Ann Inst Stat Math (2012) 64:577–613
DOI 10.1007/s10463-010-0321-6

Priors for Bayesian adaptive spline smoothing

Yu Ryan Yue · Paul L. Speckman · Dongchu Sun

Received: 1 September 2009 / Revised: 21 July 2010 / Published online: 18 January 2011

Abstract Adaptive smoothing has been proposed for curve-fitting problems where the underlying function is spatially inhomogeneous. Two Bayesian adaptive smoothing models, Bayesian adaptive smoothing splines on a lattice and Bayesian adaptive P-splines, are studied in this paper. Estimation is fully Bayesian and carried out by efficient Gibbs sampling. Choice of prior is critical in any Bayesian non-parametric regression method. We use objective priors on the first level parameters where feasible, specifically independent Jeffreys priors (right Haar priors) on the implied base linear model and error variance, and we derive sufficient conditions on higher level components to ensure that the posterior is proper. Through simulation, we demonstrate that the common practice of approximating improper priors by proper but diffuse priors may lead to invalid inference, and we show how appropriate choices of proper but only weakly informative priors yields satisfactory inference.

Keywords Adaptive smoothing · Intrinsic autoregressive · Objective priors · Penalized regression · Posterior propriety

Supported by PSC-CUNY Grant 60147-39 40 and National Science Foundation Grant 0720229.

Y. R. Yue (✉)
Department of Statistics and CIS, Baruch College, City University of New York,
One Bernard Baruch Way, New York, NY 10010, USA
e-mail: yu.yue@baruch.cuny.edu

P. L. Speckman · D. Sun
Department of Statistics, University of Missouri, 146 Middlebush Hall,
Columbia, MO 65211, USA
e-mail: speckmanp@missouri.edu

D. Sun
e-mail: sund@missouri.edu

1 Introduction

Adaptive non-parametric regression is a problem that has attracted considerable attention on a variety of fronts. Ordinary non-parametric regression (e.g., Wand and Jones 1995; Eubank 1999) is well known to perform badly when estimating highly varying functions that have peaks, jumps or frequent curvature transitions. Consequently, many authors have proposed modifications or alternative methods to circumvent these difficulties. Our aim here is to examine a special case, Bayesian non-parametric regression. We consider two related methods. The first method, suitable to data collected on a lattice, generalizes a class of intrinsic Gaussian Markov random field priors related to smoothing splines. This class of priors is attractive because the sparse nature of the precision matrix allows efficient computation. The second method is Bayesian P-splines. As with all Bayesian non-parametric methods, choices must be made for certain prior parameters. We believe that "objective" Bayesian methods (e.g., Berger 2006) are attractive to many analysts. However, it is highly problematic that one can specify completely "objective" improper priors in an infinite dimensional problem such as non-parametric regression, and proper priors at some stage are necessary. Thus our goal is to produce rigorously justified priors that are (a) reasonably objective, (b) have guaranteed proper posteriors, and (c) are parameterized in a way that necessary subjective information is easily elicited.

We term the two methods Bayesian adaptive smoothing splines (BASS) and Bayesian adaptive P-splines (BAPS). BASS is closely related to the method in Lang et al. (2002), where non-informative priors on the variance components were used with no proof that the posterior is proper. Similarly, BAPS is closely related to a series of papers (Lang and Brezger 2004; Baladandayuthapani et al. 2005; Crainiceanu et al. 2007) in which proper priors were used to ensure proper posteriors. Those priors, however, are diffuse and improper in the limit. The practice of using proper but diffuse priors is known to be dangerous (Hobert and Casella 1996) and may result in problematic MCMC. In Fig. 2 (see Sect. 3.2), we give an example showing how the diffuse priors recommended in Baladandayuthapani et al. (2005) produce MCMC trace plots that appear acceptable for the first few thousand iterations, but the simulation is far from convergence. In order to use these methods, we believe it is crucial to establish the necessary theory to avoid cases where the posterior is actually highly dependent on choices made for a supposedly "objective" proper but diffuse prior. We derive sufficient conditions on objective, partially improper priors for BASS and BAPS such that the posterior is guaranteed to be proper. With these priors, the posterior is relatively insensitive to choice of prior parameters, and Bayesian inference using our priors is rigorously justified. We also give practical guidelines for choosing among these priors in application.

An exhaustive literature review is beyond the scope of this article, but the work here is motivated by a number of authors, both frequentist and Bayesian. For example, Staniswalis (1989) and Staniswalis and Yandell (1992) used local bandwidth selection for kernel estimates and adaptive smoothing; Cummins et al. (2001) introduced a local cross-validation criterion for adaptive smoothing splines. The BASS methodology is directly related to Abramovich and Steinberg (1996) and Pintore et al. (2006), who used a reproducing kernel Hilbert space representation to derive a smoothing spline when the smoothness penalty is a function $\lambda(t)$ of the design space t. In particular,

Pintore et al. proposed a piecewise constant model for $\lambda(t)$, which provides a convenient computational framework with closed form solutions to the corresponding reproducing kernels of the Hilbert space. Adaptive P-spline models were introduced by Ruppert and Carroll (2000), who proposed penalized splines with a truncated power basis that achieve adaptivity by introducing locally smoothing parameters to difference penalties on the regression coefficients, and then taking another layer P-spline prior on the smoothing parameters.

Another large class of adaptive smoothing methods is based on wavelet shrinkage. Within a Bayesian context, a suitable prior distribution for wavelet coefficients is chosen to adaptively produce sparsity (e.g., Chipman et al. 1997; Clyde et al. 1998; Abramovich et al. 1998). Johnstone and Silverman (2005) and Pensky (2006) have shown that empirical Bayes wavelet estimators not only attain the frequentist optimality over a wide range of inhomogeneous spaces (i.e., Besov spaces) but also outperform standard wavelet estimators in finite sample situations, although their work is beyond the scope of this paper.

There are also several other Bayesian methods including the multiple hyperparameter interpolation model of Mackay and Takeuchi (1998), the mixture of splines model of Wood et al. (2002), regression splines with adaptive knot selection (Friedman 1991; Smith and Kohn 1996; Denison et al. 1998; Di Matteo et al. 2001), and non-stationary Gaussian processes (GP) regression models (Paciorek and Schervish 2004, 2006).

The rest of the paper is organized as follows. In Sect. 2, we review an application of a basic non-parametric regression model used to fit discretized smoothing splines based on a difference approximation. We then show how these processes are generalized to be spatially adaptive for BASS. Sufficient conditions are given in Sect. 2.3 for propriety of the posteriors of BASS and a more general additive model. We then adapt the theoretical results to BAPS in Sect. 3. Some issues about Bayesian computation are discussed in Sect. 4, and several examples are presented in Sect. 5. Conclusions are provided in Sect. 6.

2 Bayesian adaptive smoothing splines

2.1 Smoothing splines on a lattice

Consider the single component non-parametric regression model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $x_i \in \mathbb{R}$ and ε_i is a mean zero noise term with constant variance. The smoothing spline estimator of f is the solution to the optimization problem

$$\hat{f} = \arg\min_f \left[\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left(f^{(p)}(x) \right)^2 \mathrm{d}x \right] \tag{2}$$

for an appropriate smoothing parameter λ, where the cost function (2) trades off fidelity to the data in terms of sum squared error against roughness of the fit as measured by the

L_2 penalty on the pth order derivative (see, e.g., Wahba 1990; Green and Silverman 1994; Eubank 1999). Many authors have noted that a global λ makes smoothing splines perform poorly when estimating spatially inhomogeneous functions.

To establish a connection with a Bayesian model, assume independent and identically distributed Gaussian errors ε_i, i.e., $\varepsilon_i \overset{iid}{\sim} N(0, \tau^{-1})$. We also assume the regular lattice structure $x_1 < x_2 < \cdots < x_n$ and $x_{j+1} - x_j = h$ for $j = 1, \ldots, n-1$. Extensions to non-equally spaced design are briefly outlined at the end of this section. For future convenience, we adopt the notation $z_i = f(x_i)$, $i = 1, \ldots, n$. With $\boldsymbol{y} = (y_1, \ldots, y_n)', \boldsymbol{z} = (z_1, \ldots, z_n)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, the matrix form of model (1) is

$$\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{I}_n). \tag{3}$$

Following a number of authors including Fahrmeir and Wagenpfeil (1996) and Speckman and Sun (2003), a suitable prior on $\boldsymbol{z}$ is motivated by a difference approximation for the smoothing spline penalty term in (2) as follows.

Assuming h is small and $f^{(p)}(x)$ is continuous, $f^{(p)}(x_k) \approx h^{-p}\nabla^p f(x_k)$ for $k = p+1, \ldots, n$, where

$$\nabla^p f(x_k) = \sum_{j=0}^{p} (-1)^j \binom{p}{j} f(x_{k-j}),$$

is the pth order backward difference operator. Then a discretized version of the penalty term in (2) is

$$\int \left(f^{(p)}(x)\right)^2 \mathrm{d}x \approx h^{-(2p-1)} \sum_{k=p+1}^{n} \left[\nabla^p f(x_k)\right]^2. \tag{4}$$

Again with $z_k = f(x_k)$, the quadratic form in (4) can be written as $\boldsymbol{z}'\boldsymbol{A}^{(p)}\boldsymbol{z} = (\boldsymbol{B}_p\boldsymbol{z})'(\boldsymbol{B}_p\boldsymbol{z})$, where $\boldsymbol{B}_p$ is the $(n-p) \times n$ full rank matrix defined by

$$\boldsymbol{B}_p\boldsymbol{z} = \begin{pmatrix} \vdots \\ \nabla^p z_k \\ \vdots \end{pmatrix}_{n-p}.$$

Thus $\boldsymbol{B}_p\boldsymbol{z}$ is the vector of all pth order backward differences and $\boldsymbol{A}^{(p)} = \boldsymbol{B}_p'\boldsymbol{B}_p$ is an $n \times n$ pth order "structure matrix" of rank $n-p$. If we let $\lambda_h = \lambda h^{-(2p-1)}$, the vector $\hat{\boldsymbol{z}}$ defined by

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \left[(\boldsymbol{y} - \boldsymbol{z})'(\boldsymbol{y} - \boldsymbol{z}) + \lambda_h \boldsymbol{z}'\boldsymbol{A}^{(p)}\boldsymbol{z}\right], \tag{5}$$

is a discretized smoothing spline. The minimization criterion in (5) suggests that the prior taken on $\boldsymbol{z}$ for Bayesian smoothing splines (BSS) be

$$[\boldsymbol{z} \mid \delta] \propto \delta^{\frac{1}{2}(n-p)} |\boldsymbol{A}^{(p)}|_+^{\frac{1}{2}} \exp\left(-\frac{\delta}{2}\boldsymbol{z}'\boldsymbol{A}^{(p)}\boldsymbol{z}\right), \tag{6}$$

where δ is a precision that must be specified or estimated. In this expression and the following, we use Bayesian convention where $[\cdot]$ and $[\cdot|\cdot]$ denote unconditional and conditional densities, respectively. Notice that the determinant $|\boldsymbol{A}^{(p)}|_+$ is given by the product of the non-zero eigenvalues of $\boldsymbol{A}^{(p)}$ and is irrelevant in Bayesian computations.

Prior density (6) is an intrinsic Gaussian Markov random field (IGMRF) on a regular lattice in the sense that $\boldsymbol{z}$ is a random vector that follows an improper multivariate normal distribution and satisfies *Markov conditional independence* assumptions (Rue and Held 2005). Note that the null space of $\boldsymbol{A}^{(p)}$ is spanned by pth order polynomials. Speckman and Sun (2003) termed (6) "partially informative" because it is flat on the null space of $\boldsymbol{A}^{(p)}$ and a proper Gaussian prior on the range space of $\boldsymbol{A}^{(p)}$. This prior is also an example of an intrinsic autoregressive (IAR) model used for spatial effects in Bayesian hierarchical models (e.g., Besag and Kooperberg 1995; Fahrmeir and Wagenpfeil 1996). Since the error terms ε_i in (3) are normal and independent of $\boldsymbol{z}$, the posterior distribution of $\boldsymbol{z}$ can be shown to be $N_n\left(\boldsymbol{S}_{\lambda_h}\boldsymbol{y}, \tau^{-1}\boldsymbol{S}_{\lambda_h}\right)$, where the smoothing parameter is $\lambda_h = \delta/\tau$ and the smoother matrix is $\boldsymbol{S}_{\lambda_h} = \left(\boldsymbol{I}_n + \lambda_h \boldsymbol{A}^{(p)}\right)^{-1}$. The posterior mean $\hat{\boldsymbol{z}} = \boldsymbol{S}_{\lambda_h}\boldsymbol{y}$ satisfies (5) and is a Bayesian version of the discretized smoothing spline.

Although we will focus on prior (6) in this article, extension of the IGMRF to a non-equally spaced design is straightforward. Letting $h_k = x_k - x_{k-1}$, $k = 1, \ldots, n-1$, we define $\nabla_k f(x_k) = h_k^{-1}(f(x_k) - f(x_{k-1}))$ and $\nabla_k^2 f(x_k) = h_k^{-1}(\nabla_k f(x_k) - \nabla_{k-1} f(x_{k-1}))$. When $p = 2$, one possible approximation of the penalty (2) for non-equally spaced x is

$$\int f''(x)^2 \mathrm{d}x \approx \sum_{k=3}^{n} h_k \left(\nabla_k^2 f(x_k)\right)^2. \tag{7}$$

Hence we may derive a second order IGMRF on an irregular lattice using the quadratic form in (7). The IGMRFs with other orders can be obtained in a similar way and they are all of form (6) but with different a matrix $\boldsymbol{A}^{(p)}$. Other approaches for constructing irregularly spaced IGMRFs are available in, for instance, Rue and Held (2005) and Lindgren and Rue (2008).

2.2 Adaptive IGMRF priors

An alternative representation of (6) in the context of dynamic or state space modeling is given by

$$\sum_{j=0}^{p} (-1)^j \binom{p}{j} z_{k-j} \overset{iid}{\sim} N\left(0, \delta^{-1}\right), \quad k = p+1, \ldots, n. \tag{8}$$

As suggested by Lang et al. (2002) and Knorr-Held (2003), an adaptive extension of (8) can be achieved by replacing the constant precision δ with locally varying precisions δ_k. Heuristically, the δ_k should be large for the flat parts of the function to be estimated while they should be small for the sharp features. This procedure is equivalent to variable bandwidth selection in non-parametric density and function estimation.

To complete the specification of the adaptive prior, a further prior is taken on δ_k. Let $\delta_k = \delta e^{\gamma_k}$ for $k = p+1, \ldots, n$, where δ is a scale parameter and $\sum_{k=p+1}^{n} \gamma_k = 0$ for identifiability. We then assume the γ_k are also smooth and take a qth order IGMRF prior on $\boldsymbol{\gamma} = (\gamma_{p+1}, \ldots, \gamma_n)'$ subject to the constraint $\mathbf{1}'\boldsymbol{\gamma} = 0$. The priors on z and $\boldsymbol{\gamma}$ can now be written in matrix notation as

$$[z \mid \delta, \boldsymbol{\gamma}] \propto \delta^{\frac{1}{2}(n-p)} |\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}|_{+}^{\frac{1}{2}} \exp\left(-\frac{\delta}{2} z' \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} z\right), \tag{9}$$

where $\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} = \boldsymbol{B}_p' \boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)} \boldsymbol{B}_p$ is a pth order adaptive structure matrix with a diagonal matrix $\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)} = \text{diag}(e^{\gamma_{p+1}}, \ldots, e^{\gamma_n})$, and

$$[\boldsymbol{\gamma} \mid \eta] \propto \eta^{\frac{1}{2}(n-p-q)} \exp\left(-\frac{\eta}{2} \boldsymbol{\gamma}' \boldsymbol{A}^{(q)} \boldsymbol{\gamma}\right) I_{(\mathbf{1}'\boldsymbol{\gamma}=0)}, \tag{10}$$

where $\boldsymbol{A}^{(q)}$ is the $(n-p) \times (n-p)$ qth order structure matrix as defined in (6). Note that (10) defines a singular (proper if $q = 1$) distribution. We refer to (9) and (10) together as an adaptive IGMRF prior for fixed p and q. Since the IGMRF developed from (7) has a similar dynamic expression (with different coefficients) as in (8), the adaptive extension also applies to non-equally spaced design. The resulting adaptive IGMRF on irregular lattice is of the form (9) and (10) as well.

The adaptive IGMRF prior has appealing properties for Bayesian inference and computation. First, it often improves function estimation since the variable precisions can adapt to changes in curvature of the underlying functions (see examples in Sect. 5). Second, we do not need to compute $|\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}|_{+}$ in each MCMC iteration because it is a constant according to Lemma 2 in Appendix A. Third, the full conditional distributions of z and $\boldsymbol{\gamma}$ are normal and log concave, respectively, which is quite convenient for implementing Gibbs sampling. Finally, the sparseness of $\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}$ speeds computation. Further discussion regarding computation is in Sect. 4. In applications, the values of p and q must be decided beforehand. From our experience, $p = 1$ corresponding to linear splines does not smooth enough while $p = 2$ (cubic splines) or $p = 3$ often has good performance for the applications we have tried in function estimation. Following Lang et al. (2002), we suggest taking $q = 1$. Our experience is that the random walk prior with $q = 1$ on the γ_k works well in practice and yields fairly fast MCMC computation. The assumption also simplifies the theoretical development in Sect. 2.3. More importantly, the prior for $\boldsymbol{\gamma}$ in (10) is improper if $q \geq 2$, which leads to improper posteriors according to Sun and Speckman (2008). However, one may take a higher order proper prior on $\boldsymbol{\gamma}$ other than an IGMRF, e.g., a conditional autoregressive (CAR) prior.

Our approach differs from the locally adaptive dynamic modeling in Lang et al. (2002) with respect to the priors used on local precisions, although they appear similar. In our context, Lang et al. directly took a first order IGMRF prior for $\gamma_k = \log(\delta_k)$. Since the precision matrix ($\boldsymbol{A}^{(1)}$ in our notation) is rank deficient, the prior can be expressed as a flat prior on γ_{p+1}, for example, and a proper Gaussian prior on $\gamma_{p+2}, \ldots, \gamma_n$. Equivalently, one could write $\log(\delta_k) = \gamma_0 + \gamma_k$, $k = p+1, \ldots, n$, with a flat prior on γ_0 and a singular proper normal prior on $\gamma_{p+1}, \ldots, \gamma_n$ subject to $\sum_{j>p} \gamma_j = 0$. Identifying $\gamma_0 = \log \delta$, the first order IGMRF prior on the γ_k puts the implicit flat prior on $\log \delta$, $[\log \delta] \propto 1$. But this is equivalent to the invariance prior on the implicit variance δ^{-1}, i.e., $[\delta^{-1}] \propto 1/\delta$. Speckman and Sun (2003) considered the non-adaptive BSS with parameters τ and δ (in our notation) and showed that if the invariance prior is used for δ^{-1}, the posterior is improper for any choice of inverse-gamma prior on τ^{-1}. This would seem to imply that the more complicated setup of Lang et al. with the additional prior on the γ_k would also have an improper posterior.

The adaptive prior proposed here is related to the independent gamma priors on the δ_k introduced by Carter and Kohn (1996) (also see Lang et al. 2002; Brezger et al. 2007). The models proposed here are also related to autoregressive conditional heteroscedasticity (ARCH) models used widely in econometric time series analysis and stochastic volatility models with time-varying and autocorrelated conditional variance, which have various applications for financial data, prediction and filtering. There is a growing literature of Bayesian analysis of these models, e.g., Jacquier et al. (1994), Vrontos et al. (2000), and Nakatsuma (2000).

2.3 Propriety of the posterior for the adaptive IGMRF

To complete the hierarchial specification of BASS, we need hyperpriors on the precision components τ, δ and η. Since it is difficult to elicit subjective priors on those precision components, especially on δ and η, objective or non-informative priors might be preferred in this situation. However, priors that are improper may yield improper posteriors, resulting in invalid Bayesian inference (Hobert and Casella 1996). For non-adaptive smoothing splines, Speckman and Sun (2003) derived necessary and sufficient conditions for the propriety of the posterior for the class of PIN priors with inverse gamma type non-informative priors on both τ and δ. Sun and Speckman (2008) investigated non-informative priors in the context of additive models and concluded that the invariance (improper) prior can be taken on τ, but in this case the priors for the smoothing parameter $\xi = \delta/\tau$ must be proper. Motivated by their work, we first reparameterize the precision components and then find possible objective priors on the new parameters. Let $\xi_1 = \delta/\tau$ and $\xi_2 = \eta/\delta$, forming a one-to-one transformation between (τ, ξ_1, ξ_2) and (τ, δ, η). The adaptive IGMRF prior then can be written as

$$
\begin{aligned}
[\boldsymbol{z} \mid \tau, \xi_1, \boldsymbol{\gamma}] &\propto (\tau\xi_1)^{\frac{1}{2}(n-p)} \exp\left(-\frac{\tau\xi_1}{2}\boldsymbol{z}'\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{z}\right), \\
[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] &\propto (\tau\xi_1\xi_2)^{\frac{1}{2}(n-p-q)} \exp\left(-\frac{\tau\xi_1\xi_2}{2}\boldsymbol{\gamma}'\boldsymbol{A}^{(q)}\boldsymbol{\gamma}\right) I_{(\mathbf{1}'\boldsymbol{\gamma}=0)}.
\end{aligned}
\tag{11}
$$

Finally, we use the Jeffrey's prior (invariance prior) on τ,

$$[\tau] \propto \frac{1}{\tau}. \tag{12}$$

There are several reasons for this reparameterization. First of all, note that ξ_1 is the smoothing parameter in the non-adaptive case; ξ_2 plays a similar role at the higher level, controlling the smoothness of $\boldsymbol{\gamma}$. Thus, it may be more meaningful to deal with ξ_1 and ξ_2 than with δ and η in terms of smoothing. Second, this prior is related to Zellner's g-prior, widely used in Bayesian linear regression and model selection (e.g., Zellner 1986; Berger and Pericchi 2001; Marin and Robert 2007). Third, in very general work on testing, Dass and Berger (2003) showed that a right-Haar prior, in this context the invariance prior on τ, has optimal properties. Finally, this reparameterization is analogous to Sun and Speckman (2008), motivating the form of priors we can use.

Using the reparameterized adaptive IGMRF priors, we investigate possible objective priors on ξ_1 and ξ_2 for which the posterior is proper, i.e.,

$$\int\int\int\int\int [\boldsymbol{y} \mid \boldsymbol{z}, \tau]\,[\boldsymbol{z} \mid \xi_1, \tau, \boldsymbol{\gamma}]\,[\boldsymbol{\gamma} \mid \tau, \xi_2]\,[\tau]\,[\xi_1, \xi_2]\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\boldsymbol{\gamma}\,\mathrm{d}\tau\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 < \infty. \tag{13}$$

The following theorem gives sufficient conditions for the priors on (τ, ξ_1, ξ_2) to ensure (13).

Theorem 1 *Consider the non-parametric model* (3) *with prior distribution* $[\boldsymbol{z}, \boldsymbol{\gamma}, \tau \mid \xi_1, \xi_2]$ *given by* (11) *and* (12). *If* $[\xi_1, \xi_2]$ *is proper and* $E\xi_2^{-(n-p)/2} < \infty$ $(n > p)$, *then the joint posterior of* $(\boldsymbol{z}, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ *is proper.*

The proof is in Appendix A.1. The results on $[\tau]$ and $[\xi_1]$ in the theorem coincide with those in Sun and Speckman (2008). The strong condition on $[\xi_2]$ suggests that ξ_2 must be *a priori* bounded away from zero. This makes intuitive sense since one would never want the prior on $\boldsymbol{\gamma}$ to interpolate data in any sense. Notice that this theorem also applies to the non-equally spaced design because the proof does not require the equally spaced condition.

Remark 1 Although only IGMRF priors are considered in this paper, one may take other smooth processes, e.g., CAR models, to build this kind of two-layer adaptive prior. The strategy of the proofs in the appendix should also apply and yield similar theoretical results. We also extend this one-dimensional adaptive smoothing method to a two-dimensional spatial model in Yue and Speckman (2010).

2.4 Choice of hyperpriors for ξ_1 and ξ_2

From Sun and Speckman (2008), it's clear that the priors on ξ_1 and ξ_2 must be proper if the invariance prior (or any proper but diffuse approximation) is used for δ. In the spirit of an objective Bayesian analysis, we propose the following weakly informative priors. Following Liang et al. (2008), we suggest a Pareto prior for ξ_1, i.e.,

$[\xi_1|c] = c/(c+\xi_1)^2,\ \xi_1 \geq 0, c > 0$. The prior on ξ_2 is a little more difficult since the theorem requires negative moments. One solution is a proper inverse gamma prior, i.e., $[\xi_2|a,b] \propto \xi_2^{-(a+1)} e^{-b/\xi_2},\ \xi_2 > 0, a > 0, b > 0$, since this prior has all negative moments.

To complete the choice of hyperpriors, values for the hyperparameters a, b and c must be specified. Our strategy is based on the notion of equivalent degrees of freedom (e.d.f.), first used by White (2006). Since the priors on ξ_1 and ξ_2 must be proper, they must be subjective. The difficulty is in eliciting prior information for these hyperparameters. We believe e.d.f. is a useful way to elicit such information.

The trace of the smoother matrix is commonly used to define the degrees of freedom for a frequentist non-parametric regression method (Hastie et al. 2001). Motivated by a discussion on prior effective degrees of freedom in Hastie and Tibshirani (2000), we choose c for the prior on ξ_1 so that the median prior degrees of freedom is desirable. For the non-adaptive case ($\boldsymbol{\gamma} \equiv \mathbf{0}$), the smoother matrix is $\boldsymbol{S}_{\xi_1} = \left(\boldsymbol{I}_n + \xi_1 \boldsymbol{A}^{(p)}\right)^{-1}$. Since $\text{trace}(\boldsymbol{S}_{\xi_1})$ is a monotone function of ξ_1, the median of its distribution is $\text{trace}(\boldsymbol{S}_\kappa)$, where $\boldsymbol{S}_\kappa = \left(\boldsymbol{I}_n + \kappa \boldsymbol{A}^{(p)}\right)^{-1}$ and κ is the median of the prior on ξ_1. For the Pareto prior given above, the median is c. Therefore, we choose the value c that sets $\text{trace}(\boldsymbol{S}_c)$ equal to a desired prior degrees of freedom.

The situation for $[\xi_2]$ is more complicated. We first expand the conditional posterior of $\boldsymbol{\gamma}$ as in the Laplace approximation and then compute an equivalent smoother matrix $\left(\boldsymbol{W} + 2\xi_2 \boldsymbol{A}^{(q)}\right)^{-1} \boldsymbol{W}$, where $\boldsymbol{W}$ is a data-dependent diagonal matrix (see Appendix B.1). For a given shape parameter a, we solve for a scale parameter b giving desired degrees of freedom. We suggest $a = .5$ corresponding to Zellner–Siow priors (Zellner and Siow 1980).

The notion of "degrees of freedom" relates directly to the complexity of the model. For a parametric model, the degrees of freedom is exactly the number of parameters in the model. "Equivalent degrees of freedom" has exactly the same interpretation. Thus a researcher can use his or her knowledge of the problem to select appropriate prior distributions for ξ_1 and ξ_2. In Sect. 5, we provide more guidelines on the choice of the prior degrees of freedom.

In Fig. 1, we compare MCMC trace plots for estimation of a spatially inhomogeneous function taken from Di Matteo et al. (2001) (described in Example 2 in Sect. 5) under the prior of Lang et al. (2002) and the model here. For the Lang et al. prior, we used gamma(.0001, .0001) priors on δ and η. The trace plots of δ and η are presented in panels (a) and (b). The MCMC samples suggest two modes for the posterior, which could well be caused by the near impropriety of the posterior. On the contrary, the Markov chains of ξ_1 and ξ_2 with prior degrees of freedom equal 10 and 15, respectively, converge fast and mix well as shown in panels (c) and (d).

2.5 Adaptive IGMRF in additive models

In this section, the adaptive IGMRF is used as a prior in an additive model. The models not only are able to fit underlying inhomogeneous spatial patterns but can also include covariates and repeated measurements. We derive sufficient conditions to ensure proper posteriors.

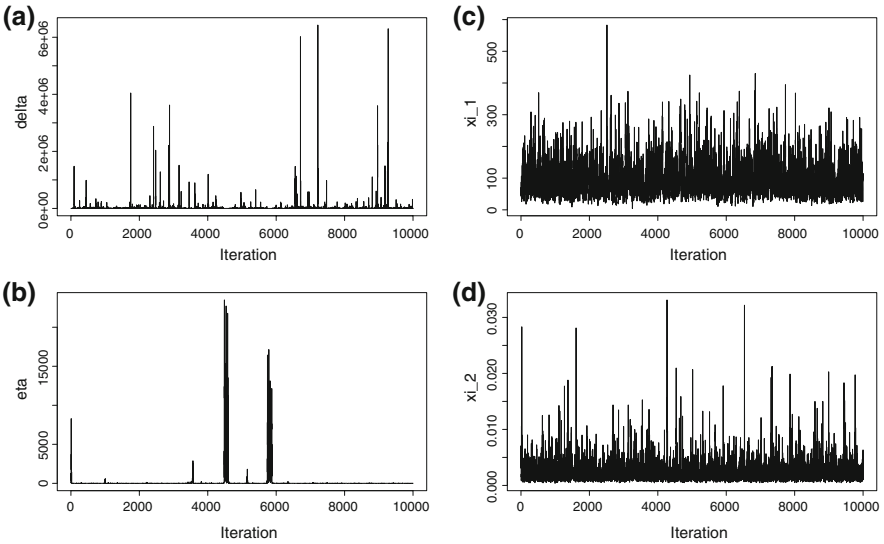

Fig. 1 Trace plots of MCMC samples of δ and η using the Lang et al. (2002) priors (**a**, **b**), and of ξ_1 and ξ_2 from Theorem 1 (**c**, **d**) for Example 2. The *plots* show one million iterations sampled every 100 times; the y-scale in **a** and **b** are truncated for clarity

Consider the additive model

$$y_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta} + \boldsymbol{x}'_{2i}\boldsymbol{z} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,\ \tau^{-1}), \tag{14}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)'$ is an $N \times 1$ vector of data, $\boldsymbol{X}_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1N})'$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2N})'$ are known design matrices with dimensions $N \times m$ and $N \times n$, $\boldsymbol{\beta}$ is an $m \times 1$ vector of fixed effects, $\boldsymbol{z}$ is an $n \times 1$ vector of nonlinear effects, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)'$ is an $N \times 1$ vector of random normally distributed errors with mean 0 and variance τ^{-1}. Typically, $\boldsymbol{X}_2$ is an incidence matrix indicating the location of each observation. Note that any location could have repeated, single or no observations. Model (14) can thus be represented as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{I}_N). \tag{15}$$

Obviously, (15) is a more general model than (3). The objective Bayesian hierarchical linear mixed model here has a constant prior on $\boldsymbol{\beta}$,

$$[\boldsymbol{\beta}] \propto 1, \tag{16}$$

and the adaptive IGMRF prior in (11) for $\boldsymbol{z}$. A similar approach has been suggested by many authors, e.g., Sun et al. (1999), who used conditional autoregressive priors on $\boldsymbol{z}$. The model proposed here can be viewed as a spatially adaptive extension of Sun et al.'s method.

Hyperpriors are required for τ, ξ_1 and ξ_2 to complete the hierarchical structure. We consider the following cases.

Case 1 There are repeated observations for at least one location.
Case 2 There is at most one observation for each location and at least one location has a missing observation.

With the invariance prior on τ, the following theorem provides sufficient conditions on the joint prior of (ξ_1, ξ_2) to ensure a proper posterior distribution in each case.

Theorem 2 *Consider the additive model* (15) *with prior distribution* $[z, \boldsymbol{\gamma}, \tau \mid \xi_1, \xi_2]$ *given by* (11), (12), *and* (16). *Let* $\boldsymbol{N}_p$ *be an* $n \times p$ *matrix whose columns span the null space of* $\boldsymbol{A}^{(p)}$, *and let* $C(\boldsymbol{X}_1)$ *and* $C(\boldsymbol{X}_2\boldsymbol{N}_p)$ *denote the column spaces of* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2\boldsymbol{N}_p$, *respectively. Assume that* $rank(\boldsymbol{X}_1) = m$ *and* $C(\boldsymbol{X}_2\boldsymbol{N}_p) \cap C(\boldsymbol{X}_1) = \emptyset$. *Then the joint posterior of* $(\boldsymbol{\beta}, z, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ *is proper if the following conditions hold for the two cases:*

Case 1 (*a*) $[\xi_1, \xi_2]$ *is proper;* (*b*) $N \geq m + p + 1$;
Case 2 (*a*) $[\xi_1, \xi_2]$ *is proper;* (*b*) $N \geq m + p$; (*c*) $E(\xi_1\xi_2)^{-(N-m-p)/2} < \infty$.

The proof is in Appendix A.2. Assumption (c) in Case 2 indicates that both ξ_1 and ξ_2 have priors with finite negative moments. Following the priors from Theorem 1, we suggest taking independent Pareto priors for ξ_1 and ξ_2 in Case 1 and proper inverse gamma priors in Case 2. Sun et al. (1999, 2001) investigated the posteriors for similar linear mixed models and motivated this work. Again, this theorem also applies to non-equally spaced designs.

3 Bayesian adaptive P-splines

In this section, we consider a BAPS model based on work by a number of authors including Lang and Brezger (2004), Baladandayuthapani et al. (2005), Crainiceanu et al. (2007), and we adapt the theoretical results from the adaptive IGMRF to BAPS.

3.1 Regression P-splines

P-splines (Eilers and Marx 1996; Ruppert et al. 2003) have become popular for non-parametric regression due to the use of a relatively small number of basis functions. In addition, P-splines can be viewed as linear mixed models and are easy to compute with widely available statistical software (e.g., Ngo and Wand 2004).

Consider again the non-parametric regression model (1) with $\varepsilon_i \overset{iid}{\sim} N(0,\ \tau^{-1})$. As defined by Ruppert et al. (2003), the P-spline method for estimating f is to use a large regression spline model,

$$f(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{j=1}^{m_t} b_j (x_i - t_j)_+^p, \tag{17}$$

where $p \geq 0$ is an integer, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ and $\boldsymbol{b} = (b_1, \ldots, b_{m_t})'$ are two vectors of regression coefficients, $t_1 < \cdots < t_{m_t}$ are fixed knots, and a_+^p denotes $a^p I(a \geq 0)$. Denote by $\boldsymbol{X}$ the $n \times (p+1)$ matrix with ith row equal to $(1, x_i, \ldots, x_i^p)$ and by $\boldsymbol{Z}$ the $n \times m_t$ matrix with ith row equal to $((x_i - t_1)_+^p, \ldots, (x_i - t_{m_t})_+^p)$, and let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{b}')'$, $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Z})$, and $\boldsymbol{y} = (y_1, \ldots, y_n)'$. We define $\hat{\boldsymbol{\theta}}$ as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{T}\boldsymbol{\theta}) + \lambda \boldsymbol{b}'\boldsymbol{b} \right]. \tag{18}$$

Then $\hat{\boldsymbol{f}} = \boldsymbol{T}\hat{\boldsymbol{\theta}}$ is called a pth order P-spline fit. To avoid overfitting, the penalty term $\boldsymbol{b}'\boldsymbol{b}$ in (18) shrinks the b_j towards zero by an amount controlled by the smoothing parameter λ.

The model proposed by Eilers and Marx (1996) and treated, for example by Lang and Brezger (2004), is a variation. The truncated power basis in (17) is replaced by a B-spline basis $B_j(x_i)$ with the same set of knots, and the quadratic penalty in (18) is replaced by $\boldsymbol{b}'\boldsymbol{A}^{(q)}\boldsymbol{b}$ for $q = 1$ or 2. We expect that results similar to those obtained in Theorem 3 below are possible for this version as well. However, for simplicity, we confine attention to the version of Ruppert et al. (2003).

The minimizing criterion in (18) reveals that P-splines have a natural linear mixed model representation given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{I}_n),$$

where $\boldsymbol{\beta}$ is a vector for fixed effects, $\boldsymbol{b} \sim N(\boldsymbol{0}, \delta^{-1}\boldsymbol{I}_{m_t})$ is a vector of random effects, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be viewed as design matrices for fixed and random effects, respectively. Inspired by this fact, Bayesian P-splines (Ruppert et al. 2003; Lang and Brezger 2004) use a stochastic process model as a prior for the regression function,

$$\begin{aligned} &(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}, \tau) \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \tau^{-1}\boldsymbol{I}_n), \\ &[\boldsymbol{\beta}] \propto \boldsymbol{1}, \quad (\boldsymbol{b} \mid \delta) \sim N(\boldsymbol{0}, \delta^{-1}\boldsymbol{I}_{m_t}). \end{aligned} \tag{19}$$

It is easy to see that the posterior mean in (19) given τ and δ provides a Bayesian version of P-splines as defined in (18), and the smoothness is controlled by the ratio δ/τ. Note that for fully Bayesian inference we also need to specify priors on τ and δ.

3.2 Spatially adaptive Bayesian P-splines

As with smoothing splines, P-splines with a single smoothing parameter are not optimal for estimating spatially adaptive functions (Wand 2000). There is a growing literature (e.g., Ruppert and Carroll 2000; Lang and Brezger 2004; Baladandayuthapani et al. 2005; Crainiceanu et al. 2007) on extending P-spline methodology to spatially adaptive smoothing parameters. Following Baladandayuthapani et al. (2005), we modify the homoscedastic prior on the random effects b_j in (19) using spatially adaptive

precisions δ_j,

$$(b_j \mid \delta_j) \overset{ind}{\sim} N(0, \delta_j^{-1}), \quad j = 1, \ldots, m_t.$$

Baladandayuthapani et al. directly modeled $\log(\delta_j)$ as a P-spline of degree $q (q < p)$,

$$\log(\delta_j) = \beta_{\gamma 0} + \beta_{\gamma 1} t_j + \cdots + \beta_{\gamma q} t_j^q + \sum_{k=1}^{m_s} b_{\gamma k} (t_j - s_k)_+^q, \tag{20}$$

where $\boldsymbol{\beta}_\gamma = (\beta_{\gamma 0}, \ldots, \beta_{\gamma q})'$, $\boldsymbol{b}_\gamma = (b_{\gamma 1}, \ldots, b_{\gamma m_s})'$, and $s_1 < \cdots < s_{m_s}$ are fixed knots. Then they used a diffuse inverse gamma prior on the error variance τ^{-1} and a normal prior with a large variance on the fixed effects $\boldsymbol{\beta}_\gamma$. In the limit, their prior is equivalent to using improper invariance priors for τ and $\delta = \exp(\boldsymbol{\beta}_{\gamma 0})$, i.e., $[\tau, \delta] \propto 1/(\tau\delta)$. However, Sun et al. (2001) proved that even ignoring the non-constant terms in (20), this limiting prior with $\delta_j \equiv \exp(\beta_{\gamma 0})$ for all j will lead to an improper posterior in the general linear mixed model. Clearly, the limiting flat priors on $\beta_{\gamma 1}, \ldots, \beta_{\gamma q}$ are also not justified. Therefore, MCMC computation is problematic using the Baladandayuthapani et al. prior. Since the limiting posterior is improper, one would expect that the posterior for proper but diffuse priors depends heavily on the priors.

We propose a modification to the Baladandayuthapani et al. prior and investigate propriety of the corresponding posteriors with improper priors. As in Sect. 2.2, define $\delta_j = \delta \exp(\gamma_j)$ and let γ_j be a P-spline of degree $q = 0$ or 1 with $\log \delta$ replacing $\beta_{\gamma 0}$ in (20). Suppose first that $q = 0$. Then

$$\gamma_j = \sum_{k=1}^{m_s} b_{\gamma k} I(t_j \geq s_k), \quad j = 1, \ldots, m_t. \tag{21}$$

Baladandayuthapani et al. used a normal prior with mean zero and variance η^{-1} on the random effects $b_{\gamma k}$,

$$(b_{\gamma k} \mid \eta) \overset{iid}{\sim} N(0, \eta^{-1}), \quad k = 1, \ldots, m_s.$$

When $q = 1$, the P-spline we use for $\boldsymbol{\gamma}$ is

$$\gamma_j = \sum_{k=0}^{m_s} b_{\gamma k} (t_j - s_k)_+, \quad j = 1, \ldots, m_t. \tag{22}$$

Model (22) is obtained by replacing the fixed effect linear term $\beta_{\gamma 1}$ with a new knot $s_0 = 0$ and a new random effect $b_{\gamma 0}$. Following the $q = 0$ case, we take independent normal priors on the $b_{\gamma k}$. In our experience, the cases $q = 0$ or 1 suffice for most applications. For $q \geq 2$, it is more difficult to specify a sensible proper prior on $\boldsymbol{\gamma}$ because of the need to specify proper priors on the low order fixed effects such as $\beta_{\gamma 1}$.

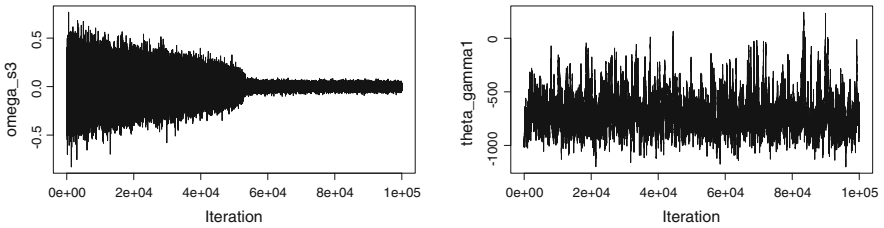

Fig. 2 Trace plots of MCMC samples for the spatially adaptive P-spline of Baladandayuthapani et al. (2005) (*left panel*) and BAPS (*right panel*) for the Doppler function

Following the BASS model, an invariance prior is taken on τ as in (12). With reparameterizations $\xi_1 = \delta/\tau$ and $\xi_2 = \eta/\delta$, the BAPS model that we propose with $q = 0$ or 1 has the matrix form

$$\begin{aligned} (\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}, \tau) &\sim N(\boldsymbol{X\beta} + \boldsymbol{Zb}, \tau^{-1}\boldsymbol{I}_n), \\ [\tau] \propto \frac{1}{\tau}, \quad [\boldsymbol{\beta}] \propto \mathbf{1}, \quad (\boldsymbol{b} \mid \boldsymbol{\gamma}, \tau, \xi_1) &\sim N(\mathbf{0}, (\tau\xi_1)^{-1}\boldsymbol{D}_\gamma^{-1}), \\ \boldsymbol{\gamma} = \boldsymbol{Z}_\gamma \boldsymbol{b}_\gamma, \quad (\boldsymbol{b}_\gamma \mid \tau, \xi_1, \xi_2) &\sim N(\mathbf{0}, (\tau\xi_1\xi_2)^{-1}\boldsymbol{I}_{m_s+q}), \end{aligned} \tag{23}$$

where $\boldsymbol{D}_\gamma = \text{diag}(e^{\gamma_1}, \ldots, e^{\gamma_{m_t}})$ and $\boldsymbol{Z}_\gamma$ is the design matrix in (21) and (22). The parameters τ, ξ_1 and ξ_2 in (23) need hyperpriors for fully Bayesian inference. Again we derive sufficient conditions on those priors to ensure the propriety of posteriors for BAPS.

Theorem 3 *For the BAPS model in* (23)*, assume* $[\xi_1, \xi_2]$ *is proper. Then the joint posterior of* $(\boldsymbol{\beta}, \boldsymbol{b}, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ *is proper.*

See the proof in Appendix A.3. Note that the theorem is valid for any choice of P-spline basis. In practice, we suggest choosing hyperpriors and hyperparameters for ξ_1 and ξ_2 in the same way as for BASS additive models, with knots selected as in Baladandayuthapani et al. (2005). Note that we can also use BAPS in additive models, where important covariates and repeated observations can be taken into account, and it is easy to generalize the theorem accordingly.

In Fig. 2 (left panel), we present trace plots of an MCMC simulation for the Doppler function example used in Baladandayuthapani et al. (2005) to illustrate the problem. The estimates are from a run of two million iterations sampled every 200 trials. The plots show that it takes very many iterations (over one million) for the Markov chain to converge, and the mixing is also very poor for some variables. Baladandayuthapani et al. suggested only 10,000 MCMC iterations with a burn-in period of 1,000. This is clearly invalid. We also present the trace plots from the BAPS analysis in the right panel of Fig. 2, in which the MCMC chains mix well and converge quickly.

4 Bayesian computation

We give explicit full conditional distributions of BASS and BAPS. The Gibbs sampler is used to estimate posterior distributions. We show how block sampling and orthogonal transformation can be used to speed convergence of the chains.

4.1 Gibbs sampler for BASS

Letting $S_1(\boldsymbol{z}) = \boldsymbol{z}'\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{z}$ and $S_2(\boldsymbol{\gamma}) = \boldsymbol{\gamma}'\boldsymbol{A}^{(q)}\boldsymbol{\gamma}$, the full conditionals of the posterior of BASS have the following properties.

(C1) $(\boldsymbol{z} \mid \cdot) \sim N_n(\boldsymbol{\mu}_z, \boldsymbol{Q}_z^{-1})$, where $\boldsymbol{\mu}_z = \tau\boldsymbol{Q}_z^{-1}\boldsymbol{y}$ and $\boldsymbol{Q}_z = \tau\boldsymbol{I}_n + \tau\xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}$.
(C2) $(\tau \mid \cdot) \sim \text{Gamma}\left(\frac{1}{2}(3n-2p-q), \frac{1}{2}\left[\|\boldsymbol{y}-\boldsymbol{z}\|^2 + \xi_1 S_1(\boldsymbol{z}) + \xi_1\xi_2 S_2(\boldsymbol{\gamma})\right]\right)$.
(C3) $(\xi_1 \mid \cdot) \sim \text{Gamma}\left(\frac{1}{2}(2n-2p-q)+1,\ \frac{1}{2}\tau[S_1(\boldsymbol{z}) + \xi_2 S_2(\boldsymbol{\gamma})] + \theta\right)$.
(C4) $(\theta \mid \cdot) \sim \text{Gamma}(2,\ \xi_1 + c)$, where θ is an auxiliary variable.
(C5) $[\gamma_\ell \mid \cdot] \propto \exp\left[-\frac{1}{2}\tau\xi_1 e^{\gamma_\ell}\left(\sum_{j=0}^{p}\binom{p}{j}(-1)^j z_{\ell-j}\right)^2 - \frac{1}{2}\tau\xi_1\xi_2\sum_{k=0}^{q}\left(\sum_{j=0}^{q}\binom{q}{j}(-1)^j\gamma_{\ell+k-j}\right)^2\right]$ subject to $\boldsymbol{1}'\boldsymbol{\gamma} = 0$.
(C6) $[\xi_2 \mid \cdot] \propto \xi_2^{\frac{1}{2}(n-p-q)-a-1}\exp\left[-\frac{1}{2}\tau\xi_1\xi_2 S_2(\boldsymbol{\gamma}) - b/\xi_2\right]$.

Note that $\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}$ in (C1) is a band matrix with bandwidth $p+1$. The entire vector $\boldsymbol{z}$ can be sampled efficiently using a band Cholesky decomposition. The full conditionals for γ_ℓ and ξ_2 are both log concave, so either adaptive rejection Metropolis sampling (ARMS) (Gilks and Wild 1992; Gilks et al. 1995) or the Metropolis–Hastings (MH) method can be used to sample them. To enforce the linear restriction on γ_ℓ, one can either re-center after each MCMC cycle by subtracting mean $\bar{\gamma}$ from each γ_ℓ (Besag et al. 1995) or employ the orthogonal transformation strategy described below.

4.2 Orthogonal transformation and dimension reduction

If we sample the $\boldsymbol{\gamma}_\ell$ from the full conditional in (C5) by ARMS, the Markov chains tend to converge rather slowly because of the correlation and autocorrelation between the IGMRF and its hyperparameters (Rue and Held 2005). To improve mixing and speed computation, we propose an orthogonal transformation combined with a dimension reduction technique as follows.

First, compute the spectral decomposition $\boldsymbol{A}^{(q)} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, where $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-p}]$ is orthogonal, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{n-p})$, $\lambda_1 = 0 \le \lambda_2 \le \cdots \le \lambda_{n-p}$, and $\boldsymbol{A}^{(q)}$ is the precision matrix in the prior density of $\boldsymbol{\gamma}$. Note that the constant vector $\boldsymbol{1} = (1, \ldots, 1)'$ is always in the null space of $\boldsymbol{A}^{(q)}$ for any $q \ge 1$. Without loss of generality, let $\boldsymbol{p}_1 = \boldsymbol{1}$. By an orthogonal change of variable, $\boldsymbol{\phi} = \boldsymbol{P}'\boldsymbol{\gamma}$ has prior

$$[\boldsymbol{\phi} \mid \tau, \xi_1, \xi_2] \propto (\tau\xi_1\xi_2)^{\frac{1}{2}(n-p-1)}\exp\left(-\frac{\tau\xi_1\xi_2}{2}\sum_{k=2}^{n-p}\lambda_k\phi_k^2\right),$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{n-p})'$, and we enforce $\phi_1 = 0$ to ensure the identifiability. Following the principal components dimension reduction for splines introduced by Van Der Linde (2003), we reduce the dimension of $\boldsymbol{\phi}$ by using the first $m \ll n$ basis vectors of $\boldsymbol{P}$.

Again, consider $p = 2$ and $q = 1$. With $\gamma_k = \boldsymbol{e}'_{k-2} \sum_{j=2}^{m} \boldsymbol{p}_j \phi_j$ for $3 \le k \le n$, where $\boldsymbol{e}_\ell = (0, \ldots, 0, 1, 0, \ldots, 0)'_{(n-2)\times 1}$ has "1" at the ℓth position, the log likelihood of the full conditional of ϕ_ℓ for $2 \le \ell \le m$ is (up to additive constant)

$$h(\phi_\ell) = -\frac{\tau\xi_1}{2}\left[\sum_{k=3}^{n} \exp\left(\boldsymbol{e}'_{k-2}\boldsymbol{p}_\ell\phi_\ell + \boldsymbol{e}'_{k-2}\sum_{j=2, j\neq\ell}^{m}\boldsymbol{p}_j\phi_j\right) \times (z_k - 2z_{k-1} + z_{k-2})^2\right] - \frac{\tau\xi_1\xi_2}{2}\lambda_\ell\phi_\ell^2.$$

Since $h(\phi_\ell)$ is also log concave, ARMS can be used again. The full conditionals of other parameters change accordingly, and the sampling scheme remains. The orthogonal transformation improves the mixing of the chains dramatically, and the simulation also becomes much faster since a small m, say $m = 10$ or $m = 20$, may be used in practice. Note that the dimension m is comparable to m_s in BAPS.

An alternative to dimension reduction is block sampling for the full $\boldsymbol{\gamma}$ vector. We have used the method of Lang et al. (2002) to obtain comparable results.

4.3 Gibbs sampler for BAPS

For simplicity, assume $\boldsymbol{\beta}$ in (23) has a normal prior with large variance, say $\sigma_\beta^2 = 10^6$, which is essentially equivalent to the constant prior. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{b}')'$ and choose $[\xi_1|c] = c/(c+\xi_1)^2$ and $[\xi_2|a,b] \propto \xi_2^{-(a+1)}e^{-b/\xi_2}$. The full conditionals of the BAPS model are listed below.

(C7) $(\boldsymbol{\theta} \mid \cdot) \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{Q}_\theta^{-1})$, where $\boldsymbol{\mu}_\theta = \tau\boldsymbol{Q}_\theta^{-1}\boldsymbol{T}'\boldsymbol{y}$, $\boldsymbol{Q}_\theta = \tau\boldsymbol{T}'\boldsymbol{T} + \boldsymbol{\Lambda}_y$, and $\boldsymbol{\Lambda}_y = \mathrm{diag}(1/\sigma_\beta^2, \ldots, 1/\sigma_\beta^2, \tau\xi_1 e^{\gamma_1}, \ldots, \tau\xi_1 e^{\gamma_{m_t}})$ is the prior precision on $\boldsymbol{\theta}$.

(C8) $(\tau \mid \cdot) \sim \mathrm{Gamma}\big(\frac{1}{2}(n + m_t + m_s + q), \frac{1}{2}(\|\boldsymbol{y} - \boldsymbol{T}\boldsymbol{\theta}\|^2 + \xi_1\boldsymbol{b}'\boldsymbol{D}_\gamma\boldsymbol{b} + \xi_1\xi_2\boldsymbol{b}'_\gamma\boldsymbol{b}_\gamma)\big)$.

(C9) $(\xi_1 \mid \cdot) \sim \mathrm{Gamma}\left(\frac{1}{2}(m_t + m_s + q) + 1, \frac{1}{2}\tau\boldsymbol{b}'\boldsymbol{D}_\gamma\boldsymbol{b} + \frac{1}{2}\tau\xi_2\boldsymbol{b}'_\gamma\boldsymbol{b}_\gamma + \rho_1\right)$.

(C10) $(\xi_2 \mid \cdot) \sim \mathrm{Gamma}\left(\frac{1}{2}(m_s + q) + 1, \frac{1}{2}\tau\xi_1\boldsymbol{b}'_\gamma\boldsymbol{b}_\gamma + \rho_2\right)$.

(C11) $(\rho_i \mid \cdot) \sim \mathrm{Gamma}(2,\ \xi_i + c_i)$ for $i = 1, 2$.

(C12) $[\boldsymbol{b}_\gamma \mid \cdot] \propto |\boldsymbol{D}_\gamma|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\tau\xi_1\boldsymbol{b}'\boldsymbol{D}_\gamma\boldsymbol{b} - \frac{1}{2}\tau\xi_1\xi_2\boldsymbol{b}'_\gamma\boldsymbol{b}_\gamma\right)$.

To speed convergence of the Markov chains, we employ a block-move MH algorithm (Carter and Kohn 1996; Knorr-Held and Richardson 2003; Lang et al. 2002) to sample $\boldsymbol{b}_\gamma$. The basic idea is that we first split $\boldsymbol{b}_\gamma$ into several blocks and then update each block using a MH sampler that has the prior for $\boldsymbol{b}_\gamma$ as the proposal distribution. The details can be found in Yue and Speckman (2010).

5 Examples

In this section, we present a simulation study to evaluate the performance of BASS and BAPS. Based on the examples in Di Matteo et al. (2001), we first compare BASS visually to its non-adaptive version, the BSS under prior (6). We then compare BASS and BAPS quantitatively with BARS (Di Matteo et al. 2001) and the non-stationary GP model of Paciorek and Schervish (2004) in terms of standardized mean squared error (MSE).

The data were generated with the three mean functions used in Di Matteo et al. (2001): a smoothly varying function, a spatially inhomogeneous function, and a function with a sharp jump. According to the strategy described in Sect. 2.3, we choose values of a, b and c to have desirable prior e.d.f. For the BASS prior and $p = 2$ ($p = 3$), we chose $c = 254(3, 926)$ for the first two examples and $c = 3,914(240, 848)$ for the third one. These produce a median prior e.d.f. of ten corresponding to ξ_1 in all situations. For ξ_2 we chose prior e.d.f. to be 5, 15, and 60 in Examples 1–3, respectively. Letting $a = .5$, the corresponding values of b are .003 (.0017), .0009 (.0004) and .0001 (.00008) for $p=2$ ($p=3$).

Example 1 We first describe a simulated smooth example. The test function is a natural spline with three knots at (.2, .6, .7) and coefficients $\beta = (20, 4, 6, 11, 6)'$ evaluated at $n = 101$ equally spaced points on [0,1] with Gaussian noise and standard deviation $\tau^{-1/2} = 0.9$. The true function is plotted as a dashed line in Fig. 3. Panels (a) and (b), respectively, show the fits from BASS and BSS when $p = 2$, from which little difference can be found. This behavior can be explained by panel (c), where the estimates of γ_k and 95% credible intervals are plotted. There is no evidence that adaptive smoothing is needed in this case. Similar simulation results are given for $p = 3$ in panels (d)–(f). Note that in this example there appears to be little loss in efficiency in using the adaptive prior when it is not needed.

Example 2 The second simulated data example is a spatially inhomogeneous function,

$$g(t) = \sin(t) + 2\exp(-30t^2), \quad t \in [-2, 2],$$

again evaluated at $n = 101$ regularly spaced points. The standard deviation of the noise is $\tau^{-1/2} = .3$. The "true" data are plotted in Fig. 4. Clearly, the smoothness of the process varies significantly due to a sharp peak in the middle. The results obtained from BSS with $p = 2$ are displayed in panel (b). The estimate shows typical fixed bandwidth behavior with serious undersmoothing in the flat portions of the graph on the left and right. A single smoothing parameter is thus clearly not adequate to accommodate the varying smoothness of the structure of the data. Panel (a) shows the estimate from BASS with $p = 2$, which is clearly able to adapt to the changes of smoothness with noticeably smoother fits on both sides and a much better fit of the peak. To see how the adaptive scheme works, panel (c) displays the estimated γ_k. The precision is high on the left and the right where the function is flat and appropriately low for the peak. From panels (d)–(f), the $p = 3$ prior appears to have similar spatially adaptive behavior as the $p = 2$ case. However, due to the higher autocorrelation of the

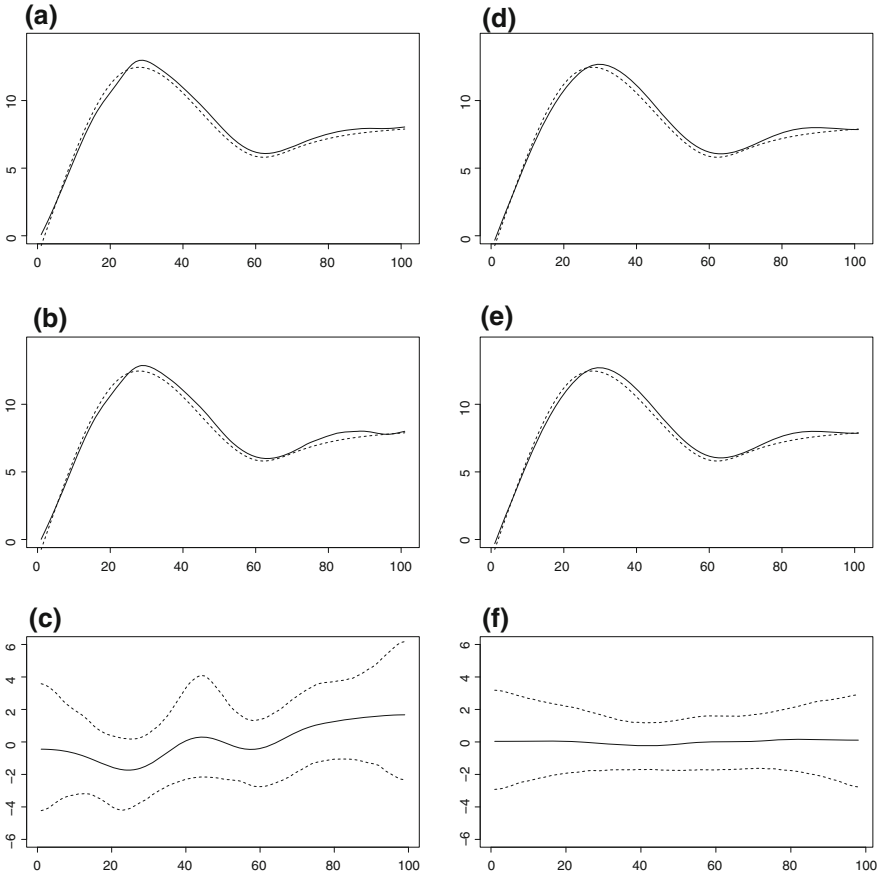

Fig. 3 Example 1: **a** BASS fit ($p=2$); **b** BSS fit ($p=2$); **c** posterior mean of γ_k and 95% credible intervals ($p=2$); **d** BASS fit ($p=3$); **e** BSS ($p=3$); **f** posterior mean of γ_k and 95% credible intervals ($p=3$) (*dotted line* true function, *solid line* fit)

$p=3$ prior, some oversmoothing is apparent. By comparing panels (a)–(d), $p=3$ has a better fit for both flat parts than $p=2$ but does not catch the peak as well as $p=2$ does. This is also demonstrated clearly by panels (c) and (f). For this case, twenty basis vectors ($m=20$) are adequate. Increasing m had no appreciable effect on the posterior means for $\boldsymbol{\gamma}$ or $\boldsymbol{z}$.

Example 3 Finally, we consider another natural spline example but with a discontinuity. The internal knots are located at (.4, .4, .4, .4, .7) and the coefficients are $(2, -5, 5, 2, -3, -1, 2)$. We generated $n = 201$ data points regularly on [0,1] with zero-mean Gaussian noise added and $\tau^{-1/2} = .55$. The true function is given as a dashed line in Fig. 5. A jump can be seen around the 80th point and the rest of the curve is smooth. As in the previous example, it appears both visually and from the behavior of the estimated γ_k given in panel (c) that BASS with $p = 2$ in panel (a) is capable of capturing the features of the data more precisely than BSS with $p = 2$,

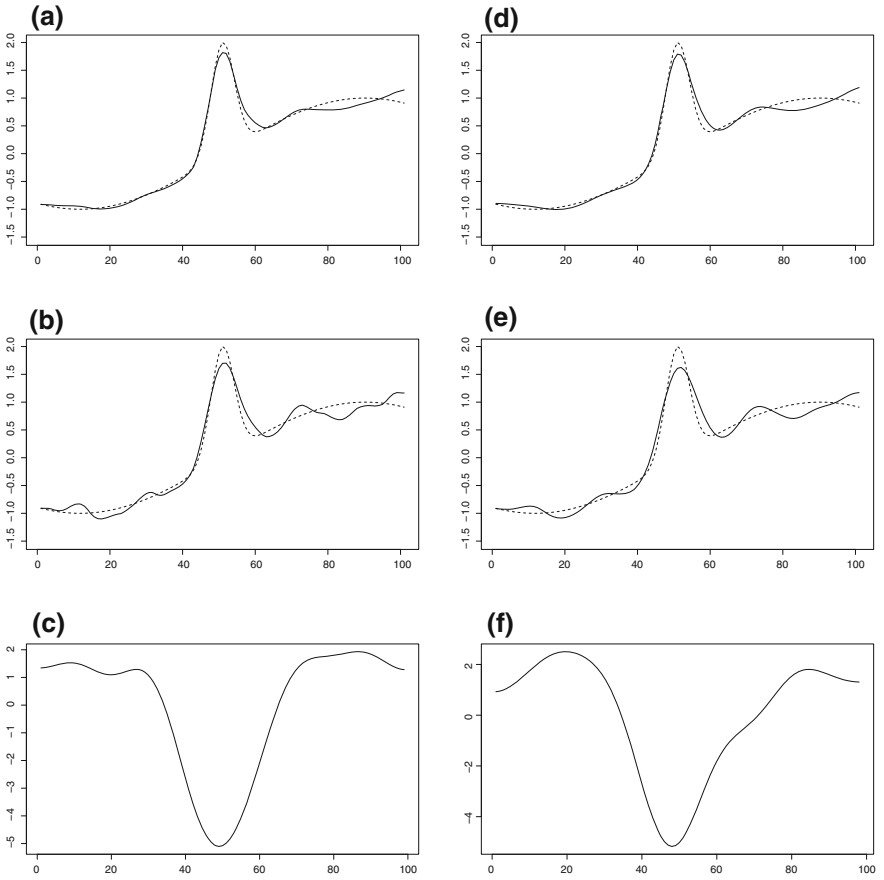

Fig. 4 Example 2: **a** BASS fit ($p=2$); **b** BSS fit ($p=2$); **c** posterior mean of $\gamma_k(p=2)$; **d** BASS ($p=3$); **e** BSS fit ($p=3$); **f** posterior mean of $\gamma_k(p=3)$ (*dotted line* true function, *solid line* fit)

which clearly undersmooths the flat regions and oversmooths the sharp jump as shown in panel (b). The result of the comparison between $p = 3$ and $p = 2$ is similar to Example 2. The prior with $p = 3$ tends to oversmooth the data. For this extreme case, we need more orthogonal basis elements from Sect. 4.2 than in Example 2, for example $m = 60$.

We also performed a small simulation study to compare BASS, BAPS, BSS, BARS and the non-stationary GP model. We generated 50 sets of noisy data and compared the above models using the means, averaged over the 50 sets, of the standardized MSE, $\sum_k (\hat{f}_k - f_k)^2 / \sum_k (f_k - \bar{f})^2$, where $\hat{f}_k$ is the posterior mean at x_k, and $\bar{f}$ is the mean of true values. MSE and its 95% confidence intervals are reported in Table 1 [the results for BARS and non-stationary GP are from Paciorek and Schervish (2004)]. BASS ($p = 2, 3$) outperforms BSS in the last two examples and even performs well in the first example where BSS is appropriate. It is hard to compare BASS to BARS and the non-stationary GP for the first two examples: there is much overlap in the

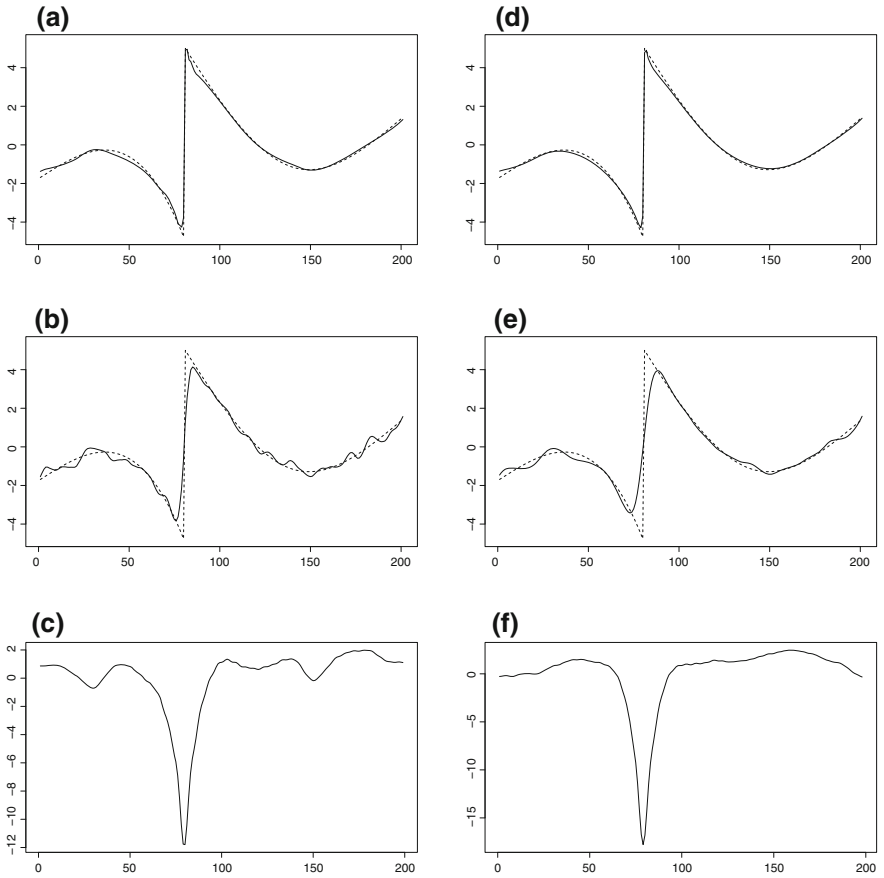

Fig. 5 Example 3: **a** BASS fit ($p = 2$); **b** BSS fit ($p = 2$); **c** posterior mean of $\gamma_k(p = 2)$; **d** BASS fit ($p = 3$); **e** BSS fit ($p = 3$); **f** posterior mean of $\gamma_k(p = 3)$ (*dotted line* true function, *solid line* fit)

confidence intervals and the number of simulations is small (only 50). However, BASS does surprisingly well (outperforming the non-stationary GP and BAPS) in estimating the jump function of Example 3. BAPS is much better than the other methods in Example 1, performs equally well as the other adaptive models in Example 2, and does a competitive job in Example 3: not as good as BASS and BARS but outperforming the non-stationary GP.

As requested by a referee, the estimates using BASS with $p = 2$ and 3 for all three examples corresponding to the 10th percentile, 50th percentile and 90th percentile MSE are also plotted in Figs. 6 and 7. The figures show consistent results with MSE: BASS with $p = 2$ performs worse, better and about the same as BASS with $p = 3$ in Examples 1–3, respectively. The figures also show small variability of the estimates based on the BASS model.

Finally, we did a small sensitivity study on the priors $[\xi_1 \mid c]$ and $[\xi_2 \mid a, b]$. In our experience, the adaptive fit is quite robust to the choice of prior for ξ_1 but somewhat

Table 1 Simulation study

Method	Function 1	Function 2	Function 3
BSS ($p = 2$)	.0100 (.0090, .0110)	.022 (.020, .024)	.097 (.0940, .0990)
BSS ($p = 3$)	.0088 (.0077, .0099)	.027 (.025, .029)	.130 (.1280, .1340)
BASS ($p = 2$)	.0097 (.0086, .0107)	.011 (.010, .013)	.008 (.0075,.0094)
BASS ($p = 3$)	.0089 (.0078, .0099)	.013 (.012, .015)	.008 (.0073, .0086)
BAPS	.0055 (.0054, .0056)	.012 (.011, .013)	.017 (.0157, .0183)
BARS	.0081 (.0071, .0092)	.012 (.011, .013)	.005 (.0043, .0056)
Non-stat. GP	.0083 (.0073, .0093)	.015 (.013, .016)	.026 (.0210, .0300)

Mean and 95% confidence interval for standardized MSE based on 50 samples obtained for the three test functions described in Sect. 5

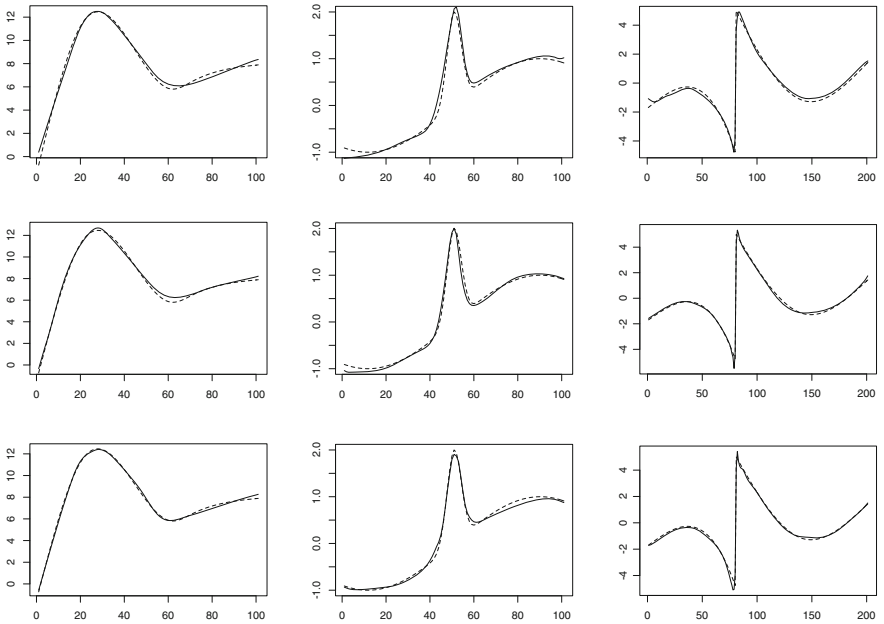

Fig. 6 *Top panels* plot the BASS estimates ($p = 2$) corresponding to the 10th worst percentile MSE for three simulated examples. *Middle* and *bottom panels* are similar plots corresponding to the 50th percentile MSE and 10th best percentile MSE, respectively. In all cases the true function is given by the *dotted line* and the estimate is given by the *solid line*

more sensitive to the choice for ξ_2. We experimented with various prior e.d.f. choices between 5 and 50 for ξ_2 when using prior e.d.f. $\approx$10 for ξ_1 in Example 2 and plotted the posterior densities of ξ_2 under prior e.d.f. $\approx$5 and 50 in Fig. 8 (left). As can be seen, the two posteriors obviously differ in shapes and posterior means. It is interesting that the overall fit, however, is much less affected by the choice of the prior e.d.f. as shown in the right panel of Fig. 8. In conclusion, the performance of BASS seems to be quite robust to the choice of hyperpriors if $[\xi_2]$ is chosen to have reasonable

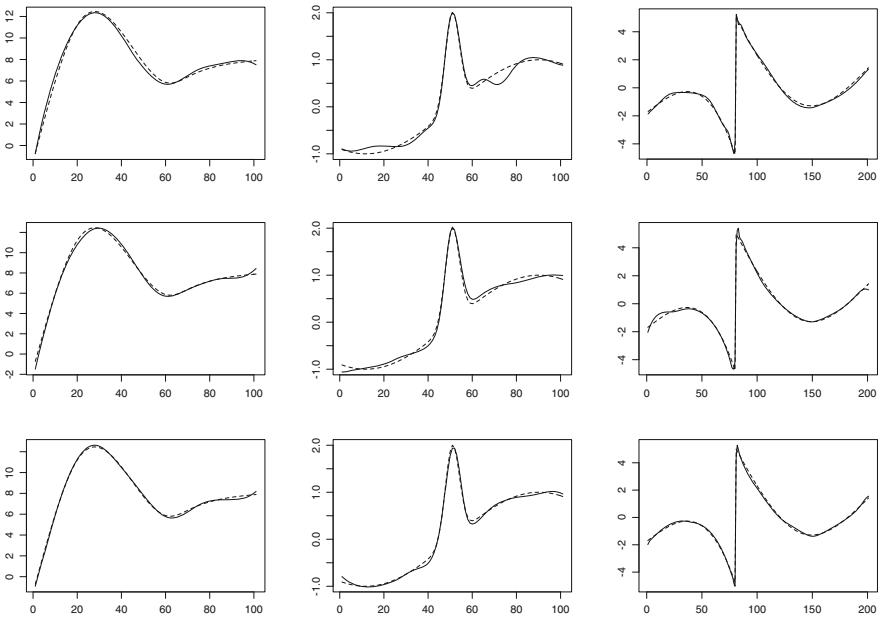

Fig. 7 *Top panels* plot the BASS estimates ($p = 3$) corresponding to the 10th worst percentile MSE for three simulated examples. *Middle* and *bottom panels* are similar plots corresponding to the 50th percentile MSE and 10th best percentile MSE, respectively. In all cases the true function is given by the *dotted line* and the estimate is given by the *solid line*

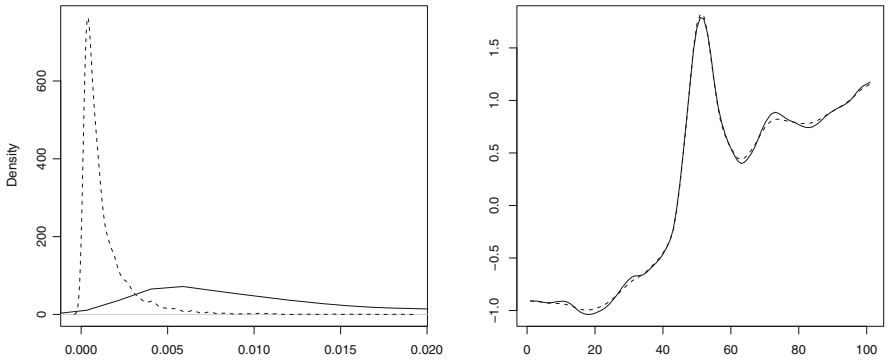

Fig. 8 Posterior densities of ξ_2 (*left*) and adaptive fits (*right*) under prior e.d.f. ≈5 (*solid line*) and prior e.d.f. ≈50 (*dashed line*)

prior e.d.f. for the approximate smoother on $\boldsymbol{\gamma}$. The value of prior e.d.f. can reflect the sample size n and the assumed complexity of the underlying process. In our simulated examples, we have had success using prior e.d.f. ≈10 with $n = 101$ for moderately varying function in Example 2 and prior e.d.f. ≈60 with $n = 201$ for the highly varying function in Example 3.

6 Summary and discussion

In this paper, we examine two spline-based Bayesian adaptive smoothing methods, BASS and BAPS. Those models allow spatial adaptivity in the spline penalty by using locally varying precisions δ_k, whose logarithm is modeled as another spline function. A reparameterization on the precision components is proposed, and sufficient conditions for the propriety of posterior distributions are derived on the proper priors for smoothing parameters ξ_1 and ξ_2 when the objective invariance prior is used for the error variance. We also suggest using the trace of appropriate smoother matrices to choose mildly informative priors on ξ_1 and ξ_2.

Based on both theoretical and empirical evidence, we conclude that one should be careful when using diffuse proper priors that are improper in the limit. Those priors may cause slow convergence and poor mixing in MCMC. Therefore, the propriety of posteriors should be rigorously checked before doing Bayesian inference with diffuse priors.

Appendix A. Proofs

To begin, we need the following lemmas.

Lemma 1 *Suppose $\boldsymbol{A}$ is an $m \times m$ full rank matrix and $\boldsymbol{B}$ is an $m \times n$ matrix with $rank(\boldsymbol{B}) = m (m < n)$. Define $|\boldsymbol{C}|_+$ to be the product of the non-zero eigenvalues of a non-negative matrix $\boldsymbol{C}$. Then*

$$|\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}|_+ = |\boldsymbol{B}\boldsymbol{B}'\boldsymbol{A}|.$$

Proof Let $\boldsymbol{B}' = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}'$ be the singular value decomposition of $\boldsymbol{B}'$, where $\boldsymbol{P}$ is an $n \times m$ matrix, $\boldsymbol{Q}$ is an $m \times m$ matrix, $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_m$, and $\boldsymbol{\Lambda}^2$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{B}\boldsymbol{B}'$. Then

$$\begin{aligned} |\boldsymbol{B}\boldsymbol{B}'\boldsymbol{A}| &= |\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}'\boldsymbol{A}| \\ &= |\boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}'\boldsymbol{A}| \\ &= |\boldsymbol{\Lambda}|^2|\boldsymbol{A}|. \end{aligned}$$

On the other hand,

$$\begin{aligned} |\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}|_+ &= |\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}'\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{P}'|_+ \\ &= |\boldsymbol{\Lambda}\boldsymbol{Q}'\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}| \\ &= |\boldsymbol{\Lambda}|^2|\boldsymbol{A}|. \end{aligned}$$

□

Lemma 2 *For the adaptive IGMRF priors defined in (9) and (10), $|\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}|_+ = |\boldsymbol{B}_p\boldsymbol{B}_p'|$.*

Proof By Lemma 1, $|\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}|_+ = |\boldsymbol{B}_p'\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{B}_p|_+ = |\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{B}_p\boldsymbol{B}_p'|$. Since $\mathbf{1}'\boldsymbol{\gamma} = 0$, it follows that

$$|\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{B}_p\boldsymbol{B}_p'| = |\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}||\boldsymbol{B}_p\boldsymbol{B}_p'| = \exp(\mathbf{1}'\boldsymbol{\gamma})|\boldsymbol{B}_p\boldsymbol{B}_p'| = |\boldsymbol{B}_p\boldsymbol{B}_p'|.$$

□

Lemma 3 *Suppose $h(x)$ is a positive non-decreasing function. If F and G are distribution functions and there exist constants $M > 0$ and $N > 0$ such that $M[1-F(t)] \geq N[1-G(t)]$ for all t, then $M\int h(x)\mathrm{d}F(x) \geq N\int h(x)\mathrm{d}G(x)$.*

Proof Let $h_n(x) = \sum_i a_i I_{A_i}$ be a simple function, where the $A_i = [t_i, t_{i+1})$ form a finite decomposition of the support of $h(x)$ and $a_{i+1} \geq a_i \geq 0$ for all i. Let $b_i = a_i - a_{i-1}$ ($b_0 = 0$) and $B_i = A_i \cup A_{i+1} \cup \cdots$. Since $b_i \geq 0$ for all i, we have

$$\begin{aligned} M\int h_n(x)\mathrm{d}F(x) &= M\sum_i a_i F(A_i) \\ &= M\sum_i b_i F(B_i) \\ &\geq N\sum_i b_i G(B_i) \\ &= N\int h_n(x)\mathrm{d}G(x). \end{aligned}$$

Since there exists a sequence $\{h_n\}$ such that $0 \leq h_n \uparrow h$, the lemma is true by the monotone convergence theorem. □

Lemma 4 *Without loss of generality, assume $0 = x_1 < x_2 < \cdots < x_n = 1$ and set $h_j = x_j - x_{j-1}$, $j = 2, \ldots, n$. Suppose $e_j \overset{\text{ind}}{\sim} N(0, h_j(\tau\xi_1\xi_2)^{-1})$ and $X_k = \sum_{j=p+1}^{k} e_j$ for $k = p+1, \ldots, n$. Let $\bar{X} = \sum_{k=p+1}^{n} X_k/(n-p)$ and $\gamma_{\max} = \max_{p+1\leq k\leq n}\gamma_k$, where the prior on γ_k is defined in (10) with $q = 1$ and possibly unequally spaced points. Then*

(a) $\gamma_{\max} \overset{D}{=} \max_{p<k\leq n}\left(X_k - \bar{X}\right)$;
(b) *There is a constant $0 < c < \infty$ such that $P(\gamma_{\max} \geq t_0) < 4\left[1 - \Phi\left(t_0 c\sqrt{\tau\xi_1\xi_2}\right)\right]$ for all $t_0 > 0$.*

Proof Part (a) is obvious. To prove (b), let $B(t)$, $0 \leq t \leq 1$, be standard Brownian motion, let $V_n = \max_{p<k\leq n} X_k$, and let $W_n = -\bar{X}$. Since

$$e_j \overset{\mathcal{D}}{=} \frac{1}{\sqrt{\tau\xi_1\xi_2}}\left[B\left(x_j - x_p\right) - B\left(x_{j-1} - x_p\right)\right] \overset{\text{ind}}{\sim} N\left(0, h_j(\tau\xi_1\xi_2)^{-1}\right),$$
$$j = p+1, \ldots, n,$$

we have

$$X_k = \sum_{j=p+1}^{k} e_j \overset{\mathcal{D}}{=} \frac{1}{\sqrt{\tau\xi_1\xi_2}} B\left(x_k - x_p\right),$$
$$\bar{X} \sim N\left(0, \frac{c_1}{\tau\xi_1\xi_2}\right),$$

for some constant $0 < c_1 < \infty$. Thus,

$$\begin{aligned}
P(\gamma_{\max} \geq t_0) &\leq P(\max(V_n, W_n) \geq t_0/2) \\
&\leq P(V_n \geq t_0/2) + P(|W_n| \geq t_0/2) \\
&= P\left(\max_{p<k\leq n} X_k \geq \frac{t_0}{2}\right) + P\left(|\bar{X}| \geq \frac{t_0}{2}\right) \\
&< P\left[\frac{1}{\sqrt{\tau\xi_1\xi_2}} \max_{0\leq t\leq 1} B(t) \geq \frac{t_0}{2}\right] + 2\left[1 - \Phi\left(\frac{t_0}{2\sqrt{c_1}}\sqrt{\tau\xi_1\xi_2}\right)\right] \\
&= 2\left[1 - \Phi\left(\frac{t_0}{2}\sqrt{\tau\xi_1\xi_2}\right)\right] + 2\left[1 - \Phi\left(\frac{t_0}{2\sqrt{c_1}}\sqrt{\tau\xi_1\xi_2}\right)\right] \\
&< 4\left[1 - \Phi\left(t_0 c\sqrt{\tau\xi_1\xi_2}\right)\right],
\end{aligned}$$

where $c = \min\{1/2, 1/(2\sqrt{c_1})\}$. □

Lemma 5 (Marshall and Olkin 1979) *Assume that two $n \times n$ symmetric matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are both non-negative definite. Let $\lambda_1(\boldsymbol{S}_i) \leq \lambda_2(\boldsymbol{S}_i) \leq \cdots \leq \lambda_n(\boldsymbol{S}_i)$ be the eigenvalues of $\boldsymbol{S}_i$ for $i = 1, 2$. Then*

$$\prod_{j=1}^{n}[\lambda_j(\boldsymbol{S}_1) + \lambda_j(\boldsymbol{S}_2)] \leq |\boldsymbol{S}_1 + \boldsymbol{S}_2| \leq \prod_{j=1}^{n}[\lambda_j(\boldsymbol{S}_1) + \lambda_{n-j+1}(\boldsymbol{S}_2)].$$

Lemma 6 *Let $\boldsymbol{F}$ and $\boldsymbol{G}$ be two $n \times n$ non-negative matrices, where $\boldsymbol{F}$ has rank $(n-q)$ and $\boldsymbol{G}$ has rank $(n-p)$ for $p, q \geq 0$. Assume $\boldsymbol{F} + \boldsymbol{G}$ is positive definite. Denote by $\lambda_{\min}(\boldsymbol{F})$ and $\lambda_{\min}(\boldsymbol{G})$ the smallest non-zero eigenvalues of $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively.*

(a) $|\boldsymbol{F}+\boldsymbol{G}| \geq \lambda_{\min}(\boldsymbol{F})^n[1+\lambda_{\min}(\boldsymbol{G})/\lambda_{\min}(\boldsymbol{F})]^{(n-p)}$ *if $\boldsymbol{F}$ is positive definite* $(q = 0)$;
(b) $|\boldsymbol{F}+\boldsymbol{G}| \geq C_0|\boldsymbol{G}|_+$ *for some constant $0 < C_0 < \infty$, where $0 < C_0 < \infty$ depends only on $\boldsymbol{F}$, and $|\boldsymbol{G}|_+$ denotes the product of the non-zero eigenvalues of $\boldsymbol{G}$.*

Proof To prove (a), let $\lambda_i(\boldsymbol{F})$ and $\lambda_i(\boldsymbol{G})$ be the eigenvalues in ascending order of $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively. Note that $\lambda_1(\boldsymbol{F}) = \lambda_{\min}(\boldsymbol{F})$ and $\lambda_1(\boldsymbol{G}) = \cdots = \lambda_p(\boldsymbol{G}) = 0$. Following Lemma 5, we have

$$\begin{aligned}
|\boldsymbol{F}+\boldsymbol{G}| &\geq \prod_{i=1}^{n}[\lambda_i(\boldsymbol{F}) + \lambda_i(\boldsymbol{G})] \\
&\geq \prod_{i=1}^{n}[\lambda_{\min}(\boldsymbol{F}) + \lambda_i(\boldsymbol{G})] \\
&= \lambda_{\min}(\boldsymbol{F})^p \prod_{i=p+1}^{n}[\lambda_{\min}(\boldsymbol{F}) + \lambda_i(\boldsymbol{G})] \\
&\geq \lambda_{\min}(\boldsymbol{F})^p[\lambda_{\min}(\boldsymbol{F}) + \lambda_{\min}(\boldsymbol{G})]^{(n-p)} \\
&= \lambda_{\min}(\boldsymbol{F})^n[1 + \lambda_{\min}(\boldsymbol{G})/\lambda_{\min}(\boldsymbol{F})]^{(n-p)}.
\end{aligned}$$

To prove (b), let $\boldsymbol{Q}_{n\times n} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2)$, where $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_n$ and $\boldsymbol{Q}_2$ spans the null space of $\boldsymbol{G}$, so $\boldsymbol{Q}_2'\boldsymbol{G} = \boldsymbol{0}$. Since $\boldsymbol{F}+\boldsymbol{G}$ is positive definite, $\boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_2$ is invertible. Then, we have

$$\begin{aligned}
|\boldsymbol{F}+\boldsymbol{G}| &= \left|\boldsymbol{Q}'\left(\boldsymbol{F}+\boldsymbol{G}\right)\boldsymbol{Q}\right| \\
&= \begin{vmatrix} \boldsymbol{Q}_1'\boldsymbol{F}\boldsymbol{Q}_1 + \boldsymbol{Q}_1'\boldsymbol{G}\boldsymbol{Q}_1 & \boldsymbol{Q}_1'\boldsymbol{F}\boldsymbol{Q}_2 \\ \boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_1 & \boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_2 \end{vmatrix} \\
&= \left|\boldsymbol{Q}_1'\boldsymbol{F}^{\frac{1}{2}}\left[\boldsymbol{I}_n - \boldsymbol{F}^{\frac{1}{2}}\boldsymbol{Q}_2\left(\boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_2\right)^{-1}\boldsymbol{Q}_2'\boldsymbol{F}^{\frac{1}{2}}\right]\boldsymbol{F}^{\frac{1}{2}}\boldsymbol{Q}_1 + \boldsymbol{Q}_1'\boldsymbol{G}\boldsymbol{Q}_1\right| \left|\boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_2\right| \\
&\geq \left|\boldsymbol{Q}_1'\boldsymbol{G}\boldsymbol{Q}_1\right| \left|\boldsymbol{Q}_2'\boldsymbol{F}\boldsymbol{Q}_2\right| \\
&= C_0\left|\boldsymbol{G}\right|_+, \quad 0 < C_0 < \infty.
\end{aligned}$$

□

A.1 Proof of Theorem 1

Since $[\tau] \propto 1/\tau$, the joint posterior density of $(z, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ is proportional to

$$\begin{aligned}
h(z, \tau, \xi_1, \xi_2, \boldsymbol{\gamma}) = \tau^{\frac{n}{2}-1}\exp\left[-\frac{\tau}{2}(\boldsymbol{y}-z)'(\boldsymbol{y}-z)\right]\left|\tau\xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}} \\
\times \exp\left(-\frac{\tau\xi_1}{2}z'\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}z\right)[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2]\,[\xi_1, \xi_2].
\end{aligned}$$

Integrating out z, let

$$\begin{aligned}
h^*(\tau, \xi_1, \xi_2, \boldsymbol{\gamma}) &= \int_{\mathbb{R}^n} h\mathrm{d}z \\
&\propto \tau^{\frac{1}{2}(n-p)-1}\exp\left\{-\left[\frac{1}{2}\boldsymbol{y}'\left(\boldsymbol{I}_n - (\boldsymbol{I}_n + \xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)})^{-1}\right)\boldsymbol{y}\right]\tau\right\} \\
&\quad \times \frac{\left|\xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}}}{\left|\boldsymbol{I}_n + \xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|^{\frac{1}{2}}}[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2][\xi_1, \xi_2].
\end{aligned}$$

Following Lemma 1, we have

$$\left|\xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}} = \left|\xi_1\boldsymbol{B}_p'\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}\boldsymbol{B}_p\right|_+^{\frac{1}{2}} = \xi_1^{\frac{1}{2}(n-p)}\left|\boldsymbol{B}_p\boldsymbol{B}_p'\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}\right|^{\frac{1}{2}} = \xi_1^{\frac{1}{2}(n-p)}\left|\boldsymbol{B}_p\boldsymbol{B}_p'\right|^{\frac{1}{2}},$$

since $|\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)}| = \exp(\sum_{k=p+1}^{n}\gamma_k) = \exp(0) = 1$. Define $u = \gamma_{\min} = \min(\gamma_{p+1}, \ldots, \gamma_n)$ and let $\boldsymbol{F} = \boldsymbol{I}_n$, $\boldsymbol{G} = \xi_1\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}$ and $\boldsymbol{H} = \xi_1 e^u\boldsymbol{A}^{(p)}$. It is trivial that $\boldsymbol{G} \geq \boldsymbol{H}$. Decompose $\boldsymbol{A}^{(p)}$ with eigenvector matrix $\boldsymbol{P}$ and eigenvalues $0 < \lambda_{p+1} \leq \cdots \leq \lambda_n$. Then, using Lemma 6,

$$\left|\boldsymbol{I}_n + \xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|^{\frac{1}{2}} \geq [1 + \lambda_{\min}(\boldsymbol{G})]^{\frac{1}{2}(n-p)}$$
$$\geq [1 + \lambda_{\min}(\boldsymbol{H})]^{\frac{1}{2}(n-p)} = \left(1 + \lambda_{p+1}\xi_1 e^u\right)^{\frac{1}{2}(n-p)}. \quad (24)$$

Since $\boldsymbol{G} \geq \boldsymbol{H}$,

$$\boldsymbol{y}'\left[\boldsymbol{I}_n - \left(\boldsymbol{I}_n + \xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right)^{-1}\right]\boldsymbol{y} \geq \boldsymbol{y}'\left[\boldsymbol{I}_n - \left(\boldsymbol{I}_n + \xi_1 e^u \boldsymbol{A}^{(p)}\right)^{-1}\right]\boldsymbol{y}$$
$$\geq \|\boldsymbol{d}^*\|^2 \left(1 - \frac{1}{1 + \lambda_{p+1}\xi_1 e^u}\right), \quad (25)$$

where $\boldsymbol{d} = \boldsymbol{P}'y$ and $\boldsymbol{d}^* = (d_{p+1}, \ldots, d_n)'$. Note that $\|\boldsymbol{d}^*\|^2 > 0$ with probability one since $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. normal. The inequalities in (24) and (25) yield an upper bound for h^*,

$$h^*(\tau, \xi_1, \xi_2, \boldsymbol{\gamma}) \leq g(\tau, \xi_1, \xi_2, u)[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2]\,[\xi_1, \xi_2],$$

where

$$g(\tau, \xi_1, \xi_2, u) = \frac{C\,\xi_1^{\frac{1}{2}(n-p)}}{(1 + \lambda_{p+1}\xi_1 e^u)^{\frac{1}{2}(n-p)}}$$
$$\times \tau^{\frac{1}{2}(n-p)-1} \exp\left[-\frac{\tau}{2}\|\boldsymbol{d}^*\|^2 \left(1 - \frac{1}{1 + \lambda_{p+1}\xi_1 e^u}\right)\right]$$

for some constant $0 < C < \infty$. Therefore, we have

$$\int_{\gamma_{p+1}} \cdots \int_{\gamma_n} h(\tau, \xi_1, \xi_2, \boldsymbol{\gamma}) \mathrm{d}\gamma_n \cdots \mathrm{d}\gamma_{p+1} \leq \int_{-\infty}^{0} g(\tau, \xi_1, \xi_2, u)\,[u \mid \tau, \xi_1, \xi_2]\,[\xi_1, \xi_2]\mathrm{d}u.$$

Thus it suffices to show

$$\int_0^\infty \int_0^\infty \int_0^\infty \int_{-\infty}^{0} g(\tau, \xi_1, \xi_2, u)\,[u \mid \tau, \xi_1, \xi_2]\,[\xi_1, \xi_2]\mathrm{d}u\,\mathrm{d}\tau\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 < \infty. \quad (26)$$

Letting $w = -u$, we have

$$\int_{-\infty}^{0} g(\tau, \xi_1, \xi_2, u)\,[u \mid \tau, \xi_1, \xi_2]\,\mathrm{d}u = \int_0^\infty g(\tau, \xi_1, \xi_2, -w)\,[w \mid \tau, \xi_1, \xi_2]\,\mathrm{d}w.$$

Note that $W = -\gamma_{\min}$ and $\gamma_{\max}$ are identically distributed. Now let F denote the cdf of W and G the cdf of V, where V has density

$$[v \mid \tau, \xi_1, \xi_2] \propto \sqrt{\tau\xi_1\xi_2} \exp\left(-\frac{c^2\tau\xi_1\xi_2}{2}v^2\right), \quad v \geq 0.$$

Noting that $g(\tau, \xi_1, \xi_2, -w)$ is non-decreasing in w, Lemmas 3 and 4 together imply

$$\int_0^\infty g(\tau, \xi_1, \xi_2, -w)\,[w \mid \tau, \xi_1, \xi_2]\,\mathrm{d}w \leq 4\int_0^\infty g(\tau, \xi_1, \xi_2, -v)\,[v \mid \tau, \xi_1, \xi_2]\,\mathrm{d}v. \tag{27}$$

Substituting (27) into condition (26), it suffices to show

$$\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty g(\tau, \xi_1, \xi_2, -v)\,[v \mid \tau, \xi_1, \xi_2]\,\mathrm{d}v\,\mathrm{d}\tau\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 < \infty. \tag{28}$$

Letting $g^*(\tau, \xi_1, \xi_2, v)$ denote the integrand in (28),

$$\begin{aligned} g^*(\tau, \xi_1, \xi_2, v) \propto\ & \frac{\xi_1^{\frac{1}{2}(n-p)}\sqrt{\xi_1\xi_2}}{(1+\lambda_{p+1}\xi_1 e^{-v})^{\frac{1}{2}(n-p)}}\,[\xi_1, \xi_2] \\ & \times\tau^{\frac{1}{2}(n-p-1)} \exp\left\{-\left[\frac{1}{2}\|\boldsymbol{d}^*\|^2\left(1-\frac{1}{1+\xi_1\lambda_{p+1}e^{-v}}\right)+\frac{c^2\xi_1\xi_2}{2}v^2\right]\tau\right\}. \end{aligned}$$

Finally,

$$\begin{aligned} &\int_0^\infty g^*\mathrm{d}\tau \\ &\propto \frac{\xi_1^{\frac{1}{2}(n-p)}}{\left(1+\xi_1\lambda_{p+1}e^{-v}\right)^{\frac{1}{2}(n-p)}}\,\frac{\sqrt{\xi_1\xi_2}\,[\xi_1, \xi_2]}{\left[\frac{1}{2}\|\boldsymbol{d}^*\|^2\left(1-\frac{1}{1+\lambda_{p+1}\xi_1e^{-v}}\right)+c^2\xi_1\xi_2v^2/2\right]^{\frac{1}{2}(n-p+1)}}. \end{aligned} \tag{29}$$

Without loss of generality, let $\lambda_{p+1} = 1, r_0 = (n-p+1)/2$ and $s = \min\{\|\boldsymbol{d}^*\|^2/2, c^2/2\}$. It suffices to show

$$\int_0^\infty\int_0^\infty\int_0^\infty g^{**}(v, \xi_1, \xi_2)\,[\xi_1, \xi_2]\,\mathrm{d}v\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 < \infty, \tag{30}$$

where, up to a multiplicative constant,

$$g^{**}(v, \xi_1, \xi_2) = \frac{\left(\frac{\xi_1}{1+\xi_1e^{-v}}\right)^{r_0-\frac{1}{2}}\sqrt{\xi_1\xi_2}}{\left(\frac{\xi_1e^{-v}}{1+\xi_1e^{-v}}+\xi_1\xi_2v^2\right)^{r_0}}$$

is an upper bound of (29). We treat two cases separately.

Case 1 $0 \le v \le 1$. On the range $\xi_1 > 1$, we have

$$\frac{\xi_1}{1+\xi_1 e^{-v}} = \frac{1}{\xi_1^{-1}+e^{-v}} < e^v < e,$$
$$\frac{\xi_1 e^{-v}}{1+\xi_1 e^{-v}} = \frac{1}{1+\xi_1^{-1}e^{v}} > \frac{1}{1+e^v} > \frac{1}{1+e}.$$

Using these two inequalities,

$$\begin{aligned}\int_0^1 g^{**}\mathrm{d}v &< \int_0^1 \frac{e^{r_0-\frac{1}{2}}\sqrt{\xi_1\xi_2}}{\left(\frac{1}{1+e}+\xi_1\xi_2 v^2\right)^{r_0}}\,\mathrm{d}v \\ &= e^{r_0-\frac{1}{2}}\int_0^{\sqrt{\xi_1\xi_2}} \frac{1}{\left(\frac{1}{1+e}+w^2\right)^{r_0}}\,\mathrm{d}w \quad (w=\sqrt{\xi_1\xi_2}v) \\ &< e^{r_0-\frac{1}{2}}\int_0^{\infty} \frac{1}{\left(\frac{1}{1+e}+w^2\right)^{r_0}}\,\mathrm{d}w = C_1 < \infty.\end{aligned}$$

On the range $0 < \xi_1 \le 1$, we have

$$\frac{e^{-v}}{1+\xi_1 e^{-v}} = \frac{1}{\xi_1+e^{v}} > \frac{1}{1+e},$$
$$1+\xi_1 e^{-v} > 1.$$

Thus for $0 < \xi_1 \le 1$,

$$\begin{aligned}\int_0^1 g^{**}\mathrm{d}v &= \int_0^1 \frac{\left(\frac{1}{1+\xi_1 e^{-v}}\right)^{r_0-\frac{1}{2}}\sqrt{\xi_2}}{\left(\frac{e^{-v}}{1+\xi_1 e^{-v}}+\xi_2 v^2\right)^{r_0}}\,\mathrm{d}v \\ &< \int_0^1 \frac{\sqrt{\xi_2}}{\left(\frac{1}{1+e}+\xi_2 v^2\right)^{r_0}}\,\mathrm{d}v \\ &= \int_0^{\sqrt{\xi_2}} \frac{1}{\left(\frac{1}{1+e}+w^2\right)^{r_0}}\,\mathrm{d}w \quad (w=\sqrt{\xi_2}v) \\ &< \int_0^{\infty} \frac{1}{\left(\frac{1}{1+e}+w^2\right)^{r_0}}\,\mathrm{d}w = C_2 < \infty.\end{aligned}$$

From this calculation, we conclude that $\int_0^1 g^{**}(v,\xi_1,\xi_2)\,dv$ has finite expectation for any proper joint prior on (ξ_1,ξ_2).

Case 2 $v > 1$. On this range, we have

$$\begin{aligned}\int_1^\infty g^{**}\mathrm{d}v &= \int_1^\infty \frac{\left(\frac{1}{1+\xi_1 e^{-v}}\right)^{r_0-\frac{1}{2}}\sqrt{\xi_2}}{\left(\frac{e^{-v}}{1+\xi_1 e^{-v}}+\xi_2 v^2\right)^{r_0}}\,\mathrm{d}v \\ &< \int_1^\infty \frac{\sqrt{\xi_2}}{(\xi_2 v^2)^{r_0}}\,\mathrm{d}v \\ &= \frac{1}{\xi_2^{r_0-\frac{1}{2}}}\int_1^\infty v^{-2r_0}\,\mathrm{d}v \\ &= \frac{C_3}{\xi_2^{r_0-\frac{1}{2}}}.\end{aligned}$$

Combining the two cases, it is clear that (30) holds for any joint prior on (ξ_1, ξ_2) satisfying the moment condition of the theorem, and the proof is complete.

A.2 Proof of Theorem 2

With $[\tau] \propto 1/\tau$, the joint posterior density of $(\boldsymbol{\beta}, z, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ is proportional to

$$\begin{aligned}h(\boldsymbol{\beta}, z, \tau, \xi_1, \xi_2, \boldsymbol{\gamma}) = \tau^{\frac{N}{2}} \exp\left[-\frac{\tau}{2}(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}-\boldsymbol{X}_2 z)'(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}-\boldsymbol{X}_2 z)\right] \\ \times \left|\tau\xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}} \exp\left(-\frac{\tau\xi_1}{2} z' \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} z\right) \frac{1}{\tau}\,[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2]\,[\xi_1, \xi_2].\end{aligned}$$

Since $\boldsymbol{X}_1$ has full rank, $\boldsymbol{X}_1'\boldsymbol{X}_1$ is invertible. Let $\boldsymbol{T} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', z')'$. The usual least squares estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{z}')' = (\boldsymbol{T}'\boldsymbol{T})^- \boldsymbol{T}'\boldsymbol{y}$, where $(\boldsymbol{T}'\boldsymbol{T})^-$ is a generalized inverse. Let the sum of squared errors be

$$\mathrm{SSE} = \boldsymbol{y}'\left(\boldsymbol{I}_N - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^-\boldsymbol{T}'\right)\boldsymbol{y},$$

and define

$$\begin{aligned}\boldsymbol{M}_1 &= \boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' \\ \boldsymbol{R}_1 &= \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2 + \xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} \\ \boldsymbol{R}_2 &= \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{R}_1^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2.\end{aligned}$$

Note that $\boldsymbol{R}_1$ is positive definite by assumption in the theorem. Following Sun et al. (1999, 2001), the posterior density after integrating out $\boldsymbol{\beta}$ and z is proportional to

$h^*(\tau, \xi_1, \xi_2, \boldsymbol{\gamma}) = \int_{\mathbb{R}^{m+n}} h \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}z$ given by

$$\tau^{\frac{1}{2}(N-m-p)-1} \left|\xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}} |\boldsymbol{R}_1|^{-\frac{1}{2}} \times \exp\left[-\frac{\tau}{2}\left(\mathrm{SSE} + \hat{\boldsymbol{z}}'\boldsymbol{R}_2\hat{\boldsymbol{z}}\right)\right] [\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] \, [\xi_1, \xi_2]. \tag{31}$$

Letting $\boldsymbol{F} = \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$ and $\boldsymbol{G} = \xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}$ in Case 1, we have from Lemma 6

$$|\boldsymbol{R}_1|^{\frac{1}{2}} = |\boldsymbol{F} + \boldsymbol{G}|^{\frac{1}{2}} \geq C_0^{\frac{1}{2}} |\boldsymbol{G}|_+^{\frac{1}{2}} = C_0^{\frac{1}{2}} \left|\xi_1 \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)}\right|_+^{\frac{1}{2}}. \tag{32}$$

Since $\mathrm{SSE} > 0$ and $\hat{\boldsymbol{z}}'\boldsymbol{R}_3\hat{\boldsymbol{z}} \geq 0$, (32) yields an upper bound for h^*,

$$h^*(\tau, \xi_1, \xi_2, \boldsymbol{\gamma}) \leq C_0^{-\frac{1}{2}} g(\tau) \, [\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] \, [\xi_1, \xi_2],$$

where

$$g(\tau) = \tau^{\frac{1}{2}(N-m-p)-1} \exp\left(-\frac{\tau}{2}\mathrm{SSE}\right).$$

Note that $g(\tau)$ is an integrable function given $N \geq m + p + 1$. Then it suffices to show

$$\int_0^\infty \int_0^\infty \int_0^\infty \int_{\mathbb{R}^{n-p}} g(\tau) \, [\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] \, [\xi_1, \xi_2] \, \mathrm{d}\boldsymbol{\gamma} \, \mathrm{d}\tau \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 < \infty. \tag{33}$$

Since $[\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2]$ is a proper density function and $g(\tau)$ is integrable, the left hand side of (33) after integrating out $\boldsymbol{\gamma}$ and τ is

$$C_1 \int_0^\infty \int_0^\infty [\xi_1, \xi_2] \, \mathrm{d}\xi_1 \mathrm{d}\xi_2, \quad \text{for some } 0 < C_1 < \infty.$$

Hence, proper priors on ξ_1 and ξ_2 ensure the propriety of the posterior in Case 1.

In Case 2, $\mathrm{SSE} = 0$ and it is necessary to find a lower bound for $\hat{\boldsymbol{z}}'\boldsymbol{R}_2\hat{\boldsymbol{z}}$ in (31). For convenience, suppress the superscript in the notation $\boldsymbol{A}^{(p)}$. Again let $u = \min_k \gamma_k$. Since $\boldsymbol{G} \geq \xi_1 e^u \boldsymbol{A}$, we have

$$\begin{aligned} \hat{\boldsymbol{z}}'\boldsymbol{R}_2\hat{\boldsymbol{z}} &= \hat{\boldsymbol{z}}' \left[\boldsymbol{F} - \boldsymbol{F}\left(\boldsymbol{F} + \boldsymbol{G}\right)^{-1} \boldsymbol{F}\right] \hat{\boldsymbol{z}} \\ &\geq \hat{\boldsymbol{z}}' \left[\boldsymbol{F} - \boldsymbol{F}\left(\boldsymbol{F} + \xi_1 e^u \boldsymbol{A}\right)^{-1} \boldsymbol{F}\right] \hat{\boldsymbol{z}}. \end{aligned} \tag{34}$$

Note that $\boldsymbol{F}$ is positive semidefinite in this case. Suppose $\mathrm{rank}(\boldsymbol{F}) = r \leq (n-1)$. Do the spectral decomposition

$$\boldsymbol{F} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}' = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_1' \\ \boldsymbol{P}_2' \end{pmatrix},$$

where $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ span the range and null spaces of $\boldsymbol{F}$, and $\boldsymbol{\Lambda}_1$ is a diagonal matrix of non-zero eigenvalues of $\boldsymbol{F}$. Let

$$\left(\boldsymbol{\Lambda}+\xi_1 e^u \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\right)^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_1+\xi_1 e^u \boldsymbol{P}_1'\boldsymbol{A}\boldsymbol{P}_1 & \xi_1 e^u \boldsymbol{P}_1'\boldsymbol{A}\boldsymbol{P}_2 \\ \xi_1 e^u \boldsymbol{P}_2'\boldsymbol{A}\boldsymbol{P}_1 & \xi_1 e^u \boldsymbol{P}_2'\boldsymbol{A}\boldsymbol{P}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}. \tag{35}$$

Using a well-known formula (e.g., Christensen 2002, p. 423), $\boldsymbol{S}_{11}=(\boldsymbol{\Lambda}_1+\xi_1 e^u \boldsymbol{K})^{-1}$, where $\boldsymbol{K}_{r\times r}=\boldsymbol{P}_1'\boldsymbol{A}^{1/2}(\boldsymbol{I}_n-\boldsymbol{M})\boldsymbol{A}^{1/2}\boldsymbol{P}_1$, and $\boldsymbol{M}=\boldsymbol{A}^{1/2}\boldsymbol{P}_2\left(\boldsymbol{P}_2'\boldsymbol{A}\boldsymbol{P}_2\right)^{-1}\boldsymbol{P}_2'\boldsymbol{A}^{1/2}$ is a projection matrix. Note that $\boldsymbol{K}$ is non-negative. Suppose it has rank $(r-\ell)$ for $\ell \geq 0$. Following (34) and (35),

$$\begin{aligned}
\hat{\boldsymbol{z}}'\boldsymbol{R}_2\hat{\boldsymbol{z}} &\geq \hat{\boldsymbol{z}}'\boldsymbol{P}\left[\boldsymbol{\Lambda}-\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}+\xi_1 e^u \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\right)^{-1}\boldsymbol{\Lambda}\right]\boldsymbol{P}'\hat{\boldsymbol{z}}. \\
&= \hat{\boldsymbol{z}}'\left(\boldsymbol{P}_1,\boldsymbol{P}_2\right)\left[\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}-\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22},\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\right]\begin{pmatrix}\boldsymbol{P}_1'\\ \boldsymbol{P}_2'\end{pmatrix}\hat{\boldsymbol{z}} \\
&= \hat{\boldsymbol{z}}'\boldsymbol{P}_1(\boldsymbol{\Lambda}_1-\boldsymbol{\Lambda}_1\boldsymbol{S}_{11}\boldsymbol{\Lambda}_1)\boldsymbol{P}_1'\hat{\boldsymbol{z}} \\
&= \hat{\boldsymbol{z}}'\boldsymbol{P}_1\boldsymbol{\Lambda}_1^{\frac{1}{2}}\left[\boldsymbol{I}_r-\left(\boldsymbol{I}_r+\xi_1 e^u \boldsymbol{\Lambda}_1^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{\Lambda}_1^{-\frac{1}{2}}\right)^{-1}\right]\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{P}_1'\hat{\boldsymbol{z}} \\
&\geq \|\boldsymbol{d}^*\|^2\left(1-\frac{1}{1+\lambda_{\min}(\boldsymbol{J})\xi_1 e^u}\right),
\end{aligned} \tag{36}$$

where $\boldsymbol{J}=\boldsymbol{\Lambda}_1^{-1/2}\boldsymbol{K}\boldsymbol{\Lambda}_1^{-1/2}$, $\lambda_{\min}(\boldsymbol{J})$ is the smallest non-zero eigenvalue of $\boldsymbol{J}$, $\boldsymbol{d}^*=(d_{\ell+1},\ldots,d_r)'$, $\boldsymbol{d}_{r\times 1}=\boldsymbol{P}_J'\boldsymbol{\Lambda}_1^{1/2}\boldsymbol{P}_1'\hat{\boldsymbol{z}}$, and $\boldsymbol{P}_J$ is the eigenvector matrix of $\boldsymbol{J}$. We then use (32) and (36) to give an upper bound for h^*,

$$h^*(\tau,\xi_1,\xi_2,\boldsymbol{\gamma}) \leq C_0^{-\frac{1}{2}} g(\tau,\xi_1,\xi_2,u)[\boldsymbol{\gamma}\mid\tau,\xi_1,\xi_2]\,[\xi_1,\xi_2],$$

where

$$g(\tau,\xi_1,\xi_2,u)=\tau^{\frac{1}{2}(N-m-p)-1}\exp\left\{-\frac{\tau}{2}\left[\|\boldsymbol{d}^*\|^2\left(1-\frac{1}{1+\lambda_{\min}(\boldsymbol{J})\,\xi_1 e^u}\right)\right]\right\}.$$

Following the proof of Theorem 1, it suffices to show

$$\int_0^\infty\int_0^\infty\int_0^\infty g^{**}(v,\xi_1,\xi_2)\,[\xi_1,\xi_2]\,\mathrm{d}v\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2<\infty,$$

where

$$g^{**}(v,\xi_1,\xi_2)=\frac{\sqrt{\xi_1\xi_2}}{\left(\frac{\xi_1 e^{-v}}{1+\xi_1 e^{-v}}+\xi_1\xi_2 v^2\right)^{r_0}},$$

for $r_0 = (N - m - p + 1)/2$. Note that r_0 is assumed to be positive. When $0 \leq v \leq 1$ and $\xi_1 > 1$,

$$\int_0^1 g^{**} \mathrm{d}v < C_2 < \infty.$$

When $0 \leq v \leq 1$ and $0 < \xi_1 \leq 1$,

$$\int_0^1 g^{**} \mathrm{d}v < C_3 \xi_1^{-r_0+\frac{1}{2}}, \quad 0 < C_3 < \infty.$$

When $v > 1$,

$$\int_1^\infty g^{**} \mathrm{d}v < C_4 (\xi_1 \xi_2)^{-r_0+\frac{1}{2}}, \quad 0 < C_4 < \infty.$$

Hence, it suffices to have proper priors on ξ_1 and ξ_2 with finite expected value of $(\xi_1 \xi_2)^{-r_0+\frac{1}{2}}$ in Case 2. The theorem has been proved.

A.3 Proof of Theorem 3

The joint posterior density of $(\boldsymbol{\beta}, \boldsymbol{b}, \tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ is proportional to

$$h(\boldsymbol{\beta}, \boldsymbol{z}, \tau, \xi_1, \xi_2, \boldsymbol{\gamma}) = \tau^{n/2} \exp\left[-\frac{\tau}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})\right] \times \mid \tau \xi_1 \boldsymbol{D}_\gamma \mid^{1/2} \exp\left(-\frac{\tau \xi_1}{2} \boldsymbol{b}' \boldsymbol{D}_\gamma \boldsymbol{b}\right) \frac{1}{\tau} [\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] [\xi_1, \xi_2].$$

Following the proof of Theorem 2, let $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Z})$ and denote the least squares estimator by $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{b}}')' = (\boldsymbol{T}'\boldsymbol{T})^- \boldsymbol{T}'\boldsymbol{y}$. Again let the sum of squared errors be $\mathrm{SSE} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^- \boldsymbol{T}')\boldsymbol{y}$ and define

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ \boldsymbol{R}_1 &= \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} + \xi_1 \boldsymbol{D}_\gamma \\ \boldsymbol{R}_2 &= \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}\boldsymbol{R}_1^{-1}\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}. \end{aligned}$$

As in (31), the posterior of $(\tau, \xi_1, \xi_2, \boldsymbol{\gamma})$ is proportional to

$$h^* = \tau^{\frac{1}{2}(n-p-1)-1} |\xi_1 \boldsymbol{D}_\gamma|^{\frac{1}{2}} |\boldsymbol{R}_1|^{-\frac{1}{2}} \times \exp\left[-\frac{\tau}{2}\left(\mathrm{SSE} + \hat{\boldsymbol{b}}' \boldsymbol{R}_2 \hat{\boldsymbol{b}}\right)\right] [\boldsymbol{\gamma} \mid \tau, \xi_1, \xi_2] [\xi_1, \xi_2].$$

Note that $\mathrm{SSE} > 0$ in P-splines since one always used far fewer knots than observations. Again let $\boldsymbol{F} = \boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ and $\boldsymbol{G} = \xi_1 \boldsymbol{D}_\gamma$. The rest of proof exactly follows Case 1 in Theorem 2.

Appendix B. Derivations of smoother matrices

B.1 Smoother matrix for $\boldsymbol{\gamma}$ in BASS

Recall that negative two times the logarithm of the full conditional density of $\boldsymbol{\gamma}$ in BASS up to additive constant is

$$L(\boldsymbol{\gamma}) = \tau\xi_1 z' \boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} z + \tau\xi_1\xi_2 \boldsymbol{\gamma}' \boldsymbol{A}^{(q)} \boldsymbol{\gamma}, \tag{37}$$

where $\boldsymbol{A}_{\boldsymbol{\gamma}}^{(p)} = \boldsymbol{B}_p' \boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)} \boldsymbol{B}_p$ and $\boldsymbol{D}_{\boldsymbol{\gamma}}^{(p)} = \text{diag}(e^{\gamma_{p+1}}, \ldots, e^{\gamma_n})$ for given p and q. Note that in (37) we ignore the linear constraint on $\boldsymbol{\gamma}$ since it is irrelevant for the purpose. Letting $\tilde{z} = \boldsymbol{B}_p z = (\tilde{z}_{p+1}, \ldots, \tilde{z}_n)'$, expand $L(\boldsymbol{\gamma})$ about the mode $\tilde{\gamma}_k$ as in the Laplace approximation,

$$\begin{aligned} L(\boldsymbol{\gamma}) &= \tau\xi_1 \sum_{k=p+1}^{n} \tilde{z}_k^2 e^{\gamma_k} + \tau\xi_1\xi_2 \boldsymbol{\gamma}' \boldsymbol{A}^{(q)} \boldsymbol{\gamma} \\ &\approx \tau\xi_1 \sum_{k=p+1}^{n} \tilde{z}_k^2 \left[e^{\tilde{\gamma}_k} + (\gamma_k - \tilde{\gamma}_k) e^{\tilde{\gamma}_k} + \frac{1}{2}(\gamma_k - \tilde{\gamma}_k)^2 e^{\tilde{\gamma}_k} \right] + \tau\xi_1\xi_2 \boldsymbol{\gamma}' \boldsymbol{A}^{(q)} \boldsymbol{\gamma} \\ &= \tau\xi_1 \left[(\mathbf{1} - \tilde{\boldsymbol{\gamma}})' \boldsymbol{W} \boldsymbol{\gamma} + \frac{1}{2} \boldsymbol{\gamma}' \boldsymbol{W} \boldsymbol{\gamma} + \xi_2 \boldsymbol{\gamma}' \boldsymbol{A}^{(q)} \boldsymbol{\gamma} \right], \end{aligned} \tag{38}$$

where $\boldsymbol{W} = \text{diag}\left(\tilde{z}_{p+1}^2 e^{\tilde{\gamma}_{p+1}}, \ldots, \tilde{z}_n^2 e^{\tilde{\gamma}_n} \right)$ and $\mathbf{1} = (1, \ldots, 1)'$. By taking the first derivative in $\boldsymbol{\gamma}$ and setting it to be zero, the approximate posterior mode $\tilde{\boldsymbol{\gamma}}$ satisfies

$$\tilde{\boldsymbol{\gamma}} = \left(\boldsymbol{W} + 2\xi_2 \boldsymbol{A}^{(q)} \right)^{-1} \boldsymbol{W} (\tilde{\boldsymbol{\gamma}} - \mathbf{1}).$$

Therefore the equivalent smoother matrix for $\boldsymbol{\gamma}$ is $\left(\boldsymbol{W} + 2\xi_2 \boldsymbol{A}^{(q)} \right)^{-1} \boldsymbol{W}$. Note that in practice we take $\tilde{\gamma}_k = 0$ when there is no prior information about the posterior mode for choosing a prior for ξ_1.

B.2 Smoother matrix for $\boldsymbol{\gamma}$ in BAPS

Negative two times the logarithm of the full conditional density of $\boldsymbol{b}_\gamma$ in BAPS (up to additive constant) is

$$L(\boldsymbol{b}_\gamma) = \tau\xi_1 \boldsymbol{b}' \boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{b} + \tau\xi_1\xi_2 \boldsymbol{b}_\gamma' \boldsymbol{b}_\gamma, \tag{39}$$

where $\boldsymbol{D}_{\boldsymbol{\gamma}} = \text{diag}(e^{\gamma_1}, \ldots, e^{\gamma_{m_t}})$ and $\boldsymbol{\gamma} = \boldsymbol{Z}_\gamma \boldsymbol{b}_\gamma$. Again we expand (39) around the mode $\tilde{\boldsymbol{\gamma}}$ as in (38)

$$
\begin{aligned}
L(\boldsymbol{b}_\gamma) &\approx \tau\xi_1 \sum_{k=1}^{m_t} b_k^2 \left[e^{\tilde{\gamma}_k} + (\gamma_k - \tilde{\gamma}_k) e^{\tilde{\gamma}_k} + \frac{1}{2}(\gamma_k - \tilde{\gamma}_k)^2 e^{\tilde{\gamma}_k} \right] + \tau\xi_1\xi_2 \boldsymbol{b}_\gamma' \boldsymbol{b}_\gamma \\
&= \tau\xi_1 \left[(\mathbf{1} - \tilde{\boldsymbol{\gamma}})' \boldsymbol{W}\boldsymbol{Z}_\gamma \boldsymbol{b}_\gamma + \frac{1}{2} \boldsymbol{b}_\gamma' \boldsymbol{Z}_\gamma' \boldsymbol{W}\boldsymbol{Z}_\gamma \boldsymbol{b}_\gamma + \xi_2 \boldsymbol{b}_\gamma' \boldsymbol{b}_\gamma \right],
\end{aligned}
$$

where $\boldsymbol{W} = \text{diag}\left(b_1^2 e^{\tilde{\gamma}_1}, \ldots, b_{m_t}^2 e^{\tilde{\gamma}_{m_t}}\right)$. Therefore, the posterior mode $\tilde{\boldsymbol{b}}_\gamma$ for the approximation satisfies

$$
\tilde{\boldsymbol{b}}_\gamma = \left(\boldsymbol{Z}_\gamma' \boldsymbol{W}\boldsymbol{Z}_\gamma + 2\xi_2 \boldsymbol{I}_{m_t} \right)^{-1} \boldsymbol{Z}_\gamma' \boldsymbol{W}(\tilde{\boldsymbol{\gamma}} - \mathbf{1}).
$$

The equivalent smoother matrix in BAPS for $\boldsymbol{\gamma}$ is $\boldsymbol{Z}_\gamma \left(\boldsymbol{Z}_\gamma' \boldsymbol{W}\boldsymbol{Z}_\gamma + 2\xi_2 \boldsymbol{I}_{m_t} \right)^{-1} \boldsymbol{Z}_\gamma' \boldsymbol{W}$, and e.d.f. is

$$
\text{trace}\left(\boldsymbol{Z}_\gamma \left(\boldsymbol{Z}_\gamma' \boldsymbol{W}\boldsymbol{Z}_\gamma + 2\xi_2 \boldsymbol{I}_{m_t} \right)^{-1} \boldsymbol{Z}_\gamma' \boldsymbol{W} \right) = \text{trace}\left(\boldsymbol{I}_{m_t} + 2\xi_2 \left(\boldsymbol{Z}_\gamma' \boldsymbol{W}\boldsymbol{Z}_\gamma \right)^{-1} \right).
$$

Following BASS, we take $\tilde{\gamma}_k = 0$ in order to choose the prior on ξ_2.

References

Abramovich, F., Steinberg, D. M. (1996). Improved inference in nonparametric regression using L_k-smoothing splines. *Journal of Statistical Planning and Inference, 49*, 327–341.

Abramovich, F., Sapatinas, T., Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 60*, 725–749.

Baladandayuthapani, V., Mallick, B. K., Carroll, R. J. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics, 14*, 378–394.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*, 385–402.

Berger, J. O., Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison (Pkg: P135-207). In *Model selection. Institute of mathematical statistics lecture notes-monograph series* (Vol. 38, pp. 135–193). Fountain Hills, AZ: IMS Press.

Besag, J., Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika, 82*, 733–746.

Besag, J., Green, P., Higdon, D., Mengersen, K. (1995). Bayesian computation and stochastic systems (Disc: P41-66). *Statistical Science, 10*, 3–41.

Brezger, A., Fahrmeir, L., Hennerfeind, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 56*, 327–345.

Carter, C. K., Kohn, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika, 83*, 589–601.

Chipman, H. A., Kolaczyk, E. D., McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association, 92*, 1413–1421.

Christensen, R. (2002). *Plane answers to complex questions: The theory of linear models*. New York: Springer.

Clyde, M., Parmigiani, G., Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika, 85*, 391–401.

Crainiceanu, C., Ruppert, D., Carroll, R., Adarsh, J., Goodner, B. (2007). Spatially adaptive penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics, 16*, 265–288.

Cummins, D. J., Filloon, T. G., Nychka, D. (2001). Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *Journal of the American Statistical Association, 96*, 233–246.
Dass, S. C., Berger, J. O. (2003). Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics, 30*, 193–210.
Denison, D. G. T., Mallick, B. K., Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B, 60*, 333–350.
Di Matteo, I., Genovese, C. R., Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika, 88*, 1055–1071.
Eilers, P., Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science, 11*, 89–121.
Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York: Marcel Dekker.
Fahrmeir, L., Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association, 91*, 1584–1594.
Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics, 19*, 1–141.
Gilks, W. R., Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics, 41*, 337–348.
Gilks, W. R., Best, N. G., Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within gibbs sampling. *Applied Statistics, 44*, 455–472.
Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. London: Chapman & Hall.
Hastie, T., Tibshirani, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science, 15*, 196–223.
Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.
Hobert, J. P., Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association, 91*, 1461–1473.
Jacquier, E., Polson, N. G., Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models (Disc: P389-417). *Journal of Business & Economic Statistics, 12*, 371–389.
Johnstone, I. M., Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics, 33*, 1700–1752.
Knorr-Held, L. (2003). Some remarks on Gaussian Markov random field models for disease mapping. In N. H. P. Green, S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 260–264). New York: Oxford University Press.
Knorr-Held, L., Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society, Series C: Applied Statistics, 52*, 169–183.
Lang, S., Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics, 13*, 183–212.
Lang, S., Fronk, E.-M., Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics, 17*, 479–499.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*, 410–423.
Lindgren, F., Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics, 35*, 691–700.
Mackay, D. J. C., Takeuchi, R. (1998). Interpolation models with multiple hyperparameters. *Statistics and Computing, 8*, 15–23.
Marin, J.-M., Robert, C. P. (2007). *Bayesian core: A practical approach to computational Bayesian statistics*. New York: Springer.
Nakatsuma, T. (2000). Bayesian analysis of ARMA-GARCH models: A Markov chain sampling approach. *Journal of Econometrics, 95*, 57–69.
Ngo, L., Wand, M. (2004). Smoothing with mixed model software. *Journal of Statistical Software, 9*, 1–54.
Paciorek, C. J., Schervish, M. J. (2004). Nonstationary covariance functions for Gaussian process regression. In S. Thrun, L. Saul, B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16). Cambridge, MA: MIT Press.
Paciorek, C. J., Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics, 17*, 483–506.

Pensky, M. (2006). Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. *The Annals of Statistics, 34*, 769–807.

Pintore, A., Speckman, P. L., Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika, 93*, 113–125.

Rue, H., Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. *Monographs on statistics and applied probability* (Vol. 104). London: Chapman & Hall.

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

Ruppert, D., Wand, M., Carroll, R. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Smith, M. S., Kohn, R. J. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics, 75*, 317–343.

Speckman, P. L., Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika, 90*, 289–302.

Staniswalis, J. G. (1989). Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association, 84*, 284–288.

Staniswalis, J. G., Yandell, B. S. (1992). Locally adaptive smoothing splines. *Journal of Statistical Computation and Simulation, 43*, 45–53.

Sun, D., Speckman, P. L. (2008). Bayesian hierarchical linear mixed models for additive smoothing splines. *Annals of the Institute of Statistical Mathematics, 60*, 499–517.

Sun, D., Tsutakawa, R. K., He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica, 11*, 77–95.

Sun, D., Tsutakawa, R. K., Speckman, P. L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika, 86*, 341–350.

Van Der Linde, A. (2003). PCA-based dimension reduction for splines. *Journal of Nonparmetric Statistics, 15*, 77–92.

Vrontos, I. D., Dellaportas, P., Politis, D. N. (2000). Full Bayesian inference for GARCH and EGARCH models. *Journal of Business & Economic Statistics, 18*, 187–198.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

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

Wand, M., Jones, M. (1995). *Kernel smoothing*. London: Chapman & Hall.

White, G. (2006). *Bayesian semi-parametric spatial and joint spatio-temporal smoothing*. Ph.D. thesis, University of Missouri-Columbia, Columbia, MO.

Wood, S., Jiang, W., Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika, 89*, 513–528.

Yue, Y., Speckman, P. L. (2010) Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics, 19*, 96–116.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel, A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). New York/Amsterdam: Elsevier/North-Holland.

Zellner, A., Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)* (pp. 585–603). Valencia: University of Valencia.