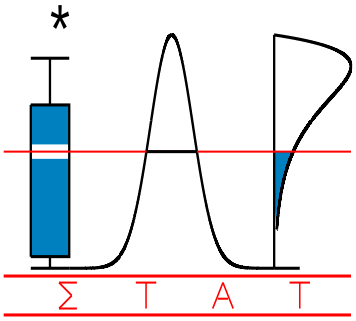
T E C H N I C A L
R E P O R T

0559

ROBUST SPECIFICATION OF THE ROUGHNESS PENALTY PRIOR DISTRIBUTION IN SPATIALLY ADAPTIVE BAYESIAN P-SPLINES MODELS

JULLION A., and P. LAMBERT



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models

Astrid Jullion[1] and Philippe Lambert[1,2*]
[1] *Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

[2] *Unité d'épidémiologie, biostatistique et méthodes opérationnelles, Faculté de Médecine, Université catholique de Louvain, Belgium.*

November 24, 2005

Abstract

We start the paper by pointing the potential important role of the prior distribution of the roughness penalty parameter in the resulting smoothness of Bayesian P-splines models (Ruppert *et al.* 2003 ; Lang and Brezger 2004). The recommended specification for that distribution yields models that can lack flexibility in specific circumstances. In such instances, these are shown to correspond to a frequentist P-splines model (Eilers & Marx, 1996) with a predefined and severe roughness penalty parameter, an obviously undesirable feature.

We show that the specification of a hyperprior distribution for one parameter of that prior distribution provides the desired flexibility. Alternatively, a mixture prior can also be used.

An extension of these two models by enabling adaptive penalties is provided. All the proposed models can be fitted quickly using the convenient Gibbs algorithm.

1 Introduction

Bayesian P-splines have recently become a widely used tool to describe the conditional mean of a response. Various authors have studied them either in normal (Ruppert *et al.* 2003 ; Berry *et al.* 2002) or non-normal contexts (Fahrmeir *et al.* 2004 ; Brezger and Lang 2005 ; Lambert and Eilers 2005 ; Lambert 2005). In the Bayesian P-splines model described in Lang and Brezger (2004), the prior distribution of the roughness penalty parameter τ_λ is taken to be a gamma with mean a/b and variance a/b^2 with a small value for b. What we highlight in this paper is the influence that the choice of b can have on the smoothness of the fitted curve. Indeed, we show that, in some specific circumstances, the results are highly sensitive to the value picked for b.

*Correspondence to: Philippe Lambert, Université catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve (Belgium). E-mail: lambert@stat.ucl.ac.be Phone: +32-10-47.28.01 Fax: +32-10-47.30.32

We propose two solutions that avoid the need of a sensitive choice for b. With the first solution, we consider the parameters of the roughness penalty gamma conjugate prior as parameters to be estimated. This solution requires a reparametrisation beforehand. The second solution suggests to use as prior distribution for the penalty parameter a weighted sum of gamma distributions with different values for b. These two models make the fitting procedure automatic since we do not require a prespecification of a value for b. It is fast and easy to implement since one can simulate from the posterior using the Gibbs sampler.

We also propose an extension to adaptive penalties. This extension can be useful when the underlying function has second derivative varying with x. In this case, adaptive penalties provide more flexibility and increase the quality of the fit. Some suggestions have already been given to work in an adaptive way (Denison *et al.* 2002 ; Lang and Brezger 2004 ; Baladandayuthapani *et al.* 2005). The approach that we propose here is to consider a different penalty parameter at each knot and to obtain the vector of these penalty parameters sequentially, by multiplying the previous one by a gamma random variable with mean 1 and a large variance. This construction yields a smooth evolution of the penalty parameter with x. With this specification, one can still use the Gibbs sampler. The presented techniques are illustrated by smoothing pharmacokinetics data.

The plan of our paper is as follows. In Section 2, we review the basic Bayesian P-splines model and we highlight the crucial role of the hyperparameters a and b. In Section 3, we present two solutions to avoid the choice of b. There, we comment the results of a simulation study realised to compare the two proposed methods with the one in Lang and Brezger (2004). Section 4 provides the extension to adaptive penalties. We conclude our presentation in Section 5 with a discussion.

2 Basic Bayesian P-splines model

The reader is expected to be familiar with P-splines. A first introduction can be found for example in Ruppert *et al.* (2003). We give here a brief summary of the ideas provided by Eilers and Marx (1996). A B-spline of degree q consists of $q+1$ polynomial pieces, each of degree q. These polynomial pieces join at q inner knots. The B-spline is positive on a domain spanned by $q+2$ knots and it is zero everywhere else. A property of B-splines is that the derivatives up to order q-1 are continuous at the joining points. Let $B_j(x;q)$ denote the value at x of the jth B-spline of degree q for a given equidistant grid of knots. A fitted curve $\hat{y}$ to data $\{(x_i, y_i)\}$ is a linear combination $\hat{y}(x) = \sum_{j=1}^{n} \hat{\theta}_j B_j(x;q)$. If we consider the regression of m data points (x_i, y_i) on a set of n B-splines $B_j(.)$, the least squares objective function to minimise is :

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} \theta_j B_j(x_i)\}^2$$

The fitted curve will show more variation than is justified by the data if we let the number of knots be relatively large. To make the result less flexible, Eilers and Marx (1996) propose to introduce a penalty on finite differences of the coefficients of adjacent B-splines :

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} \theta_j B_j(x_i)\}^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k \theta_j)^2$$

In terms of likelihood, the penalty appears as a term that we subtract from the

log-likelihood $l(y;\theta)$. The penalised likelihood function has the following form :

$$l_{\text{pen}} = l(y;\theta) - \frac{\lambda}{2} \sum_{j=k+1}^{n} (\Delta^k \theta_j)^2$$

2.1 Model specification

Let us remind the specification of the basic Bayesian P-splines model (Lang and Brezger 2004). The roughness penalty from the frequentist penalised likelihood approach translates into a prior distribution for the rth order differences of successive B-splines parameters, θ_j, yielding for a conditional normal response.

$$\begin{aligned}
(Y_x|\boldsymbol{\theta},\tau) &\sim \mathcal{N}\left(\boldsymbol{b}_x'\boldsymbol{\theta},\ \tau^{-1}\right) \\
p(\tau) &\propto \tau^{-1} \\
p(\boldsymbol{\theta}) &\propto \exp\left[-0.5\ \tau_\lambda\ \boldsymbol{\theta}' P \boldsymbol{\theta}\right] \\
\tau_\lambda &\sim \mathcal{G}\left(a,\ b\right)
\end{aligned}$$

where

- Y_x is a vector of responses, depending on x,
- $\boldsymbol{b}_x$ is the B-splines basis evaluated at x and associated to a large number of equidistant knots,
- $\boldsymbol{\theta}$ is the vector of B-splines coefficients,
- $P = D'D$ is the penalty matrix and D the rth-order difference matrix, yielding $\boldsymbol{\theta}' P \boldsymbol{\theta} = \sum_k (\Delta^r \theta_k)^2$ where Δ is the first-order difference operator. Thus, for $r = 2$, we have
$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix}$$
- τ_λ is the roughness penalty parameter,
- $\mathcal{G}\left(a,\ b\right)$ denotes a gamma distribution with mean a/b and variance a/b^2. A large variance conjugate prior distribution is usually recommended and specified for τ_λ, as suggested by Lang and Brezger (2004) by setting a equal to 1 and b equal to a small quantity (10^{-5}, say).

2.2 Conditional posterior distributions

Given a set $\boldsymbol{y} = \{y_{x_1}, \dots, y_{x_n}\}$ of independent observations, one can derive the conditional posterior distributions :

$$\begin{aligned}
(\boldsymbol{\theta}|\tau,\tau_\lambda;\boldsymbol{y}) &\sim \mathcal{N}\left(\tau\ \Sigma_\theta B' R^{-1} \boldsymbol{y},\ \Sigma_\theta\right) \\
(\tau|\text{rest};\boldsymbol{y}) \equiv (\tau|\boldsymbol{\theta};\boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ n,\ 0.5\ (y - B\boldsymbol{\theta})' R^{-1} (y - B\boldsymbol{\theta})\right) \\
(\tau_\lambda|\text{rest};\boldsymbol{y}) \equiv (\tau_\lambda|\boldsymbol{\theta};\boldsymbol{y}) &\sim \mathcal{G}\left(a + 0.5\ \rho(P),\ b + 0.5\ \boldsymbol{\theta}' P \boldsymbol{\theta}\right)
\end{aligned}$$

where $\rho(P)$ is the rank of P and

$$B = [\boldsymbol{b}_{x_1}, \dots, \boldsymbol{b}_{x_n}]', \quad R = I_n \ \text{ and } \ \Sigma_\theta^{-1} = \tau B' R^{-1} B + \tau_\lambda P$$

and 'rest' generically denotes all the other parameters from the joint distribution. These formulas can be used to sample from the posterior distributions using the Gibbs sampler.

2.3 Marginal posterior distributions

The roughness penalty parameter τ_λ can be integrated out (Lambert 2005) yielding the marginal posterior distribution

$$p(\boldsymbol{\theta}, \tau | \boldsymbol{y}) \propto \frac{L(\boldsymbol{\theta}, \tau; \boldsymbol{y})\ p(\tau)}{\left(1 + \frac{1}{2b}\ \boldsymbol{\theta}' P \boldsymbol{\theta}\right)^{a+0.5\rho(P)}} \tag{1}$$

where $L(\boldsymbol{\theta}, \tau; \boldsymbol{y})$ is the likelihood. In a classical (frequentist) likelihood framework, this suggests working with a log-likelihood from which the following penalty is subtracted

$$\text{pen}(a, b) \doteq \left[a + \frac{1}{2}\ \rho(P)\right]\ \log\left(1 + \frac{1}{2b}\ \boldsymbol{\theta}' P \boldsymbol{\theta}\right)$$

where "$\doteq$" indicates equality up to an additive constant. This is to be compared with the classical penalised log-likelihood (Eilers and Marx 1996)

$$l_{\text{pen}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\lambda\ \theta' P \boldsymbol{\theta}$$

where λ is usually selected using cross-validation or an information criterion such as the AIC. There, the final λ is a function of $\boldsymbol{y}$.

It is interesting to note the limiting behaviour of the conditional posterior distribution for $\boldsymbol{\theta}$ when the prior variance for τ_λ tends to infinity, as obtained by letting b tend to 0^+. In that case, the denominator in Equation (1) tends to infinity whatever the value of a, except if $\boldsymbol{\theta}' P \boldsymbol{\theta}$ also tends to 0. With a second order penalty, this happens when the fitted mean $B\boldsymbol{\theta}$ tends to a line. Thus, what was first thought to be an expression of our ignorance concerning the appropriate penalty actually translates to an extremely severe roughness penalty.

Another interesting limiting behaviour of the penalty is

$$\lim_{\substack{a, b \to +\infty \\ \mathrm{E}(\tau_\lambda) = \frac{a}{b} \to \mu_\lambda < \infty}} \text{pen}(a, b) \doteq \frac{1}{2}\ \mu_{\tau_\lambda}\ \boldsymbol{\theta}' P \boldsymbol{\theta}$$

It is associated to an informative gamma prior distribution for τ_λ with given mean μ_{τ_λ} and a variance, μ_{τ_λ}/b, tending to zero. It corresponds to the classical penalised log-likelihood with penalty parameter equal to the prespecified prior mean μ_{τ_λ}.

These two extreme situations reveal the potential sensitivity of the results to the choice of b. This suggests that it should also be seen as a parameter in the model.

Another way to get the posterior distribution in Equation (1) is to consider the following equivalent model:

$$\begin{aligned} (Y_x | \boldsymbol{\theta}, \tau) &\sim \mathcal{N}\left(\boldsymbol{b}_x' \boldsymbol{\theta},\ \tau^{-1}\right) \\ p(\tau) &\propto \tau^{-1} \\ D\boldsymbol{\theta} &\sim t_{\nu=2a}\left(\mathbf{0}, \frac{b}{a}\ I_{\rho(P)}\right) \end{aligned}$$

where $t_\nu(\boldsymbol{\mu}, \sigma)$ is the multivariate Student-t distribution with ν degrees of freedom, mean $\boldsymbol{\mu}$ and variance-covariance matrix $\frac{\nu}{\nu-2}\Sigma$ when these two moments exist.

It highlights the crucial roles of a and b. A small value for $2a = \nu_d$ allows an occasionally very large second-order difference between successive components of $\boldsymbol{\theta}$, while the ratio $b/a = \delta_d$ determines the 'marginal' prior dispersion of these differences (with a variance equal to $b/(a-1)$ when it exists).

It suggests a reparametrisation of the basic P-splines model in Section 2.1 obtained by replacing a and b by $\nu_d/2$ and $\delta_d \nu_d/2$, respectively.

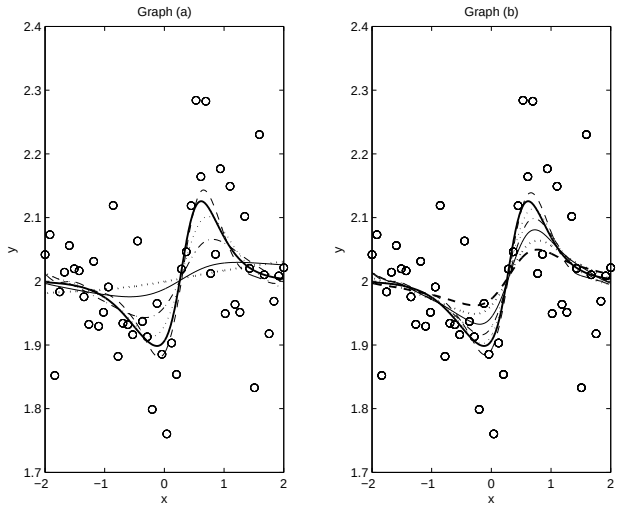


Figure 1: Illustrative data: fitted P-splines curves using the basic Bayesian P-splines model. Graph (a) : $a = 1$ combined with $b = 1$ (dashed), $b = 0.1$ (thin dotted), $b = 0.01$ (dash-dotted), $b = 0.001$ (thin solid). The fitted curves with $b = 0.0001$ and $b = 0.00001$ are confused on the thick dotted line. Graph (b) : $a = b = 1$ (dashed), $a = b = 0.1$ (thin dotted), $a = b = 0.01$ (dash-dotted), $a = b = 0.001$ (thin solid), $a = b = 0.0001$ (thick dotted), $a = b = 0.0001$ (thick dashed). On both graphs, the underlying μ_x corresponds to the thick solid line.

2.4 Illustration

An illustration of the limitations of the basic Bayesian P-splines model is obtained by applying it on 50 simulated data from the function $y_x = \mu_x + \epsilon_x$ with $\epsilon_x \sim \mathcal{N}(0,\ 0.0169)$ and

$$\mu_x = \left(1 + \mathrm{e}^{-4(x-0.3)}\right)^{-1} + \left(1 + \mathrm{e}^{3(x-0.2)}\right)^{-1} + \left(1 + \mathrm{e}^{-4(x-0.7)}\right)^{-1} + \left(1 + \mathrm{e}^{5(x-0.8)}\right)^{-1} \quad (2)$$

Different P-splines curves were fitted using the cubic B-splines associated to 20 equally spaced knots on $(-2, 2)$. It is suggested in Lang and Brezger (2004) to standardise the vector of responses y before estimation and to retransform the results afterwards. On Figure 1, Graph (a) shows the fitted curves with $a = 1$, $b = 1$ (dashed), $b = 0.1$ (thin dotted), $b = 0.01$ (dash-dotted), $b = 0.001$ (thin solid). The fitted curves with $b = 0.0001$ and $b = 0.00001$ are confused on the thick dotted line. The underlying μ_x corresponds to the thick solid line. One can see the strong influence of the choice of the hyperparameter on the resulting fit. A too small value for b, initially thought to express our ignorance about the smoothing variance parameter τ_λ, leads to an oversmoothed curve for the reasons explained in Section 2.3. A larger value for b than suggested in the literature yields a satisfactory fit.

Note that, if we consider the case where $a = b$, we still observe the same dependence of the results on the choice of the value picked for a and b (see Figure 1, graph (b)). Thus, some guidance should be provided to choose these hyperparameters (see Section 3).

3 Specification of the roughness penalty prior distribution

3.1 First method : hyperpriors on the roughness prior

As shown in the previous section, the choice of the parameters a and b for the prior distribution of the penalty parameter τ_λ has an important influence on the smoothness of the fitted curve.

Section 2.3 has highlighted the role of these quantities leading to a reparametrisation in terms of ν_d and δ_d. These two parameters are difficult to prespecify and, hence, it is desirable to see them as parameters to be estimated. This is the topic of this section.

3.1.1 Prior distribution on δ_d

A possible uninformative proper prior distribution for the dispersion parameter δ_d is

$$\delta_d \sim \mathcal{G}\left(a_\delta,\ b_\delta\right)$$

where we may take for instance $a_\delta = b_\delta$ equal to a small value. For a fixed value of ν_d, we have the following conditional posterior distributions :

$$\begin{aligned}
(\boldsymbol{\theta}|\text{rest};\boldsymbol{y}) \equiv (\boldsymbol{\theta}|\tau,\tau_\lambda;\boldsymbol{y}) &\sim \mathcal{N}\left(\tau\ \Sigma_\theta B' R^{-1}\boldsymbol{y},\ \Sigma_\theta\right)\\
(\tau|\text{rest};\boldsymbol{y}) \equiv (\tau|\boldsymbol{\theta};\boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ n,\ 0.5\ (y-B\boldsymbol{\theta})' R^{-1}(y-B\boldsymbol{\theta})\right)\\
(\tau_\lambda|\text{rest};\boldsymbol{y}) \equiv (\tau_\lambda|\boldsymbol{\theta},\delta_d,\nu_d;\boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ \nu_d + 0.5\ \rho(P),\ 0.5\ \delta_d\nu_d + 0.5\ \boldsymbol{\theta}' P\boldsymbol{\theta}\right)\\
(\delta_d|\text{rest};\boldsymbol{y}) \equiv (\delta_d|\tau_\lambda,\nu_d;\boldsymbol{y}) &\sim \mathcal{G}\left(a_\delta + 0.5\ \nu_d,\ b_\delta + 0.5\ \nu_d\tau_\lambda\right)
\end{aligned}$$

These can be used directly to set up a Gibbs algorithm.

3.1.2 Prior distribution on ν_d

We propose to take a uniform prior for ν_d on $(0, K)$:

$$p(\nu_d) \propto I_{(0,K)}(\nu_d)$$

where $K > 0$ is a large degrees of freedom yielding a Student-t density hardly distinguishable from the normal one.

Thus, the conditional posterior distribution for ν_d is

$$p(\nu_d|\boldsymbol{\theta},\tau,\tau_\lambda,\delta_d;\boldsymbol{y}) \equiv p(\nu_d|\tau_\lambda,\delta_d;\boldsymbol{y}) \propto \frac{1}{\Gamma\left(\frac{\nu_d}{2}\right)}\ \left(\delta_d\tau_\lambda\frac{\nu_d}{2}\right)^{\nu_d/2}\ \mathrm{e}^{-\delta_d\tau_\lambda\frac{\nu_d}{2}}\ I_{(0,K)}(\nu_d) \quad (3)$$

In principle, any prior distribution can be considered for ν_d. But none of them will provide an identifiable conditional posterior distribution for ν_d. Therefore, a Metropolis-Hastings step will be required to generate a chain for ν_d.

Alternatively, the Stirling's formula could be used in Equation (3) to approximate $\Gamma(\nu_d/2)$:

$$\Gamma(\nu_d/2) = \frac{2}{\nu_d}\ \Gamma(\nu_d/2+1) \approx \sqrt{2\pi}\left(\frac{\nu_d}{2}\right)^{\nu_d/2-1/2}\mathrm{e}^{-\frac{\nu_d}{2}} \quad (4)$$

Substituting Equation (4) in Equation (3), one obtains

$$\frac{1}{2\sqrt{\pi}}\ \nu_d^{1/2}\ \mathrm{e}^{-\frac{1}{2}(\delta_d\tau_\lambda - \log\delta_d - \log\tau_\lambda - 1)\nu_d}\ I_{(0,K)}(\nu_d) \quad (5)$$

As the Stirling's formula provides an excellent approximation[1] to the gamma function for most of the relevant values of ν_d, one can use the gamma density

$$\mathcal{G}\left(1.5,\ 0.5\ [\delta_d\tau_\lambda - \log\delta_d - \log\tau_\lambda - 1]\right)$$

(as suggested by Equation (5)) truncated to $(0, K)$ in an independence sampler to generate from the posterior distribution in Equation (3).

But, in our experience, no relevant information concerning the degrees of freedom can be obtained in practice, our MCMC simulations yielding a posterior distribution very close to a uniform on $(0, K)$. This is not surprising as, when $a_\delta = b_\delta$ are small,

$$E(\delta_d|\boldsymbol{\theta}, \tau, \tau_\lambda, \delta_d, \nu_d; \boldsymbol{y}) = \frac{a_\delta + 0.5\ \nu_d}{b_\delta + 0.5\ \nu_d\tau_\lambda} \approx \frac{1}{\tau_\lambda}$$

suggesting that the second parameter of the truncated gamma approximating the conditional posterior distribution of ν_d is expected to take small values. This corresponds to a posterior distribution with a large variance, as observed in our unreported examples.

Therefore, we simply suggest to fix ν_d to some value and to evaluate the sensitivity of the fitted curve to that choice.

3.2 Alternative method: a mixture prior for the penalty

An alternative solution to avoid the sensitive choice of b is to consider as prior distribution for τ_λ a weighted sum of M gamma distributions with different values for b.

This leads to the following model specification :

$$\begin{aligned}
(Y_x|\boldsymbol{\theta}, \tau) &\sim \mathcal{N}\left(\boldsymbol{b}_x'\boldsymbol{\theta},\ \tau^{-1}\right)\\
p(\tau) &\propto \tau^{-1}\\
p(\boldsymbol{\theta}) &\propto \exp\left[-0.5\ \tau_\lambda\ \boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta}\right]\\
\tau_\lambda &\sim \sum_{m=1}^{M} p_m\mathcal{G}\left(a,\ b_m\right)\\
\boldsymbol{p} &\sim \mathcal{D}(\boldsymbol{u})
\end{aligned}$$

where $\{b_1, ..., b_M\}$ is a set of prespecified values, $\mathcal{D}$ stands for the Dirichlet distribution, and $\boldsymbol{u}' = \{u_1, \ldots, u_M\}$ is a set of (small and equal) hyperprior parameters expressing our likely prior ignorance about the optimal choice for b.

The conditional posterior distributions are :

$$\begin{aligned}
(\boldsymbol{\theta}|\text{rest}; \boldsymbol{y}) \equiv (\boldsymbol{\theta}|\tau, \tau_\lambda; \boldsymbol{y}) &\sim \mathcal{N}\left(\tau\ \Sigma_\theta B'R^{-1}\boldsymbol{y},\ \Sigma_\theta\right)\\
(\tau|\text{rest}; \boldsymbol{y}) \equiv (\tau|\boldsymbol{\theta}; \boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ n,\ 0.5\ (y - B\boldsymbol{\theta})'R^{-1}(y - B\boldsymbol{\theta})\right)\\
(\tau_\lambda|\text{rest}; \boldsymbol{y}) \equiv (\tau_\lambda|\boldsymbol{\theta}, \boldsymbol{p}; \boldsymbol{y}) &\sim \sum_{m=1}^{M} p_m\mathcal{G}\left(a + 0.5\ \rho(P),\ b_m + 0.5\ \boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta}\right)\\
(\boldsymbol{p}|\text{rest}; \boldsymbol{y}) \equiv (\boldsymbol{p}|\tau_\lambda; \boldsymbol{y}) &\propto \sum_{m=1}^{M} \frac{c_m}{\sum_{j=1}^{M} c_j}\mathcal{D}(u_1, ..., u_{m-1}, u_m + 1, u_{m+1}, ...u_M)
\end{aligned}$$

where

$$c_m = \exp(-\tau_\lambda b_m)b_m^a\frac{\sum_{j=1}^{M} u_j}{u_m}$$

[1]Stirling's formula underestimates the exact value of the gamma function by about 4 (3, 2) percents for an argument of the gamma function greater of equal to 3 (4, 5).

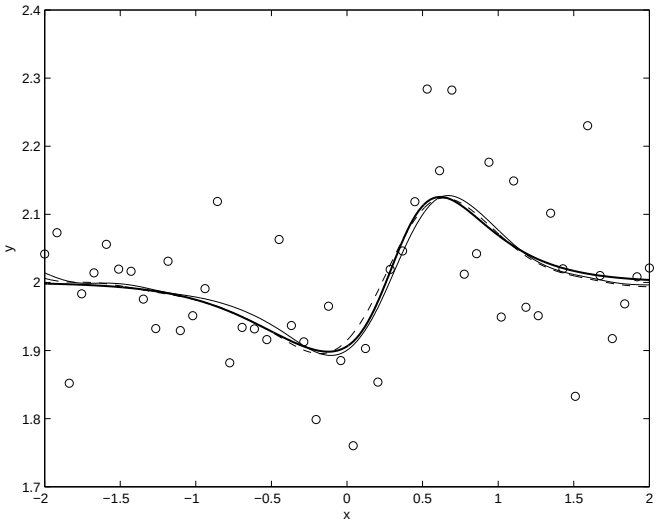


Figure 2: Illustrative data: fitted curves obtained using a Bayesian P-splines model combined with a mixture prior (dashed) or with a hyperprior on δ_d where $\nu_d = 2a = 2$ and $a_\delta = b_\delta = 0.0001$ (thin solid). The underlying μ_x corresponds to the thick solid line.

3.3 Illustration

Let us consider the same example as in Section 2.4. We use a Gibbs simulation with a chain of length $3,000$ (and a burn-in of $1,000$) to get the fitted curves shown in Figure 2. The curve fitted with the first method is the thin solid one while the dashed one is the curve fitted with the mixture prior method. The underlying μ_x corresponds to the thick solid line. For the mixture prior method, we consider a grid of 33 values for b, logarithmically equally spaced between 10^{-6} and 10^2. We can see that the two fitted curves are close to each other and that both methods provide a satisfactory fit. Concerning the first method, a sensitivity analysis shows that the results do not depend on the choices made for ν, a_δ or b_δ (see Figure 3).

In Figure 4, graph (a) shows the distribution of $b = \delta_d \nu_d/2$, as obtained from the Gibbs simulation with the first method. The posterior distribution suggests pretty large value for b (compared to the values recommended in the literature for that quantity). The mode of this distribution is 0.0046. Graph (b) represents the weights associated with each value of the grid of b for the mixture prior method. The value of b with the largest weight is 0.0032.

3.4 Simulation study

We have performed a simulation study to compare the performances of the two proposed specifications for the penalty prior with the basic Bayesian P-splines model with different values for b (0.1, 0.01, 0.001 and 0.0001). To simulate the data, we consider the same functions as in Lang and Brezger (2004), i.e. a linear function, $f(x) = \frac{1}{1.758}x$, a quadratic one, $f(x) = \frac{1}{2.75}x^2 - 1.5$, and a sinusoidal one $f(x) = \frac{1}{0.72}\sin(x)$. We also take the same values for the overall variance parameter σ^2, i.e. σ=1, 0.5 and 0.33. We simulated 100 repetitions for each of the nine combinations with $n = 20$ design points[2] on an equidistant grid between -3 and 3. We also considered the 'illustration function' presented in Section 2.4, i.e. $y_x = \mu_x + \epsilon_x$ with

[2]100 design points were considered in Lang and Brezger (2004)

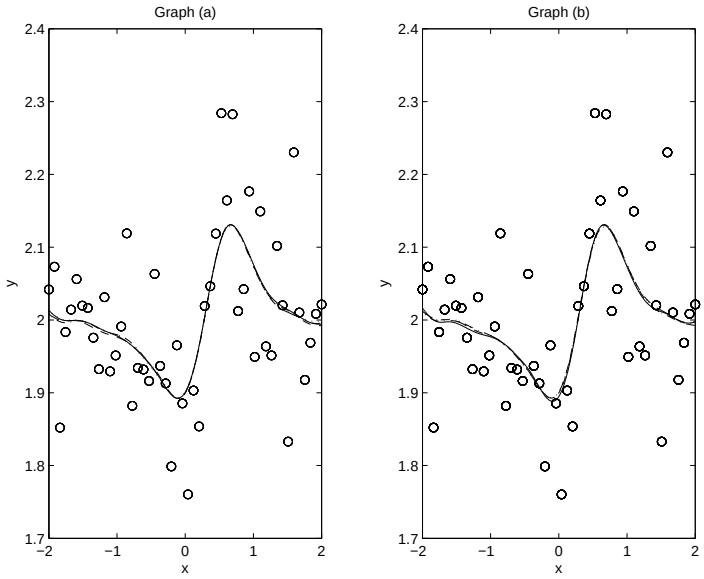


Figure 3: Illustrative data: sensitivity analysis of the fitted curve obtained with a Bayesian P-splines model combined with a hyperprior on δ_d. Graph(a): sensitivity to the choice of $a_\delta = b_\delta$: $a_\delta = b_\delta = 0.00001$ (solid line), $a_\delta = b_\delta = 0.0001$ (dashed line), $a_\delta = b_\delta = 0.001$ (dotted line), $a_\delta = b_\delta = 0.01$ (dashed-dotted line). Graph (b) : sensitivity to the choice of ν : $\nu = 2$ (solid line), $\nu = 5$ (dashed line), $\nu = 10$ (dotted line), $\nu = 20$ (dashed-dotted line).

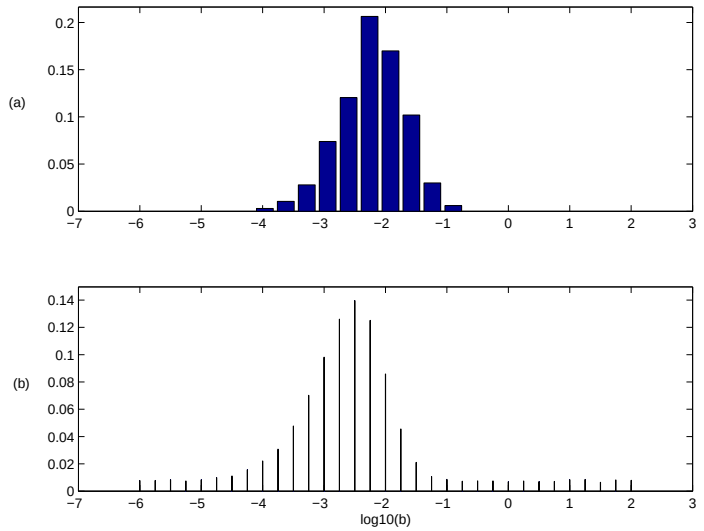


Figure 4: Illustrative data: posterior distribution of b in the Bayesian P-splines model with (a) a hyperprior on δ_d where $b = \delta_d \nu_d/2$; (b) a mixture prior.

Function	σ	Hyperprior on δ_d	Mixture prior
Linear	1	0.049	0.010
	0.5	0.012	0.003
	0.33	0.005	0.003
Quadratic	1	0.238	0.056
	0.5	0.125	0.032
	0.33	0.096	0.032
Sine	1	0.170	0.056
	0.5	0.158	0.056
	0.33	0.140	0.056
Illustration		0.003	0.001

Table 1: Simulation study: median of the posterior modes for b under our two proposed priors in the Bayesian P-splines model.

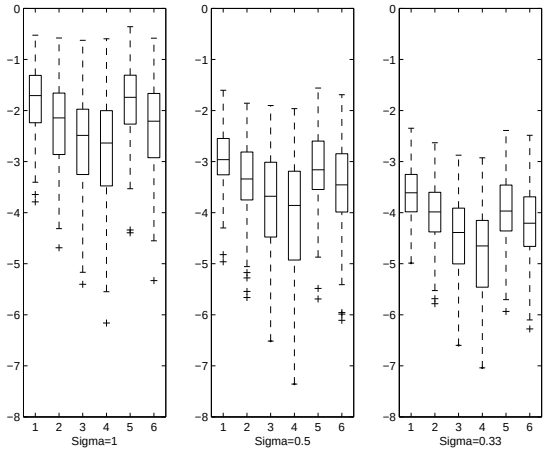


Figure 5: Simulation study: boxplots of $\log(MSE)$ for the linear function. The considered models are the basic Bayesian P-splines model with b=0.1, 0.01, 0.001 or 0.0001 (Priors 1 to 4) ; the Bayesian P-splines model with a hyperprior on δ_d (Prior 5) or a mixture prior (Prior 6).

$\epsilon_x \sim \mathcal{N}(0,\ 0.0169)$ and μ_x given by Equation (2) with $n = 50$ design points between -2 and 2. In all cases, we took cubic B-splines with 20 equidistant knots.

For the first specification (cf. hyperprior on δ_d), we take $a_\delta = b_\delta = 0.0001$. Note that, in the nine simulation settings, a sensitivity analysis revealed no significant influence of $a_\delta = b_\delta$ on the fit.

For the mixture prior method, a grid of 33 values logarithmically equally spaced between 10^{-6} and 10^2 was taken for b.

The quality of the fit is measured by the logarithm of the empirical mean squared error given by :

$$\log(MSE(\hat{f})) = \log\{\frac{1}{100}\sum_{i=1}^{100}(f(x_i) - \hat{f}(x_i))^2\}$$

smaller values indicating better performances.

The results of the simulations are summarised in Figs 5 to 8. Table 1 also provides the median of the posterior modes for b under our two proposed priors.

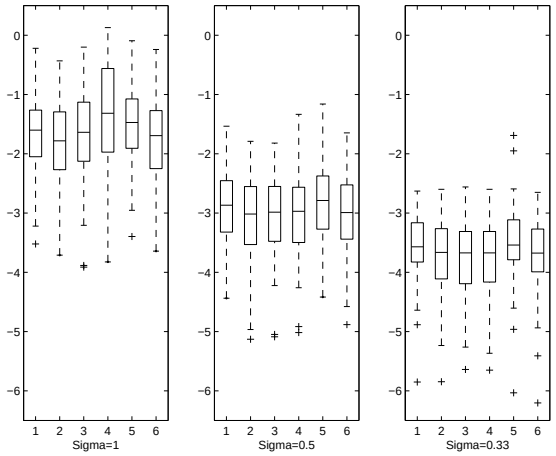


Figure 6: Simulation study: boxplots of $\log(MSE)$ for the quadratic function. The considered models are the basic Bayesian P-splines model with b=0.1, 0.01, 0.001 or 0.0001 (Priors 1 to 4) ; the Bayesian P-splines model with a hyperprior on δ_d (Prior 5) or a mixture prior (Prior 6).

In the linear case (see Fig 5), the best results are obtained with the basic Bayesian P-splines model with the smallest value for b (= 0.0001). That small value for b implies a large penalty (cf. Section 2.3) with an expected linear fit at the limit when b tend to 0^+. Our two priors provide higher $\log(MSE)$ with better results for the mixture prior.

The influence of b on the fit in the basic Bayesian P-splines model is negligible in the quadratic case, all methods performing equally well (see Fig 6).

It is not true anymore with the sine function (see Fig 7) where the recommended small values for b generate relatively large MSE's when the signal-to-noise ratio is low or very low. Then, larger values for b should be considered to be competitive with our proposals for the prior.

The same conclusions apply for the "illustration function" (see Fig 8).

This suggests that the recommendation to take a very small value for b may reveal not to be a good choice in specific circumstances (such as a small sample size and/or data with a moderate signal-to-noise ratio). This would, of course, be revealed by a sensitivity analysis of the results to the choice of b. However, the ease with which the mixture prior method can be implemented makes it attractive while relieving us from such an analysis.

Finally, note that the hyperprior on δ_d method generally yields MSE's that are a bit larger than the ones obtained with the mixture prior. Nevertheless, it often performs better that the basic Bayesian P-splines model with the recommended values for b.

4 P-splines model with smoothly adapting penalties

An adaptive penalty is desirable to smooth function with a second derivative significantly varying with x.

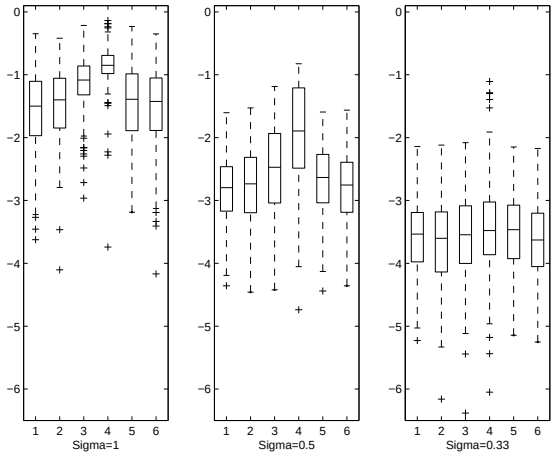


Figure 7: Simulation study: boxplots of $\log(MSE)$ for the sine function. The considered models are the basic Bayesian P-splines model with b=0.1, 0.01, 0.001 or 0.0001 (Priors 1 to 4) ; the Bayesian P-splines model with a hyperprior on δ_d (Prior 5) or a mixture prior (Prior 6).

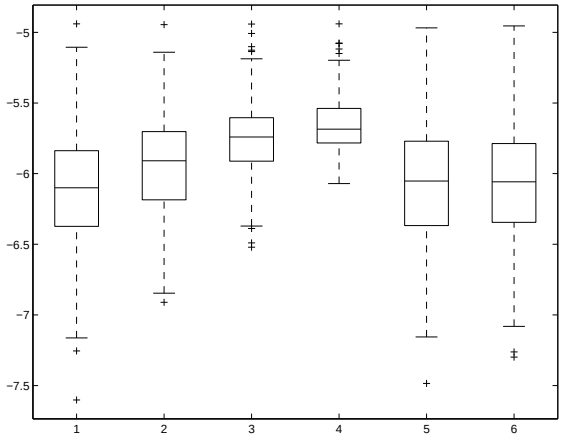


Figure 8: Simulation study: boxplots of $\log(MSE)$ for the illustration function. The considered models are the basic Bayesian P-splines model with b=0.1, 0.01, 0.001 or 0.0001 (Priors 1 to 4) ; the Bayesian P-splines model with a hyperprior on δ_d (Prior 5) or a mixture prior (Prior 6).

4.1 Smoothly adapting penalties combined with a hyperprior for δ_d

Smoothly adapting penalties can be added with the following model specification:

$$\begin{aligned}
(Y_x|\boldsymbol{\theta},\tau) &\sim \mathcal{N}\left(\boldsymbol{b}_x'\boldsymbol{\theta},\ \tau^{-1}\right)\\
p(\tau) &\propto \tau^{-1}\\
p(\boldsymbol{\theta}) &\propto \exp\left[-0.5\ \tau_\lambda \sum_{k=r+1}^{K}\left(\prod_{l=r+1}^{k}\lambda_l\right)(\Delta^r\theta_k)^2\right]\\
&= \exp\left[-0.5\ \tau_\lambda \sum_{k=r+1}^{K}\lambda^{(k)}(\Delta^r\theta_k)^2\right]\\
&= \exp\left[-0.5\ \tau_\lambda\ \boldsymbol{\theta}'D'\Lambda D\boldsymbol{\theta}\right]\\
\lambda_k &\sim \mathcal{G}\left(\omega,\ \omega\right) \quad \text{when} \quad k > r+1 \ \ ; \ \ \lambda_{r+1}=1\\
\tau_\lambda &\sim \mathcal{G}\left(0.5\ \nu_d,\ 0.5\ \delta_d\nu_d\right)\\
\delta_d &\sim \mathcal{G}\left(a_\delta,\ b_\delta\right)
\end{aligned}$$

where

$$\Lambda = \text{diag}\left(\lambda^{(r+1)},\ldots,\lambda^{(K)}\right)$$

That diagonal matrix contains a penalty parameter for each rth-order difference between successive components of $\boldsymbol{\theta}$. They are obtained sequentially by multiplying the previous one by a gamma random variable with mean 1 and (an arbitrarily large) variance ω^{-1}. That construction yields a smooth evolution of the penalty parameters with x.

Note that this proposal differs from Lang and Brezger (2004) where no smoothness is imposed on the roughness penalty coefficient. Such a smoothness was imposed in Baladandayuthapani *et al.* (2005) on the log-scale of the variance. Unfortunately, it required the use of the Metropolis-Hastings as some conditional distributions could not be identified anymore.

Here, the Gibbs sampler can be used as all the conditional distributions can be identified:

$$\begin{aligned}
(\boldsymbol{\theta}|\tau,\tau_\lambda,\delta_d,\boldsymbol{\lambda};\boldsymbol{y}) &\sim \mathcal{N}\left(\tau\ \Sigma_\theta B'R^{-1}\boldsymbol{y},\ \Sigma_\theta\right)\\
(\tau|\text{rest};\boldsymbol{y}) \equiv (\tau|\boldsymbol{\theta};\boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ n,\ 0.5\ (y-B\boldsymbol{\theta})'R^{-1}(y-B\boldsymbol{\theta})\right)\\
(\lambda_l|\text{rest};\boldsymbol{y}) \equiv (\lambda_l|\boldsymbol{\theta},\tau_\lambda,\boldsymbol{\lambda}_{-l};\boldsymbol{y}) &\overset{l>r+1}{\sim} \mathcal{G}\left(\omega+\frac{K-l+1}{2},\ \omega+\frac{\tau_\lambda}{2}\sum_{k=l}^{K}\frac{\lambda^{(k)}}{\lambda_l}(\Delta^r\theta_k)^2\right)\\
(\tau_\lambda|\text{rest};\boldsymbol{y}) \equiv (\tau_\lambda|\boldsymbol{\theta},\delta_d,\boldsymbol{\lambda};\boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ \nu_d+0.5\ \rho(P),\ 0.5\ \delta_d\nu_d+0.5\ \boldsymbol{\theta}'D'\Lambda D\boldsymbol{\theta}\right)\\
(\delta_d|\tau,\tau_\lambda,\boldsymbol{\lambda};\boldsymbol{y}) &\sim \mathcal{G}\left(a_\delta+0.5\ \nu_d,\ b_\delta+0.5\ \nu_d\tau_\lambda\right)
\end{aligned}$$

where

$$\Sigma_\theta^{-1} = \tau B'R^{-1}B + \tau_\lambda D'\Lambda D$$

4.2 Smoothly adapting penalties with a mixture prior for the reference penalty

If, instead, a mixture prior for the reference penalty is considered (cf. Section 3.2), we get the following model specification:

$$(Y_x|\boldsymbol{\theta},\tau) \sim \mathcal{N}\left(\boldsymbol{b}_x'\boldsymbol{\theta},\ \tau^{-1}\right)$$

x	y	x	y	x	y
0	0	4.47e-02	3.142	2.89e-01	3.278
2.74e-03	0.399	5.98e-02	3.742	3.72e-01	2.628
6.25e-03	1.138	7.90e-02	3.519	4.77e-01	2.292
1.07e-02	1.511	1.03e-01	3.067	6.11e-01	2.359
1.64e-02	2.005	1.34e-01	3.870	7.82e-01	2.011
2.36e-02	2.957	1.74e-01	2.977	1.00e+00	1.717
3.29e-02	3.421	2.25e-01	3.093		

Table 2: Simulated pharmacokinetics data corresponding to a two-compartment model with multiplicative log-normal error: $y_x = \mu_x \exp(\epsilon_x)$ with $\epsilon_x \sim \mathcal{N}(0,\ 0.01)$ and $\mu_x = \frac{A\ k_a}{k_a - k_e}[\exp(-k_e x) - \exp(-k_a x)]$ where $A = 3.74,\ k_e = 0.78,\ k_a = 50$.

$$
\begin{aligned}
p(\tau) &\propto \tau^{-1} \\
p(\boldsymbol{\theta}) &\propto \exp\left[-0.5\ \tau_\lambda\ \boldsymbol{\theta}' D' \Lambda D \boldsymbol{\theta}\right] \\
\lambda_k &\sim \mathcal{G}(\omega,\ \omega) \quad \text{when} \quad k > r+1 \quad ; \quad \lambda_{r+1} = 1 \\
\tau_\lambda &\sim \sum_{m=1}^{M} p_m \mathcal{G}(a,\ b_m) \\
\boldsymbol{p} &\sim \mathcal{D}(\boldsymbol{u})
\end{aligned}
$$

The conditional posterior distributions are:

$$
\begin{aligned}
(\boldsymbol{\theta}|\text{rest}; \boldsymbol{y}) \equiv (\boldsymbol{\theta}|\tau, \tau_\lambda \boldsymbol{\lambda}; \boldsymbol{y}) &\sim \mathcal{N}\left(\tau\ \Sigma_\theta B' R^{-1} \boldsymbol{y},\ \Sigma_\theta\right) \\
(\tau|\text{rest}; \boldsymbol{y}) \equiv (\tau|\boldsymbol{\theta}; \boldsymbol{y}) &\sim \mathcal{G}\left(0.5\ n,\ 0.5\ (y - B\boldsymbol{\theta})' R^{-1} (y - B\boldsymbol{\theta})\right) \\
(\lambda_l|\text{rest}; \boldsymbol{y}) \equiv (\lambda_l|\boldsymbol{\theta}, \tau_\lambda, \boldsymbol{\lambda}_{-l}; \boldsymbol{y}) &\overset{l>r+1}{\sim} \mathcal{G}\left(\omega + \frac{K-l+1}{2},\ \omega + \frac{\tau_\lambda}{2} \sum_{k=l}^{K} \frac{\lambda^{(k)}}{\lambda_l} (\Delta^r \theta_k)^2\right) \\
(\tau_\lambda|\text{rest}; \boldsymbol{y}) \equiv (\tau_\lambda|\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{p}; \boldsymbol{y}) &\sim \sum_{m=1}^{M} p_m \mathcal{G}\left(a + 0.5\ \rho(P),\ b_m + 0.5\ \boldsymbol{\theta}' D' \Lambda D \boldsymbol{\theta}\right) \\
(\boldsymbol{p}|\text{rest}; \boldsymbol{y}) \equiv (\boldsymbol{p}|\tau_\lambda; \boldsymbol{y}) &\propto \sum_{m=1}^{M} \frac{c_m}{\sum_{j=1}^{M} c_j} \mathcal{D}(u_1, ..., u_m + 1, ...u_M)
\end{aligned}
$$

where

$$
c_m = \exp(-\tau_\lambda b_m) b_m^a \frac{\sum_{j=1}^{M} u_j}{u_m}
$$

4.3 Illustration

A demanding illustration of the model performances is obtained by applying it on pharmacokinetics data giving the measured evolution of the concentration of a drug in the plasma over time (see Table 2). The measurement times are approximately equally spaced on the log-scale. Most measurements are taken at early times where the underlying curvature has the largest gradient.

In Figure 9, the curves are fitted with the first method (thin solid line) and with the mixture prior method (dashed line) but without smoothly adapting penalties. The underlying μ_x corresponds to the thick solid line. These curves were obtained from a Gibbs simulation with a chain of length $3,000$ (and a burn-in of $1,000$). For the first method, we use $\nu = 2$ and $a_\delta = b_\delta = 0.0001$. For the mixture

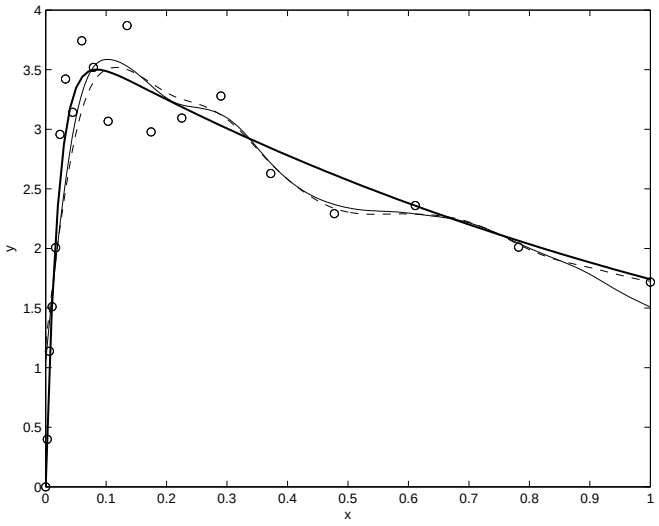


Figure 9: Simulated pharmacokinetics data from Table 2: fitted curves using the Bayesian P-splines models with a hyperprior on δ_d (thin solid line) or a mixture prior (dashed line). The underlying μ_x corresponds to the thick solid line.

prior method, we consider a grid of 33 values for b, logarithmically equally spaced between 10^{-6} and 10^2. For both methods, we obtain a wiggly curve that captures part of the early quick rise in the response, but with overfitting at later times. It is obviously a compromise between ideally a large value for b for small times (where the curvature is large) and a small value for b for later times (where the target curve is approximately a line).

Figure 10 shows the fitted curves obtained with smoothly adapting penalties combined with the hyperprior on δ_d (thin solid line, cf. Section 4.1) and with the mixture prior (dashed line, cf. Section 4.2). The two fitted curves are hardly distinguishable from the target curve in the rising phase that requires flexibility and, thus, a small penalty. A linear pattern is obtained for later times as a consequence of large penalty parameters.

The posterior distribution of $\log(b = \delta_d \nu_d/2)$ under each of the two prior specifications with adaptive penalties are given in Fig 11. The posterior modes under the hyperprior for δ_d and the mixture prior methods are equal to 2.07 (see Fig 11-a-) and 2.25 (see Fig 11-b-), respectively. Note that a larger grid (33 values for b, logarithmically equally spaced between 10^{-6} and 10^5) than before (with an upper limit set previously at 10^2) had to be considered for b.

5 Discussion

In Bayesian P-splines models, the prior distribution for the roughness penalty parameter is usually taken to be a gamma with a fixed to 1 and a small value for b or with $a = b$ equal to small value in order to have a large variance. We have shown that the choice of the gamma parameters can have a critical influence on the resulting smoothness of the fit in some specific circumstances i.e. for some functions (see Section 2.4) or/and when the sample size is small. Indeed, the simulation study has shown that even for simple functions like linear, quadratic or sine ones, the choice of b can have an influence on the fit if the number of observations is small. When the sample size is large, a sensitivity analysis for the choice of the hyperparameters

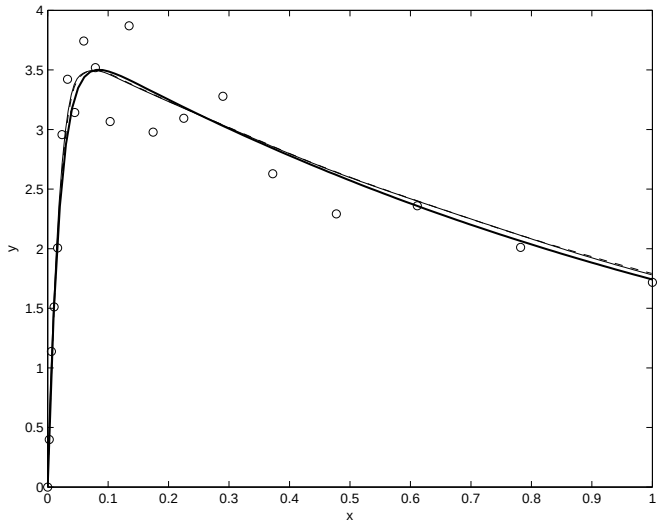


Figure 10: Simulated pharmacokinetics data from Table 2: fitted curves using the Bayesian P-splines models with a hyperprior on δ_d (thin solid line) or a mixture prior (dashed line) in combination with smoothly adapting penalties. The underlying μ_x corresponds to the thick solid line.

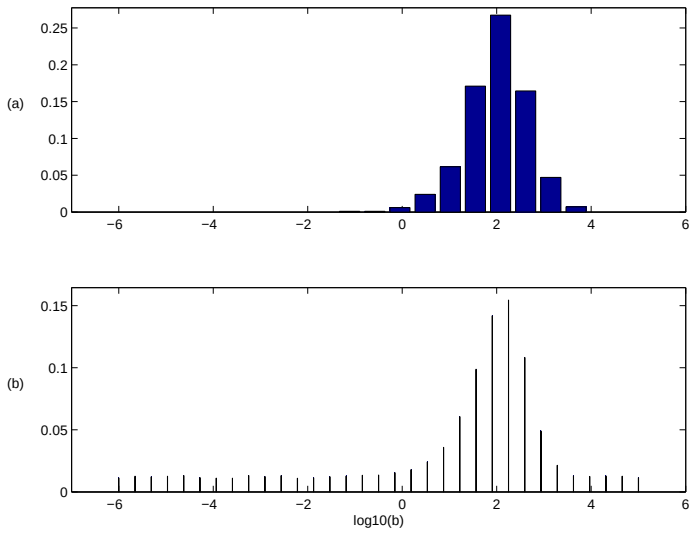


Figure 11: Simulated pharmacokinetics data from Table 2: posterior distribution of $\log(b = \delta_d \nu_d/2)$ under the hyperprior (for δ_d) and the mixture prior methods with adaptive penalties.

often lead to the conclusion that results hardly depend on them. However, we have provided an illustration where the number of observations is large and where the choice of the hyperparameters still has an influence on the fit. In order to spare a likely useless sensitivity analysis and to warrant against the possible consequences of its neglect, it is desirable to make the fitting procedure automatic. The two solutions proposed in this paper do not require a sensitive choice of hyperparameters in the prior distribution. The simulation study suggests that the mixture prior approach performs better.

We have also provided an extension enabling smoothly adapting penalties. Its usefulness was illustrated by fitting pharmacokinetics data. The data of interest show a sharp linear increase of drug concentration at early times, reach a peak and then present a slow linear decrease. Ideally, one should have a large penalty for early and later times combined with a small penalty in between to enable a smooth description of the gradient when its sign changes at the concentration peak.

Combined with one of the two proposed specifications for the reference roughness penalty prior distribution (see Section 3), we end up with a very powerful, easy to set up (cf. Gibbs sampling) and quick Bayesian smoother.

Acknowledgements

Astrid Jullion thanks Eli Lilly for financial support through a patronage research grant and the UCL for a FSR research grant. Financial support from the IAP research network nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) is also gratefully acknowledged by Philippe Lambert.

References

Baladandayuthapani, V., Mallick, B., and Carroll, R. J. (2005). Spatially adaptive Bayesian penalized regression splines. *Journal of Computational Statistics and Data Analysis*. (in press).

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, **97**, 160–169.

Brezger, A. and Lang, S. (2005). Generalized structured additive regresion based on Bayesian P-splines. *Journal of Computational Statistics and Data Analysis*. (in press).

Denison, D., Smith, A. F. M., and Mallick, B. K. (2002). *Bayesian methods for nonlinear classification and regression*. Wiley, Chichester, West Sussex, England.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 715–745.

Lambert, P. (2005). Archimedean copula estimation using Bayesian splines smoothing techniques. Discussion Paper 05-27, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium. http://www.stat.ucl.ac.be/pub/papers/dp/dp05/.

Lambert, P. and Eilers, P. H. (2005). Bayesian proportional hazards model with time varying regression coefficients: a penalized poisson regression approach. *Statistics in Medicine*, **24**. (in press).

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge, UK.