$\frac{d\psi_i}{dx} = \begin{cases} \frac{1}{d_{i-1}}, & s_{i-1} \le x < s_i \\ -\frac{1}{d_i}, & s_i \le x < s_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

For $d_i = s_{i+1} - s_i$, the tri-banded matrix $\mathbf{H}$ is given and

$$\mathbf{H}_{i,i} = -b(s_i)\left(\frac{1}{d_{i-1}} + \frac{1}{d_i}\right) \tag{4.14}$$

$$\mathbf{H}_{i,i+1} = b(s_{i+1})\frac{1}{d_i} \tag{4.15}$$

$$\mathbf{H}_{i,i-1} = b(s_{i-1})\frac{1}{d_{i-1}}, \tag{4.16}$$

since ψ_i only overlap for neighboring basis functions. This is very similar to the RW2 model, the only difference being the function $b(x)$ depended on $\boldsymbol{\beta}$.

As stated by Lindgren and Rue (2008), the precision matrix for the specified field, with an approximated matrix $\mathbf{A}$ for the covariance of the white noise, is $\mathbf{Q}(\boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta})^T\mathbf{A}^{-1}\mathbf{H}(\boldsymbol{\beta})$, a five-banded symmetric sparse matrix. Then the hierarchical model becomes

$$\mathbf{y} \sim \mathcal{N}(\mathbf{B}\mathbf{w}, \sigma_\epsilon^2 I) \tag{4.17}$$

$$\mathbf{w} \sim \mathcal{N}\left(0, \mathbf{Q}(\boldsymbol{\beta})^{-1}\right), \tag{4.18}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \Sigma_\beta). \tag{4.19}$$

The Bayesian formulation is given

$$\pi(\mathbf{w}, \boldsymbol{\beta}|\mathbf{y}) \propto \pi(\boldsymbol{\beta}) \cdot \pi(\mathbf{w}|\boldsymbol{\beta}) \cdot \pi(\mathbf{y}|\mathbf{w}) \tag{4.20}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\mathbf{y}-\mathbf{B}\mathbf{w})^T(\mathbf{y}-\mathbf{B}\mathbf{w}) - \frac{1}{2}\mathbf{w}^T\mathbf{Q}(\beta)\mathbf{w} - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_\beta\boldsymbol{\beta}\right\}. \tag{4.21}$$

We find a solution of $\mathbf{w}, \boldsymbol{\beta}$ by a maximum *a posteriori* probability (MAP) estimate, which is the mode of the posterior distribution. The estimated is computed via a numerical optimization routine. This way we find the estimate $\hat{f}(x|\mathbf{w}) = \mathbf{B}\hat{\mathbf{w}}$, where $\hat{\mathbf{w}}$ is the minimizer of the following expression

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\beta}}) = \arg\min\left[\frac{1}{2\sigma_\epsilon^2}(\mathbf{y}-\mathbf{B}\mathbf{w})^T(\mathbf{y}-\mathbf{B}\mathbf{w}) + \frac{1}{2}\mathbf{w}^T\mathbf{Q}(\boldsymbol{\beta})\mathbf{w} + \frac{1}{2}\boldsymbol{\beta}^T\Sigma_\beta\boldsymbol{\beta}\right] \tag{4.22}$$

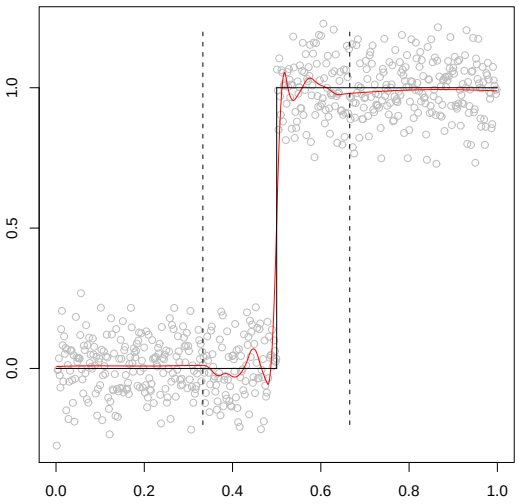


Figure 4.2: Step function with a three part piecewise smoothing parameter, $\beta_1 = 2.7 \cdot 10^{-2}, \beta_2 = 1.4 \cdot 10^{-7}, \beta_3 = 4.1 \cdot 10^{-2}$.

4.2 Results

With the general method described, the key step is the parametrization of the smoothing function $\nu(x)$. The simplest choice is a piecewise linear function, seen in Figure 4.2 with three different intervals as

$$\nu(x) = \begin{cases} \beta_1, & x \in [0, 1/3) \\ \beta_2, & x \in [1/3, 2/3) \\ \beta_3, & x \in [2/3, 1). \end{cases} \tag{4.23}$$

The smoothing parameter is given by $\lambda(x) = \sigma_\epsilon^2 e^{2\nu(x)}$ with three different smoothing levels.

In Figure 4.2, we can clearly see where the smoothing parameter changes, which is an undesired property. Nevertheless, the jump is very nicely estimated and we achieve two straight lines in the outer regions, which corresponds exactly to the step function. The estimate can be improved by increasing the number of intervals, as done in (Pintore et al., 2006) or have intervals of unequal size.

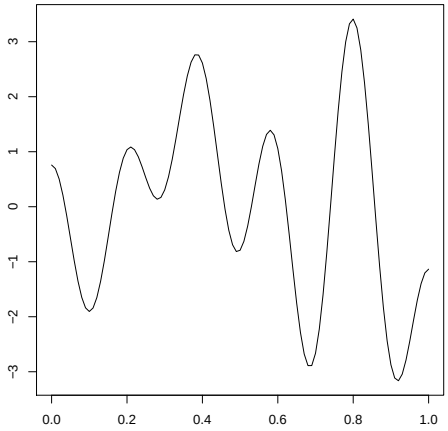


Figure 4.3: Example of Fourier representation, $\nu(x)$, with $l = 10$, $\boldsymbol{\beta} \sim \mathcal{N}(0, I)$.

Fourier parametrization

To achieve a smooth function $\nu(x)$, we can use smooth basis functions instead of discontinuous piecewise constants. One possibility is to parametrize along a Fourier series expansion, using a cosine basis $b_i(x) = \cos \pi i x$ and random amplitudes, given as

$$\nu(x) = \sum_{i=1}^{l} \beta_i \cos(\pi i x), \quad \boldsymbol{\beta} \sim \mathcal{N}(0, \Sigma), \tag{4.24}$$

where Σ is a covariance matrix. Figure 4.3 shows a realization of (4.24) with β_i being independent normally distributed, $\boldsymbol{\beta} \sim \mathcal{N}(0, I)$. With the Fourier representation the smoothing function becomes smooth, making gradually adjustments, unreasonable jumps. By choosing, $l = 20$, we allow for high frequency cosines,which makes it possible to achieve quite rapid changes. Figure 4.4 shows a step function with the spline estimate $\hat{f}(x)$ as a red curve. The fit to the true function, the black curve is particularly good. The smoothing $\nu(x)$ is fairly symmetric and high at the boundaries and drops substantially at the jump.

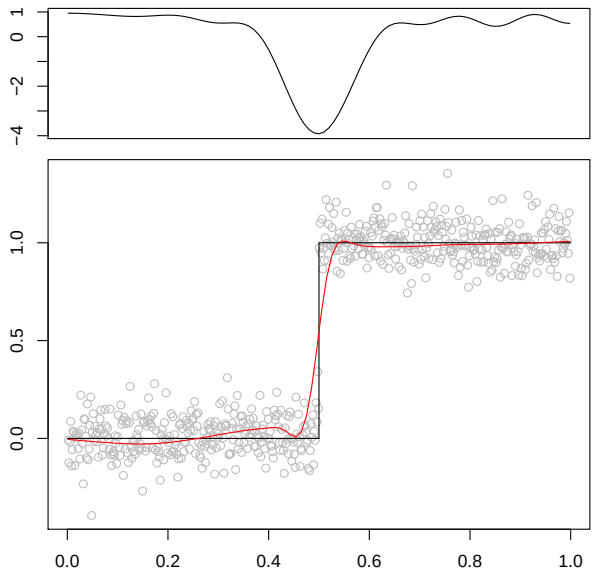


Figure 4.4: An estimated step function based on 1000 data points and 60 Galerkin basis functions. Smoothing function $\nu(x)$ in the top tile has a Fourier representation with 20 basis functions and the amplitude prior is $\boldsymbol{\beta} \sim \mathcal{N}(0, 10^2 I)$

Field prior on $\boldsymbol{\beta}$

In Figure 4.4, $\boldsymbol{\beta}$ is *iid* distributed $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 I)$, which does not have a any practical interpretation. It is instead possible to have a prior distribution modeled by an underlying field, adding a new level to the hierarchical model. We look at two situations with an O'Sullivan-type penalty, mimicking the behavior of a first and second order random walk. The $\boldsymbol{\beta}$s are the weights of the cosine basis functions $b_i(x) = \cos \pi i x$ and the penalties are given

$$\Omega_{1,ij} = \int_\Omega b_i'(x) b_j'(x) dx, \tag{4.25}$$

$$\Omega_{2,ij} = \int_\Omega b_i''(x) b_j''(x) dx. \tag{4.26}$$

Since both $\sin(\pi i x)$ and $\cos(\pi i x)$ build, in this case, orthogonal systems, the penalty matrices Ω_1 and Ω_2 will be diagonal with the following elements

$$\Omega_{1,ii} = \int_0^1 ((\cos \pi i x)')^2 dx = (\pi i)^2 \int_0^1 (\sin \pi i x)^2 dx = \frac{1}{2}(\pi i)^2 \tag{4.27}$$

$$\Omega_{2,ii} = \int_0^1 ((\cos \pi i x)'')^2 dx = (\pi i)^4 \int_0^1 (\cos \pi i x)^2 dx = \frac{1}{2}(\pi i)^4 \tag{4.28}$$

This gives a diagonal penalty matrix for both $\Omega_1 = \frac{\pi}{2} \operatorname{diag}\{1^2, 2^2, \ldots, l^2\}$ and $\Omega_2 = \frac{\pi}{2} \operatorname{diag}\{1^4, 2^4, \ldots, l^4\}$. These penalties are the resulting precision matrices

from the following two underlying fields

$$(1) \qquad \frac{d\nu(x)}{dx} = \frac{dW(x)}{dx}, \tag{4.29}$$

$$(2) \qquad \frac{d^2\nu(x)}{dx^2} = \frac{dW(x)}{dx}, \tag{4.30}$$

The diagonal precision matrix corresponds to independent $\boldsymbol{\beta}$s with increasing precision for increasing frequency. This means the amplitude $\boldsymbol{\beta}$ is suppressed for high frequency basis functions, since the mean is zero. The estimated function, $\nu(x)$, will therefore be smoother, than in the *iid* case. Since the precision increases with i^4 for Ω_2, the parameter function will be even smoother for this penalty.

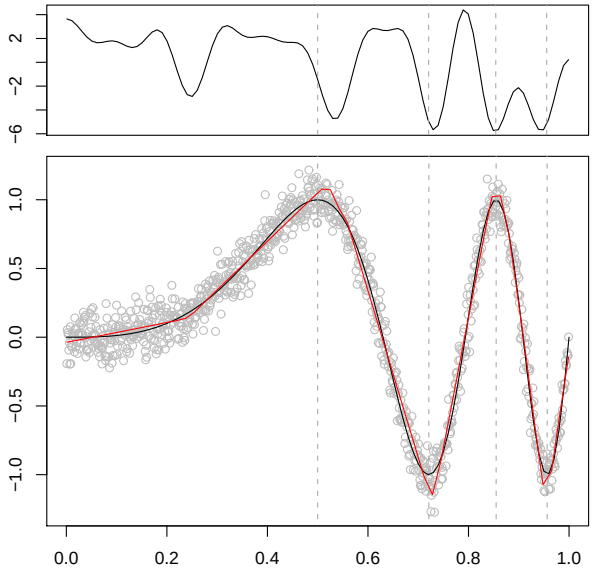


Figure 4.5: Sine function with prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 I)$ where $\sigma = 1$, based on 1000 data points and 60 Galerkin basis function.

In Figure 4.8 and 4.9, we see the step function with the introduced precision matrices Ω_1 and Ω_2. The result are slowly changing, smoother functions without unreasonable fluctuations. The smoothing function with a second-order penalty changes slowly and looses the wanted flexibility, when estimating the discontinuity.

The other important example is the sine function $\sin(4\pi x^3)$ with increasing frequency. There are no discontinuities, but the second derivative changes rapidly and there are several inflection points. Figure 4.5, 4.6 and 4.7 display the sine function with different penalty matrices $\Sigma^{-1} = \Omega$. In Figure 4.5, we use identically distributed βs, $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2 I)$ where $\sigma = 0.3$. The result is a rapidly changing

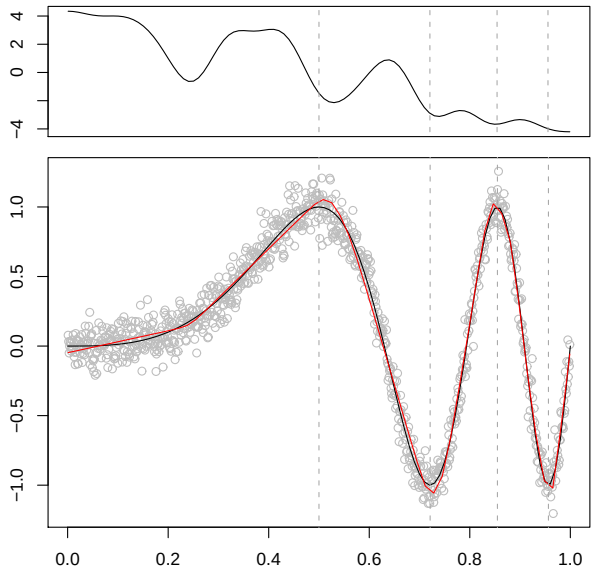


Figure 4.6: Sine function with prior penalty $\Omega_1 = \frac{\pi}{2}\,\mathrm{diag}\{1^2, 2^2, \ldots, l^2\}$ with $l = 20$ giving $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 \Omega_1^{-1})$ where $\sigma_\beta = 1.5$. The estimate is based on 1000 data points and 60 Galerkin basis function.

function, where 15 cosine basis functions allow the rapid changes. The smoothing is fairly high close to zero with a quick dip at the first inflection point, which change the direction of the estimated straight line. Then the smoothing changes from a high value 10^3 at the inflection points to 10^{-6} at the minima and maxima. This gives a very flexible fit, alternating between data interpolation and straight lines.

The next example uses a penalty matrix corresponding to a continuous random walk, as (4.29). With this specific precision matrix, the high frequency basis functions of $\nu(x)$ will be suppressed, giving less rapid changes. This is clear in Figure 4.6. The smoothing follows the same pattern as earlier with low smoothing at stationary points and higher smoothing at inflection points. The difference between high and low smoothing is smaller, due to the smoother function.

In Figure 4.7 we use the penalty matrix based on a continuous second order random walk. The general pattern is the same, but the changes are much slower. After the first maximum the smoothing stabilizes at a low level and does not exhibit any of the fluctuations seen in Figure 4.5. The overall fit to the underlying function is however fairly good.

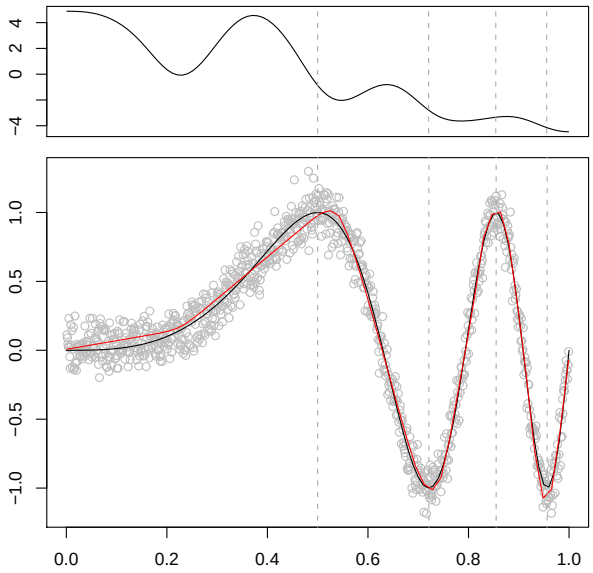


Figure 4.7: Sine function with prior based on penalty $\Omega_2 = \frac{\pi}{2}\,\mathrm{diag}\{1^4, 2^4, \ldots, l^4\}$ with $l = 20$ giving $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 \Omega_2^{-1})$ where $\sigma_\beta = 3$. The estimate is based on 1000 data points and 60 Galerkin basis functions.

4.3 Summary

Introducing an adaptive smoothing parameter is an easy task with the underlying field framework used in (Lindgren and Rue, 2008). We utilize the presented method and integrated the parameter as a function in the field. The important model adaption is the parametrization of the smoothing function $\nu(x)$. We have seen examples of a piecewise constant and a Fourier cosine basis formulation, where the Fourier expansion type functions yielded the best results. It seems that a smooth function is the best solution. Apart from choosing a good formulation for the function, the proper prior for the hyperparameters are an important point of discussion. We have seen the use of three different precision matrices for the Gaussian prior of $\boldsymbol{\beta}$ giving different behavior for the resulting function. A precision matrix based on an O'Sullivan type penalty of the first derivate gave good results. This eliminated unreasonable fluctuations, but kept the ability to have rapid changes, giving a flexible overall fit.

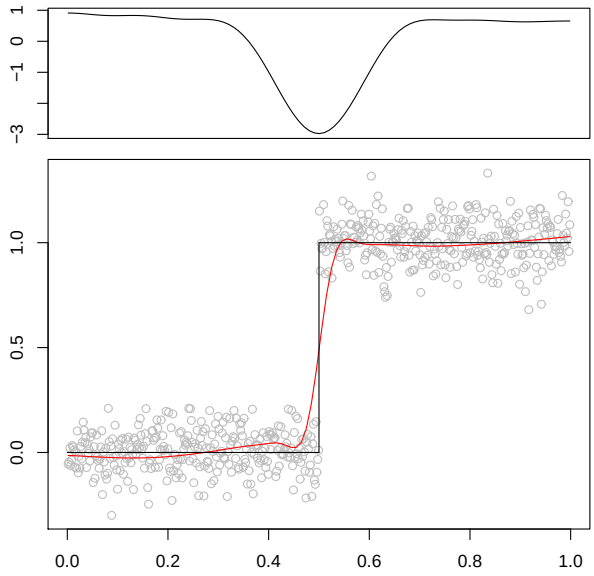


Figure 4.8: An estimated step function based on 600 data points and 100 Galerkin basis functions. The smoothing function use a Fourier representation with a first-order penalty on the amplitudes, $\beta \sim \mathcal{N}(0, \Omega_1^{-1})$.

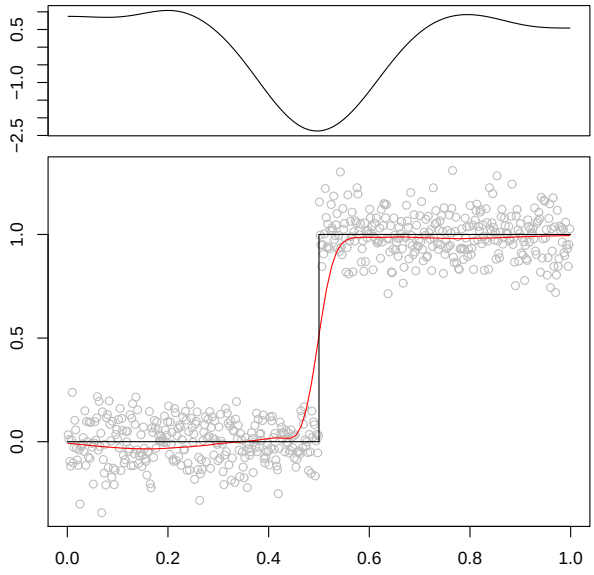


Figure 4.9: An estimated step function based on 600 data points and 100 Galerkin basis functions. The smoothing function use a Fourier representation with a second-order penalty on the amplitudes, $\beta \sim \mathcal{N}(0, 10^2\Omega_2^{-1})$.

Chapter 5

Conclusion

In this thesis, we have presented three different approaches to spline smoothing, two frequentist methods involving a penalty approximation and one Bayesian method approximating a latent field in form of a stochastic differential equation. P-splines and O'Sullivan splines approximate the square integrated second derivative by different penalty matrices, $\mathbf{\Omega}$, as a quadratic form, $\mathbf{w}^T\mathbf{\Omega}\mathbf{w}$. P-splines use a higher-order difference penalty on the spline coefficients, while the O'Sullivan splines base the penalty on the integrated second derivative of the basis functions.

P-splines can only work with uniform basis functions, a quite restrictive requirement, and the penalty itself has no natural interpretation. For cubic splines with a second-order difference, the penalty can be interpreted as an approximation to the integrated second derivative, where the effect from neighboring basis functions is left out, as seen in (2.21). On the other hand, the method allows for higher-order penalties, which result, when over-smoothing, in quadratic and cubic fits.

The penalty used in O'Sullivan splines has a very natural interpretation, along the lines of a finite element approach. The function, $f(x)$, is approximated in terms of a weighted set of basis functions, which is inserted in the penalty integral to obtain the penalty. This allows for non-uniform basis functions, usually quantiles-based knots, which better utilize calculation power. The basis functions will be placed, where there is high density of data, making it possible to capture more fine-scaled structures.

The RW2 method is a Bayesian approach to smoothing splines. The prior for the Bayesian hierarchical model is a latent field specified as a stochastic differential equation. An approximate solution of the SDE is found with a standard finite element Galerkin method. This supplies a precision matrix describing the statistical properties of the latent field. Each sample path of the stochastic process represents a possible prior function. The basis functions used in the Galerkin approximation are specified by a possibly, irregular location and observation sequence. The most

important feature of the RW2 model is the continuous interpretation, as an exact solution to the Bayesian formulation of smoothing splines. When we increase the number of basis functions, we get closer to the exact solution. For P-splines, this is not true, since the method eliminate the effect of overlapping, neighboring basis functions.

One important aspect of regression, is the degree of uncertainty. The random nature of the data, makes quantification of the uncertainty crucial to properly assessing the estimates. However, for frequentist minimization formulation of smoothing splines, there are no straightforward way to achieve this. The Bayesian formulation, however, can be very useful in this respect.

If a full Bayesian analysis is carried out, we obtain the posterior densities and can calculate the 95% credible bands for the spline weights. This will quantify how certain we are in our estimates in a proper way. Closely observed data decrease the uncertainty between the observations and, from (2.48), we see they give high values for the precision matrix, resulting in narrow posterior densities and credible bands. This highlights the usefulness of INLA, which performs a complete Bayesian analysis and supply the wanted posterior densities and credible bands. INLA utilizes integrated nested Laplace approximation to preform an approximate Bayesian analysis and is based on a general, Gaussian hiearchical model, making it easy to specify to the smoothing methods.

From graphical results, we have seen that all the methods give fairly similar estimates both for regular and irregular data. The pointwise credible bands for each spline coefficient are plotted to display the uncertainty of the estimated function, even though a credible band for the function as a whole, would be slightly more restrictive.

A main focus has been, that the frequentist methods can be reformulated as Bayesian hierarchical models. This reformulation is very useful, both in terms of interpretation and flexibility. With a Bayesian model the important concept of the smoothing parameter can be understood as the precision of a prior distribution. The precision expresses the proportion of the continuous second-order random walk in the prior, as seen in (2.33). Since the smoothing parameter λ is proportional to the prior precision b, $\lambda = \sigma^2 b$, smoothing becomes directly connected to the strength of the prior. The smoothing must balance the observations against the assumed stochastic process and this interpretation is gain through the Bayesian formulation.

The introduction of a latent field gives great flexibility, for instance, in terms of an adaptive smoothing parameter. Adaptive smoothing will decrease in areas with rapid fluctuations, capturing more of the fine scale structure. When estimating a step functions, the smoothing would drop significantly at the jump to capture the real behavior of the function. In terms of the field, the smoothing can be seen

as an instantaneous variance, which stretch and compress the scale of the prior function. Two examples in form of a step function and an increasing frequency sine have be explored and we got good results in both cases. We also specified priors for the smoothing function, as latent fields given by Gaussian white noise, Wiener process and an integrated Wiener process, giving different degrees of smoothness. A first order penalty, seem to yield the best results.

The SDE approach is connected to both the P-splines and O'Sullivan splines. We have seen that the RW2 method will for uniform basis functions, have the same penalty matrix, as second-order P-splines, which means the methods will supply the same weights $\hat{\mathbf{w}}$. The estimated function $\hat{f}$ will, however, be different, since RW2 use linear basis functions to display the solution and P-splines use quadratic or cubic B-splines. The RW2 model has a clear continuous interpretation, which is not true for P-splines.

In Chapter 3, we saw that the construction of a Petrov-Galerkin method with different the test and solution space, bridged a gap between the O'Sullivan splines and the SDE method. When the solution space is spanned by cubic B-splines, we can make a clever choice for the construction of the test space. If the test functions are the second derivative of the B-splines spanning the solution space, the penalty matrix for the SDE approach becomes exactly that of the O'Sullivan splines. This makes the O'Sullivan penalty particularly well founded mathematically. It explains the good qualities of O'Sullivan splines and the similarity with original smoothing splines. Further, it underlines the strength of a SDE formulation, as there is great flexibility in choice of test and solution space. This flexibility is expressed, when the SDE approach and O'Sullivan splines coincide for a specific choice of basis functions.

With all the methods presented in this thesis, there are specific choices, which are difficult to evaluate. The analysis will always depend on the choice of prior distributions of hyperparameters and choice of basis functions and knots. The issue concerning proper choice of priors is difficult to assess. In connection with the spline smoothing in INLA, it is important to have a prior on the smoothing parameter, that is neither too specific or general. One possible approach is to find a proper smoothing value with a general cross-validation routine and construct the prior to have this value as expectation in combination with a reasonable variance. This issue is can be discussed in great lengths, but the pragmatic approach is in many ways to use a prior that works.

With the choice of knot sequence for the mainstream spline methods, several considerations should be made. The placement of knots is done uniformly for P-splines and based on data quantiles for O'Sullivan splines. Of these two approaches, the quantile-based seems most appropriate, when we consider the possible waste of computational power. It is unlikely that important fine-scaled structures are

uniformly distributed in your data, making a plea for strategic placement of basis functions. With P-splines, we will do unnecessary large computation, but the quantile-based approach can be improved by taking into account the changes in the response, as well. A good example is the step function with uniform data, where the placement based on quantiles will be uniform. Nevertheless, we want a higher density of basis functions around the jump, than by the edges. By basing the placement on the first derivative, this can be avoided. The derivative could be estimated directly from the data, as a finite difference, or by first finding an estimate for the function and use the derivative of this. Generally, the placement of basis functions, deserve a detailed consideration.

In terms of future work, latent fields can be extended to higher dimensions and include cyclic priors, oscillating, anisotropic and non-stationary fields. In Chapter 4, we explore one extension, a non-stationary field with the following model

$$\frac{d^2}{dx^2}\left(b(x)^{\frac{1}{2}}f(x)\right) = W'(x), \quad \lambda(x) > 0, \tag{5.1}$$

resulting in an adaptive smoothing approach with good results for functions with jumps and quick curvature changes.

Another extension as a non-stationary field introduces a variable coefficient $a(x)$, requiring a numerical integration and derivation scheme, in calculating the precision matrix elements. Due to the approximation error, it could, however, be sufficient with some sort of discretization at knot points. The coefficient $a(x)$ will together with $b(x)$ control the scale and range of the function, for which the analysis could benefit.

$$\frac{d}{dx}\left(a(x)\frac{d}{dx}\left(b(x)f(x)\right)\right) = W'(x), \quad a(x), b(x) > 0, \tag{5.2}$$

Another important work for the future, would be to incorporate adaptive smoothing in INLA, to get a full Bayesian analysis and the possibility of credible bands. In the Chapter 4, we only use a simple numerical optimization routine to find the estimates, which does not supply credible bands. With INLA, we can achieve the credible bands by incorporating another level in hierarchical model, being dependent on a second latent field. It would also be practical to have different parameterizations of $b(x)$, in terms of cosine, piecewise linear or other basis functions. A Bayesian analysis providing the posterior densities of the coefficients, makes it possible to calculate the pointwise credible bands for the estimated function.

In conclusion, the Bayesian formulation of smoothing splines give a substantial contribution in terms of interpretation and flexibility. The possibility of calculating posterior densities using INLA is the greatest advantage, because credible bands will quantify the uncertainty associated with the estimate. The flexibility is

expressed with a non-stationary field extension, allowing for adaptive smoothing, a better method for estimating jumps and quick curvature changes. In addition, it provides an explanation of the O'Sullivan penalty in terms of a finite element, Petrov-Galerkin approximation.

Bibliography

F. Abramovich and D.M. Steinberg. Improved inference in nonparametric regression using Lk-smoothing splines. *Journal of statistical planning and inference*, 49(3):327–341, 1996.

T.J. Bralower, P.D. Fullagar, C.K. Paull, G.S. Dwyer, and R.M. Leckie. Mid-Cretaceous strontium-isotope stratigraphy of deep-sea sections. *Geological Society of America Bulletin*, 109(11):1421, 1997.

S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Verlag, 2008.

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, (2):89–102, 1996.

P.H.C. Eilers and B.D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2005.

A. Ern and J.L. Guermond. *Theory and practice of finite elements*. Springer Verlag, 2004.

R.M. Jones. *Buckling of bars, plates, and shells*. Bull Ridge Corporation, 2006.

F. Lindgren and H. Rue. On the Second-Order Random Walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700, 2008.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B forthcoming*, 2011.

G. Lindgren. *A second course on stationary stochastic processes*. 2010.

D.J.C. Mackay and R. Takeuchi. Interpolation models with multiple hyperparameters. *Statistics and Computing*, 8(1):15–23, 1998.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.

C.J. Paciorek and M.J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006.

A. Pintore, P. Speckman, and C.C. Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113, 2006.

S. Roberts and M. Hegland. Approximation of a thin plate spline smoother using continuous piecewise polynomial functions. *SIAM journal on numerical analysis*, 41(1):208–234, 2004.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*, volume 104. Chapman & Hall, 2005.

H. Rue and S. Martino. Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, 2009.

D. Ruppert and R.J. Carroll. Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, 42(2):205–223, 2000.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*, volume 12. Cambridge Univ Pr, 2003.

I.J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America*, 52(4):947, 1964.

G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372, 1978.

G. Wahba. *Spline models for observational data*, volume 59. Society for Industrial Mathematics, 1990.

M.P. Wand. A comparison of regression spline smoothing procedures. *Computational Statistics*, 15(4):443–462, 2000.

M.P. Wand and J.T. Ormerod. On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198, 2008.