Fully Bayesian spline smoothing and intrinsic autoregressive priors

BY PAUL L. SPECKMAN AND DONGCHU SUN

Department of Statistics, University of Missouri, Columbia, Missouri 65211, U.S.A.

speckman@stat.missouri.edu  dsun@stat.missouri.edu

SUMMARY

There is a well-known Bayesian interpretation for function estimation by spline smoothing using a limit of proper normal priors. The limiting prior and the conditional and intrinsic autoregressive priors popular for spatial modelling have a common form, which we call partially informative normal. We derive necessary and sufficient conditions for the propriety of the posterior for this class of partially informative normal priors with noninformative priors on the variance components, a condition crucial for successful implementation of the Gibbs sampler. The results apply for fully Bayesian smoothing splines, thin-plate splines and L-splines, as well as models using intrinsic autoregressive priors.

Some key words: Autoregressive Gaussian process; Gibbs sampling; Linear mixed model; Multivariate normal; Penalised likelihood methods; Spatial correlation; Spline smoothing.

1. INTRODUCTION

There is a large literature on the frequentist approach to nonparametric regression (Eubank, 1988; Wand & Jones, 1995; Fan & Gijbels, 1996). The Bayesian view is not so well developed, although the Bayesian approach has a long history. In work fundamental to this paper, Kimeldorf & Wahba (1970) and Wahba (1978) demonstrated that certain forms of spline smoothing correspond to Bayesian estimates under a class of improper Gaussian prior distributions on function spaces. This framework is fully explained in Wahba (1990) and forms the basis for this paper. We discuss a fully Bayesian approach with noninformative priors on critical variance components. Our purpose is to derive conditions under which the posterior distribution is proper.

Bayesian function estimation is closely related to the treatment of spatial components in hierarchical models. With appropriate proper priors, researchers have found powerful Markov chain Monte Carlo tools such as Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990) for estimation in Bayesian models with spatial components. The conditional autoregressive model of Besag (1974) has spawned a large number of papers. Since stationary distributions are sometimes inappropriate, Künsch (1987) introduced intrinsic autoregressions for spatial data, models where differences are stationary autoregressive processes. These intrinsic priors can be associated with versions of spline smoothing. We show that the limiting Bayes prior of Wahba and coauthors and the intrinsic autoregressive priors have a common form, which we term partially informative normal. We use this distribution to unify results on the propriety of the posterior for fully
Bayesian smoothing splines, thin-plate splines and $L$-splines among others, as well as for discretised versions.

In § 2, we introduce the basic models. We begin with a class of penalised least squares estimators including smoothing splines and thin-plate splines, and we review their Bayesian justification as limits of posterior means with normal priors. These priors are closely related to a large class of spatial priors currently popular in the Bayesian literature including conditional autoregressive processes and intrinsic autoregressive processes. We then show how these processes and the limiting improper priors derived from a Hilbert space framework for penalised likelihood problems including smoothing splines are instances of partially informative normal priors. The main results of this paper, given in § 3, provide necessary and sufficient conditions for the propriety of the joint posterior in this common framework for hierarchical Bayesian models with partially informative normal priors and commonly used noninformative priors on two variance components. These results extend earlier work in Sun et al. (1999), where sufficient conditions were derived for a class of proper priors on the variance components. Finally, several computational algorithms are listed and discussed in § 4.

2. PARTIALLY INFORMATIVE PRIORS FOR SMOOTHING SPLINES AND RELATED MODELS

2.1. Spline smoothing and Bayesian estimation

Consider the basic one-dimensional nonparametric regression problem

$$y_i = g(t_i) + \varepsilon_i \quad (i = 1, \ldots, n),$$

where the $\varepsilon_i$ are independent $N(0, \sigma_0^2)$ random variables and $g$ is an unknown function. A Bayesian version of this problem (Kimeldorf & Wahba, 1970; Wahba, 1978) is to take a Gaussian process prior on $g(.)$,

$$g(t_i) = \beta_0 + \beta_1 t_i + \ldots + \beta_{p-1} t_i^{p-1} + x_i \quad (i = 1, \ldots, n),$$

(1)

with $x_i = X(t_i)$, where $X(t)$ is a zero-mean autoregressive Gaussian process whose $p$th derivative, $D^p X(t) = \delta_t^p \, dW(t)$, is scaled white noise. Wahba (1978) showed that the extended Bayes estimate $\hat{g}$ with improper prior on $\beta = (\beta_0, \ldots, \beta_{p-1})'$ is exactly the solution to the spline smoothing problem,

$$\min_{g} \left[ \sum_{i=1}^{n} (y_i - g(t_i))^2 + \eta \int_0^1 g^{(p)}(t)^2 \, dt \right],$$

(2)

where $\eta = \sigma_0^2/\delta_1$ and the minimum is taken over all sufficiently differentiable functions $g$. This solution is the smoothing spline of degree $2p - 1$. Wahba (1983) extended the Bayesian interpretation and derived the posterior distribution of $g$ with corresponding Bayesian credibility intervals.

Many other authors have exploited this connection between nonparametric regression and Bayesian estimation. For example, Ansley & Kohn (1985) computed a marginal likelihood for $\eta$ using a diffuse prior on $\beta$, Kohn & Ansley (1987) used state-space techniques to develop efficient algorithms for spline smoothing, and Carter & Kohn (1994, 1996) employed Gibbs sampling in a fully Bayesian hierarchical model to obtain estimates.

The theory of penalised least squares as illustrated by (2) can be greatly generalised. The general theory depends in part on the representation of such problems in terms of reproducing kernel Hilbert spaces. The following description is a greatly compressed
Bayesian spline smoothing

account of the general theory. Details can be found in Wahba (1990) or Gu (2002). A reproducing kernel Hilbert space is a space of functions $\mathcal{H}$ on a domain $\mathcal{X}$ together with an inner product such that the point evaluation is a bounded linear functional. It follows that there is a symmetric bivariate function $R(s, t)$ such that $R(s, .) \in \mathcal{H}$ for all $s \in \mathcal{X}$, and $g(s) = \langle g, R(s, .) \rangle$ for any $g \in \mathcal{H}$. For the polynomial spline example of (2), one can take $R(s, t) = R_0(s, t) + R_1(s, t)$, where $R_0$ is a kernel associated with the $p$-dimensional space of polynomials of degree less than $p$, and $R_1$ is a kernel associated with the penalty $J(g) = \int_0^1 (g(t)^{2p})^2 \, dt$. One such kernel is

$$R_1(s, t) = \int_0^1 \frac{(s-u)^k(t-1)^{k-1}}{(p-1)! (p-1)!} \, du.$$ 

It can be shown that the solution (2) has the form

$$g(t) = \sum_{i=1}^n d_i R_1(t_i, t) + \sum_{k=1}^p c_k \frac{t^k}{(k-1)!}.$$ 

Moreover,

$$J(g) = \sum_{i=1}^n \sum_{j=1}^n d_i d_j R(t_i, t_j).$$

This set-up generalises. In a number of important cases, there is a seminorm $J(g)$ on a function space over a domain $\mathcal{X}$ with null space spanned by functions $\phi_j(t)$ ($j = 1, \ldots, p$) and an associated kernel $R_1(s, t)$ such that

$$g(t) = \sum_{i=1}^n d_i R_1(t_i, t) + \sum_{k=1}^p c_k \phi_k(t)$$

implies (3). If $w_1, \ldots, w_n$ are positive weights, it can be shown that the penalised least squares problem of minimising

$$\sum_{i=1}^n w_i (y_i - g(t_i))^2 + \eta J(g)$$

over a suitable space $\mathcal{X}$ of functions reduces to the finite-dimensional problem of minimising (5) over functions of the form (4).

Now let $T$ be the $n \times p$ matrix with elements $T_{ij} = \phi_j(t_i)$, let $W = \text{diag}(w_1, \ldots, w_n)$, and let $\Sigma$ be the $n \times n$ matrix with elements $R(t_i, t_j)$. Then the solution to (5) is obtained by minimising

$$(y - Td - \Sigma c)' W (y - Td - \Sigma c) + \eta c' \Sigma c.$$ 

Differentiation with respect to $c$ and $d$ shows that the solution satisfies

$$\Sigma W (\Sigma + \eta W^{-1}) c + Td - y = 0, \quad T' W (\Sigma c + T d - y) = 0.$$ 

Wahba (1978) showed that under certain circumstances this solution has a Bayesian interpretation. Since the kernel $R(s, t)$ is positive semidefinite, $\Sigma$ is a covariance matrix. Suppose that $x \sim N_n(0, \delta \Sigma)$ and $\theta \sim N_p(0, \delta_2 I_p)$. Consider the prior $g = T \theta + x$, where $x$, $\theta$ and $\epsilon \sim N_n(0, \delta_0 W^{-1})$ are independent, and let $\delta_2 \to \infty$. Clearly

$$\text{cov}(g, y) = \text{var}(g) = \delta_1 \Sigma + \delta_2 T T', \quad \text{var}(y) = \delta_0 W^{-1} + \delta_1 \Sigma + \delta_2 T T'.$$

Thus, if we use multivariate normality and set $a = \delta_2 / \delta_1$ and $\eta = \delta_0 / \delta_1$, the posterior
distribution of $g$ is multivariate normal with
\[ E(g \mid y, a, \eta) = \hat{g}_a = (aTT' + \Sigma)(aTT' + \Sigma + \eta W^{-1})^{-1} y, \]
\[ \text{var}(g \mid y, a, \eta) = \delta_1 \{(aTT' + \Sigma) - (aTT' + \Sigma)(aTT' + \Sigma + \eta W^{-1})^{-1}(aTT' + \Sigma)\}. \]

Letting $a \to \infty$, Wahba showed that $\lim_{a \to \infty} \hat{g}_a = \hat{g}_0$ is exactly the solution to system (6). Writing this solution in terms of an influence matrix $\hat{g}_0 = S(\eta)\tilde{y}$, Wahba (1983) showed that $\lim_{a \to \infty} \text{var}(g \mid y, a, \eta) = \delta_0 S(\eta)$. Details can also be found in Wahba (1990, § 5.1) or Gu (2002, §§ 2.5, 3.3). The general theory also easily extends to produce posterior distributions at points other than the observation points.

There are other important applications. For smoothing in one dimension, the penalty can be generalised to $J(g) = \int (Lg)^2 \, dx$, where $L$ is a linear differential operator. The minimiser of (5) is called an $L$-spline. Examples including null spaces and reproducing kernels are given in Gu (2002, § 4.3), for example. Another important example is the thin-plate spline (Wahba, 1990, § 2.4; Gu, 2002, § 4.4) for smoothing data on the $d$-dimensional domain $\mathcal{X} = (-\infty, \infty)^d$ with
\[ J^*_p(g) = \sum_{s_1 + \ldots + s_d = p} \int \ldots \int \left( \frac{\partial^p g}{\partial s_1^{s_1} \ldots \partial s_d^{s_d}} \right)^2 \, ds_1 \ldots ds_d. \]
The reproducing kernel is somewhat complicated but fits into the framework here; see Wahba (1990; § 2.4) or Gu (2002, § 4.4.2) for details.

In all these cases, the general version of the smoothing spline corresponds to a Bayes estimate under a partially diffuse prior. The proof goes through in complete generality, and this Bayesian interpretation for penalised estimates holds for a large class of estimation problems in function spaces as well (van der Linde, 1995).

### 2.2. Partially informative and autoregression normal priors

We will show that the priors appropriate for spline smoothing belong to a class we have termed partially informative normal (Sun et al., 1999). Let $B$ be a symmetric positive semidefinite matrix, and consider the possibly improper density
\[ f(x) \propto |B|^\frac{1}{2} \exp \left( -\frac{1}{2} x'Bx \right). \]
Here $|B|^\frac{1}{2}$ is defined to be the product of the nonzero eigenvalues of $B$. If $B > 0$, this distribution is of course multivariate normal with mean zero and covariance matrix $B^{-1}$. In this context, the inverse of the covariance matrix is sometimes referred to as the precision matrix. The interesting cases occur when $B$ is singular and the distribution is improper. As in Sun et al. (1999), the distribution can be interpreted as having two parts, a constant noninformative prior on the null space of $B$ and a proper degenerate normal on the range of $B$. Equivalently, the distribution can be thought of as the convolution of independent priors, a constant prior on the null space and the multivariate $N(0, B^{-1})$ distribution, where $B^{-1}$ is the Moore–Penrose inverse of $B$. In practice, $B$ is often only known up to an unknown parameter $\delta_1$. This results in the partially informative normal $\text{PIN}(\delta_1^{-1} A)$ prior with improper density
\[ f(z) \propto |A|^\frac{1}{2} \frac{1}{\delta_1^{(n-p)}} \exp \left( -\frac{1}{2\delta_1} z'Az \right), \quad z \in \mathbb{R}^n, \]
where $A$ is a nonnegative definite symmetric matrix with rank $n - p$. Note that this definition differs slightly from the one in Sun et al. (1999). The power of $\delta_1^q$ is the rank of $A$ here rather than the dimension of $z$.

There is another way of looking at the partially informative normal prior specified by (9). It follows that

$$f(z_i | z_j, j \neq i) = \left( \frac{a_{ii}}{2\pi \delta_1} \right)^{\frac{1}{2}} \exp \left\{ -\frac{a_{ii}}{2\delta_1} \left( z_i - \sum_{j=1}^{n} c_{ij}z_j \right)^2 \right\},$$

(10)

where $c_{ii} = 0$ and $c_{ij} = -a_{ij}/a_{ii}$ for $i \neq j$. This is exactly the form of the conditional autoregressive, CAR, process of Besag (1974), who argued in the opposite direction. Suppose that the conditional distributions of $z_i$ given the remaining $z_j$ all have the form

$$f(z_i | z_j, j \neq i) \sim N \left( \sum_{j=1}^{n} c_{ij}z_j, \frac{\delta_1}{d_i} \right),$$

where $c_{ii} = 0$. Suppose further that the matrix $A$ with elements defined by $a_{ii} = d_i$, $a_{ij} = d_i c_{ij}$ is positive definite. Then $z$ is $N_p(0, \delta_1 A^{-1})$. A CAR process by definition has a proper distribution, meaning that $A$ must be positive definite. Künsch (1987), see also Besag & Kooperberg (1995), related this condition by allowing a differenced process to be stationary, creating 'intrinsic autoregressive' or IAR priors.

One increasingly popular application of IAR models is in estimating smooth curves in various Bayesian hierarchical models. For example, Fahrmeir & Wagenpfeil (1996) used the model $z_i = 2z_{i-1} - z_{i-2} + \xi_i$ for independent $\xi_i \sim N(0, \delta_1)$. If a diffuse prior is taken on $z_1$ and $z_2$, the joint prior can be expressed as

$$f(z) \propto \delta_1^{-\frac{1}{2}(n-2)} \exp \left\{ -\frac{1}{2\delta_1} \sum_{k=3}^{n} (z_k - 2z_{k-1} + z_{k-2})^2 \right\}, \quad z \in \mathbb{R}^n.$$

This clearly has the form of (9). Fahrmeir & Lang (2001) have generalised the discrete model to approximate divided differences for non-equally spaced points. For the equally spaced case, the second-order form above easily generalises to a $p$th-order form,

$$f(z) \propto \delta_1^{-\frac{1}{2}(n-p)} \exp \left\{ -\frac{1}{2\delta_1} \sum_{k=p+1}^{n} \left\{ \sum_{i=0}^{p} (-1)^i \binom{p}{i} z_{k-i} \right\}^2 \right\}, \quad z \in \mathbb{R}^n.$$

(11)

This corresponds to (9) with a banded matrix $A$ such that $a_{ij} = 0$ if $|i - j| > p$. In general, except for edge effects,

$$a_{ij} = \begin{cases} (-1)^{|i-j|} n^2 p \binom{2p}{p-|i-j|}, & \text{for } |i - j| \leq p, \\ 0, & \text{otherwise}. \end{cases}$$

We will refer to (11) as the IAR($p$) prior. The IAR(1) prior depending only on immediately neighbouring points corresponds to linear smoothing splines, $p = 1$, and is not smooth enough for most applications in function estimation.

An appealing feature of the IAR($p$) prior is that it has a Markov property since the coefficients $c_{ij}$ in (11) are zero whenever $|i - j| > p$. Thus (11) has the form of the CAR process of Besag (1974), namely that the conditional distribution of $z_i$ depends only on neighbours. Similar IAR models have become extremely popular in Bayesian analysis involving spatial data (Besag et al., 1991; Waller et al., 1997; Ghosh et al., 1998), and they are all examples of partially informative normal priors.
To see the differences among the IAR\((p)\) models when we change \(p\), in Fig. 1 we graph three sample paths of the IAR\((p)\) process prior for \(z\) when \(n = 100\) and \(\delta_1 = 0.1\) for each of \(p = 1, 2\) and 3. The graphs were plotted as follows. Note that the rank of \(A\) is \(n - p\). Let \(0 < \lambda_{p+1} \leq \ldots \leq \lambda_n\) be the positive eigenvalues of \(A\). Let \(\gamma_j\) be the eigenvector of \(A\) corresponding to \(\lambda_j\), for \(j = p + 1, \ldots, n\). Then

\[
A = \sum_{j=p+1}^{n} \lambda_j \gamma_j \gamma_j', \quad A^{-1} = \sum_{j=p+1}^{n} \lambda_j^{-1} \gamma_j \gamma_j'.
\]

Sample \(w = (w_1, \ldots, w_{n-p})'\) from \(N_{n-p}(0, \delta_1 I_{n-p})\). Then \(z = \sum_{j=p+1}^{n} \lambda_j^{-1/2} \gamma_j'w\), an \(n \times 1\) vector, will have the proper part of the IAR\((p)\) process prior with variance component \(\delta_1\). When \(p = 1\), the sample paths have essentially only the smoothness of Brownian motion and are too rough to be a prior for a smooth function. Note that the proper part of the sample path is orthogonal to polynomials of degree less than \(p\).

### 2.3. Partially informative normal priors and smoothing splines

In this section, we give an alternative form for the limiting posterior normal distribution specified by (7) and (8), showing how the smoothing spline priors are related to IAR models. This representation will allow a unified treatment of posterior distributions in §3. Summarising the assumptions in §2.1 with \(g\) replaced by \(z\), suppose that

\[
y = z + \varepsilon, \quad z = T\theta + x, \quad \varepsilon \sim N_n(0, \delta_0 W^{-1}), \quad x \sim N_n(0, \delta_1 \Sigma), \quad \theta \sim N_p(0, \delta_2 I_p),
\]
Bayesian spline smoothing

where $\Sigma \succeq 0$ and $W > 0$ are known symmetric matrices, $T$ is a known full-rank $n \times p$
matrix, $\varepsilon$, $x$ and $\theta$ are independent, and for the moment $\delta_0$, $\delta_1$ and $\delta_2$ are known. We also
assume that the null space of $\Sigma$ is contained in the column span of $T$. Note that we relax
the usual assumption that $\Sigma$ is invertible; all that is required is that $(aTT' + \Sigma)^{-1}$ exist
for any $a > 0$. As in § 2.1, the prior on $z$ is $N(0, \delta_1(\Sigma + aTT'))$. To specify a limiting
partially informative prior, we need the following lemma.

**Lemma 1.** (i) If $\Sigma$ is positive definite,

$$\lim_{a \to \infty} (aTT' + \Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1}T(T'^{-1}T)\Sigma^{-1}. \quad (12)$$

(ii) Let $Q = \Sigma + I$. If the column span of $T$ contains the null space of $\Sigma$,

$$\lim_{a \to \infty} (aTT' + \Sigma)^{-1} = [I - (Q^{-1} - Q^{-1}T(T'Q^{-1}T)^{-1}T'Q^{-1})^{-1} - I]. \quad (13)$$

(iii) In either case, the rank of the limit is $n - p$.

**Proof.** The first limit can be found in several places including Wahba (1990, eqn 1.5.11).
To prove (ii), note that $(G - I)^{-1} = (I - G^{-1})^{-1} - I$ when $G$ and $(G - I)$ are both
invertible. Then

$$(aTT' + \Sigma)^{-1} = (aTT' + Q - I)^{-1} = (I - (aTT' + Q)^{-1} - I).$$

Taking the limit as $a \to \infty$ and applying part (i) gives the desired result.

Finally, the matrix on the right-hand side of (12) clearly has rank $n - p$. For case (ii),
let $B = I - (Q^{-1} - Q^{-1}T(T'Q^{-1}T)^{-1}T'Q^{-1})$. If $B^{-1}$ exists, the right-hand side of (13) is

$$B^{-1} - I = B^{-1}(I - B) = B^{-1}(Q^{-1} - Q^{-1}T(T'Q^{-1}T)^{-1}T'Q^{-1}),$$

which again clearly has rank $n - p$.

To see that $B$ is invertible, assume without loss of generality that $\Sigma$ is a diagonal matrix
$D$, and let $\hat{T} = (D + I)^{-\frac{1}{2}}T$. By assumption, $TT' + D > 0$, which is true if and only if
$\hat{T}\hat{T}' + D(D + I)^{-1} > 0$. Similarly, $B$ is positive definite if and only if

$$\hat{B} = (D + I)^{\frac{1}{2}}B(D + I)^{\frac{1}{2}} = D + \hat{T}(\hat{T}'\hat{T})^{-1}\hat{T}' > 0.$$ 

Suppose that $a'\hat{B}a = 0$ for some $a \neq 0$. Then $a'Da = a'D(D + I)^{-1}a = 0$, since $D$ is diagonal
with nonnegative elements. However, by assumption, $a'(\hat{T}\hat{T}' + D(D + I)^{-1})a = a'\hat{T}\hat{T}'a > 0$, which implies that $a'\hat{T}(\hat{T}'\hat{T})^{-1}\hat{T}'a > 0$, contradicting the assumption that $a'\hat{B}a = 0$. Thus $B$ has full rank.

For the rest of this section, define

$$A = \lim_{a \to \infty} (aTT' + \Sigma)^{-1}.$$ 

Then up to a proportionality constant, the prior density on $z$ satisfies

$$\pi_a(z|\delta_1) \propto \exp \left\{-\frac{1}{2\delta_1}z'(aTT' + \Sigma)^{-1}z\right\} \to \exp \left(-\frac{1}{2\delta_1}z'Az\right),$$

as $a \to \infty$. Thus it is natural to assume the partially informative normal prior

$$\pi^*(z|\delta_1) \sim \text{PIN}(\delta_1^{-1}A).$$

We can now consider two posterior distributions, the limiting posterior distribution of $z$
given $y, \delta_0$ and $\delta_1$, as $a \to \infty$, and the posterior distribution of $z$ given $y, \delta_0$ under the improper
limit prior \( \pi^* \). The possibility that these two distributions could be different appears to have been overlooked in the spline literature; see for example Gu (2002, Theorem 3.6, p. 73). However, because the limit prior \( \pi^* \) is improper, one cannot immediately pass limits inside integrals, and it should be verified that the distributions are the same. The two posterior distributions are in fact the same, as shown by the following theorem.

**Theorem 1.** Let \( h_a \) and \( h^* \) be the posterior densities of \( z \) given \( (y, \delta_0, \delta_1) \) under \( \pi_a \) and \( \pi^* \) respectively.

(i) Then \( h^*(z|y, \delta_0, \delta_1) \) is the density of the proper normal distribution

\[
N\{ (W + \eta A)^{-1} W y_0, \delta_0 (W + \eta A)^{-1} \}.
\]

(ii) We have that \( \lim_{a \to \infty} h_a(z|y, \delta_0, \delta_1) = h^*(z|y, \delta_0, \delta_1) \) for all \( z \).

**Proof.** To prove (i), note that the joint density of \( (y, z) \) under \( \pi^* \) is

\[
f(y, z|\delta_0, \delta_1) \propto \exp \left\{ -\frac{1}{\delta_0} (y - z)' W (y - z) - \frac{1}{2\delta_1} z' A z \right\}.
\]

Completing the square, we obtain

\[
h^*(z|y, \delta_0, \delta_1) \propto \exp \left[ -\frac{1}{2\delta_0} \left\{ (-2z' W y + z'(W + \eta A) z) \right\} \right]
\]

\[
= \exp \left[ -\frac{1}{2\delta_0} \left\{ (y - z)' (W + \eta A)^{-1} W y + (W + \eta A)(z - (W + \eta A)^{-1} W y) \right\} \right].
\]

To prove (ii), fix \( a \). It suffices to prove convergence of the posterior mean and variance. From (7),

\[
\hat{z}_a = (a TT' + \Sigma) (a TT' + \Sigma + \eta W^{-1})^{-1} y = \left\{ I + \eta W^{-1} (a TT' + \Sigma)^{-1} \right\}^{-1} y
\]

\[
\to (W + \eta A)^{-1} W y,
\]

as \( a \to \infty \), using Lemma 1. Similarly, using (8), we have

\[
\text{var}(z|y, a) = \delta_1 \left\{ (a TT' + \Sigma) - (a TT' + \Sigma)(a TT' + \Sigma + \eta W^{-1})^{-1}(a TT' + \Sigma) \right\} = \delta_1 \left\{ (a TT' + \Sigma) - (a TT' + \Sigma + \eta W^{-1} - \eta W)^{-1} \right\} \times (a TT' + \Sigma + \eta W^{-1})^{-1}(a TT' + \Sigma)
\]

\[
= \delta_1 \eta W^{-1} (a TT' + \Sigma + \eta W^{-1})^{-1}(a TT' + \Sigma)
\]

\[
= \delta_0 W^{-1} \left\{ I + \eta (a TT' + \Sigma)^{-1} W^{-1} \right\}^{-1} = \delta_0 \left\{ W + \eta (a TT' + \Sigma)^{-1} \right\}^{-1} \to \delta_0 (W + \eta A)^{-1},
\]

as \( a \to \infty \), using Lemma 1 again. \( \square \)

**Remark 1.** The limiting posterior distribution in (ii) was first derived by Wahba (1983) for the case \( \Sigma > 0 \) and \( W = I \). Gu (2002, p. 74) pointed out the modifications needed for general \( W \). With some manipulation, their expressions for the posterior mean and variance can be shown to be equivalent to (i). The form given here is not necessarily best for computation, but it expedites the theory in § 3.
Bayesian spline smoothing

3. Propriety of posteriors

3.1. Model and notation

To be general, we assume the possibility of repeated observations at distinct points \( t_i \), either in an interval for spline smoothing or in a more general domain. Assume now that

\[
y_{ij} = z_i + \epsilon_{ij} \quad (i = 1, \ldots, n, j = 1, \ldots, r_i),
\]

where the \( \epsilon_{ij} \) are independent \( N(0, \delta_0) \). In the context of § 2.1, \( z_i = g(t_i) \) \( (i = 1, \ldots, n) \). Then there are \( r_1 + \ldots + r_n = N \) observations. For simplicity in the proofs of Theorems 2 and 3 below, we assume that \( r_i \geq 1 \) for all \( i \). We adopt the following notation:

\[
\bar{y}_i = r_i^{-1} \sum_j y_{ij}, \quad \bar{y} = (\bar{y}_1, \ldots, \bar{y}_n), \quad \text{SSE} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2, \\
\eta = \delta_0/\delta_1, \quad W = \text{diag}(r_1, \ldots, r_n), \quad H_\eta = W - W(W + \eta A)^{-1}W.
\]

For given \( \delta_1, z \) has a \( \text{PIN}(\delta_1^{-1}A) \) partially informative distribution with precision matrix \( A \) having rank \( n - p \) and variance component \( \delta_1 \). Note that this prior includes the limiting Bayes models for smoothing splines, thin-plate splines and the IAR models. The set-up may be regarded as a one-way analysis of variance model with correlated random effects.

3.2. General priors for the variance components

We begin with results about the propriety of the posterior using partially informative normal priors for the unknown function \( g \) and general priors for \( \delta_0 \) and \( \delta_1 \). If \( w(\delta_0, \delta_1) \) is a prior density of \( (\delta_0, \delta_1) \), the joint posterior density of \( (z, \delta_0, \delta_1) \) is

\[
h(z, \delta_0, \delta_1) \propto \frac{|A|^\frac{1}{2}}{\delta_0^{(N-n)}(2\pi\delta_1)^{\frac{1}{2}(n-p)}} \exp \left[ -\frac{1}{2\delta_0} \{(\bar{y} - z)W(\bar{y} - z) + \text{SSE}\} - \frac{1}{2\delta_1} z^T A z \right] w(\delta_0, \delta_1).
\]

**Theorem 2.** Consider model (14) with prior (9).

(i) If \( w(\delta_0, \delta_1) \) is a proper density, the joint posterior of \( (z, \delta_0, \delta_1) \) is proper.

(ii) If \( w(\delta_0, \delta_1) \) is improper, the joint posterior of \( (z, \delta_0, \delta_1) \) is proper if and only if

\[
\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{(2\pi)^{p/2}}{\delta_0^{(N-n)+1}(2\pi\delta_1)^{\frac{1}{4}(n-p)}|W + \eta A|^\frac{1}{2}} \exp \left( -\frac{\text{SSE} + \bar{y}^T H_\eta \bar{y}}{2\delta_0} \right) w(\delta_0, \delta_1) d\delta_0 d\delta_1 < \infty.
\]

**Proof.** Note that

\[
\int_{\mathbb{R}^n} h dz \propto \frac{|A|^\frac{1}{2}}{\delta_0^{(N-n)}(2\pi\delta_1)^{\frac{1}{2}(n-p)}|W + \eta A|^\frac{1}{2}} \exp \left( -\frac{\text{SSE} + \bar{y}^T H_\eta \bar{y}}{2\delta_0} \right) w(\delta_0, \delta_1).
\]

Let \( 0 < \lambda_{p+1} \leq \ldots \leq \lambda_n \) be the \( n - p \) positive eigenvalues of \( W^{-\frac{1}{2}} A W^{-\frac{1}{2}} \). Then

\[
|W + \eta A| = |W| \prod_{j=p+1}^{n} (1 + \eta \lambda_j) \geq |W| \left( \frac{\delta_0}{\delta_1} \right)^{n-p} \prod_{j=p+1}^{n} \lambda_j.
\]

We obtain

\[
\int_{\mathbb{R}^n} h dz \leq C \frac{|A|^\frac{1}{2}}{\delta_0^{(N-n)}} \exp \left( -\frac{\text{SSE}}{2\delta_0} \right) w_0(\delta_0, \delta_1),
\]

for some positive constant \( C \). Part (i) follows immediately, and part (ii) is obvious. \( \square \)
Note that, in both cases, the joint marginal posterior density of \((\delta_0, \delta_1)\) is proportional to the integrand in (15).

**Corollary 1.** Assume that \(\delta_0\) and \(\delta_1\) are a priori independent. If the prior for \(\delta_1\) is proper and \(w_0(\delta_0)\), the prior of \(\delta_0\), satisfies

\[
\int_0^\infty \delta_0^{-\frac{1}{2}(N-p)} \exp \left( -\frac{\text{SSE}}{2\delta_0} \right) w_0(\delta_0) d\delta_0 < \infty,
\]

then the joint posterior distribution of \((z, \delta_0, \delta_1)\) exists.

### 3.3. Inverse gamma and related priors

To complete the Bayesian formulation, we must specify the prior for \((\delta_1, \delta_0)\). In practice, it might be difficult to elicit the prior, especially \(\delta_1\). One would then want to search for a noninformative prior. With \(r = 1\), Carter & Kohn (1994) applied the prior

\[
w(\delta_0, \delta_1) \propto \frac{1}{\delta_0} \exp \left( -\frac{b_0}{\delta_0} \right) \tag{18}
\]

for some small \(b_0 > 0\). Note that a constant prior for \(\delta_1\) was used.

There has been a great deal of recent development on noninformative priors; see for example Berger & Bernardo (1989, 1992) and Ye (1994). In work that applies here, van der Linde (2000) found reference priors for a general normal variance components model with two variance components. Consider the balanced case \(W = r I_n\). When \(\delta_1\) is the parameter of interest and \(\delta_0\) is the nuisance parameter, van der Linde’s noninformative prior is

\[
w_*(\delta_0, \delta_1) \propto \frac{1}{\delta_1^{\frac{3}{2}}} \left( \frac{(r - 1)n}{\delta_0^2} + \sum_{i=p+1}^{n} \frac{1}{(r \delta_i^{-1} \delta_1 + \delta_0)^2} \right)^{\frac{1}{2}}.
\]

Such a prior is inconvenient in applications, especially when \(n\), the number of points, becomes large. However, for large \(n\), with the help of approximations given for example in Speckman (1985), it is possible to show that

\[
w_*(\delta_0, \delta_1) \approx \frac{1}{\delta_1^{\frac{3}{2}} \delta_0}, \tag{19}
\]

which corresponds to an invariant prior for \(\delta_0\) and a constant prior for \(\delta_1^{\frac{3}{2}}\).

In general, we assume that the priors of \(\delta_0\) and \(\delta_1\) are independent and that, for \(i = 0, 1\), the prior density of \(\delta_i\) is of inverse Gamma type, with

\[
w_i(\delta_i) \propto \frac{1}{\delta_i^{a_i + 1}} \exp \left( -\frac{b_i}{\delta_i} \right), \tag{20}
\]

for some real constants \((a_i, b_i)\). If \(a_i\) and \(b_i\) are both positive, that is \(\delta_i\) has a proper inverse gamma prior, it follows from Theorem 2 that the joint posterior is proper. This kind of prior is convenient for computation. In practice, many authors choose \(a_i = b_i = \xi\), a very small number such as 0.0001. Hobert & Casella (1996) showed that the posterior may be improper for a normal linear mixed model with a constant prior for fixed effects and invariance prior for all the blockwise independent random effects, that is \(a_i = b_i = 0\) in our case. The following theorem characterises all possible choices of \((a_i, b_i)\) for our model. Related results were obtained by Sun et al. (1999), who provided sufficient conditions for
the propriety of the posterior in a class of linear models with partially informative normal priors for the random effects but proper priors for the variance components. The following conditions are used.

**Condition A.** One of the following holds:

(A1) \( b_1 > 0 \) and \( N - p + 2a_0 > 0 \);

(A2) \( b_1 = 0 \) and \( a_1 < 0 \).

**Condition B.** One of the following holds:

(B1) \( \text{sse} + 2b_0 > 0 \) and \( n - p + 2a_1 > 0 \);

(B2) \( \text{sse} + 2b_0 = 0 \) and \( N - n + 2a_0 < 0 \).

**Condition C.** We require that \( N - p + 2a_0 + 2a_1 > 0 \).

**Theorem 3.** Consider model (14) with prior given by (9) and (20). For fixed \( p \), the posterior of \((z, \delta_0, \delta_1)\) is proper if and only if Conditions A, B and C hold.

**Proof.** First, \( h_1 \) defined in (16) has the form

\[
h_1(\delta_0, \delta_1) \propto \frac{(2\pi)^{p/2} |A|^{1/2}}{\delta_0^{(N-n)+a_0+1} \delta_1^{(n-p)+a_1+1}} \exp \left( - \frac{\text{sse} + y^\prime H_y y + b_0 - b_1}{2\delta_0} \right). \tag{21}
\]

The posterior is proper if and only if \( \int_0^\infty \int_0^\infty h_1(\delta_0, \delta_1) \, d\delta_0 \, d\delta_1 < \infty \). Since \( y^\prime H_y y \leq y^\prime W y \), we have

\[
h_1 \geq \frac{C(2\pi)^{p/2} |A|^{1/2}}{\delta_0^{(N-n)+a_0+1} \delta_1^{(n-p)+a_1+1}} \exp \left( - \frac{\text{sse} + y^\prime W y + 2b_0 - b_1}{2\delta_0} \right). \tag{22}
\]

Here and in the following \( C \) or \( C_i \) denotes a positive constant depending on \( y \) only. Clearly, when \( b_1 \) is negative, the integrals of the right-hand side with respect to \( \delta_1 \) is not finite. Thus \( b_1 \) must be nonnegative.

Now make the transformation \( \eta = \delta_0/\delta_1 \) in (21). Clearly the marginal posterior density of \((\delta_0, \eta)\), if it exists, satisfies

\[
h_2(\eta | y) \propto \frac{(2\pi)^{p/2} |A|^{1/2} \eta^{\frac{(n-p)+a_1-1}{2}}}{\delta_0^{\frac{(N-n)+a_0+a_1+1}{2}} |W+\eta A|^{1/2}} \exp \left\{ - \frac{\frac{1}{2}(\text{sse} + y^\prime H_y y + b_0 + b_1 \eta)}{\delta_0} \right\}. \tag{23}
\]

For any fixed \( \eta > 0 \), the integral of the right-hand side with respect to \( \delta_0 \) is finite if and only if Condition C holds. In this case, the marginal posterior density of \( \eta \) given \( y \), if it exists, has the form

\[
h_2(\eta | y) \propto \frac{(2\pi)^{1/2} |A|^{1/2} \eta^{\frac{(n-p)+a_1-1}{2}}}{|W+\eta A|^{1/2} (\text{sse} + y^\prime H_y y + 2b_0 + 2b_1 \eta)^{\frac{(N-n)+a_0+a_1}{2}}}.
\]

Arguing as in the proof of Theorem 2, use the orthogonal decomposition

\[\begin{align*}
W^{-1/2}AW^{-1/2} &= \Gamma \text{diag}(0, \ldots, 0, \lambda_{p+1}, \ldots, \lambda_n) \Gamma', \\
\end{align*}\]

and let \( d \equiv (d_1, \ldots, d_n)' = \Gamma' W^{1/2} y \). Then

\[
|W|(1 + \eta \lambda_{p+1})^{p-n} \leq |W+\eta A| \leq |W|(1 + \eta \lambda_n)^{p-n}, \quad y^\prime H_y y = \sum_{k=p+1}^n d_k^2 \frac{\eta \lambda_k}{1 + \eta \lambda_k}.
\]

Clearly, if \( \text{sse} + b_0 < 0 \), \( h_2(\eta | y) \) would be negative or imaginary when \( \eta \) is small. So we need \( \text{sse} + b_0 \geq 0 \). We have the following four cases.
Case 1: \( b_1 > 0 \) and \( \text{sse} + b_0 > 0 \).
Case 2: \( b_1 > 0 \) and \( \text{sse} + b_0 = 0 \).
Case 3: \( b_1 = 0 \) and \( \text{sse} + b_0 > 0 \).
Case 4: \( b_1 = 0 \) and \( \text{sse} + b_0 = 0 \).

In Case 1, when \( \epsilon > 0 \) is small enough, there are constants \( C_1 \) and \( C_2 \) such that
\[
\int_{0}^{\epsilon} h_2(\eta \mid y) \, d\eta \sim C_1 \int_{0}^{\epsilon} \eta^{\frac{1}{n}} \, d\eta, \quad \int_{\epsilon}^{\infty} h_2(\eta \mid y) \, d\eta \sim C_2 \int_{\epsilon}^{\infty} \frac{1}{\eta^{\frac{1}{n}}} \, d\eta.
\]

Here \( g_1(\epsilon) \sim g_2(\epsilon) \) as \( \epsilon \to 0 \) means that \( \lim_{\epsilon \to 0} g_1(\epsilon)/g_2(\epsilon) = 1 \). Thus both terms are finite if and only if \( n - p + 2a_1 > 0 \) and \( N - p + 2a_0 > 0 \). Similarly, for Case 2 and \( \epsilon > 0 \) near zero, there are constants \( C_1 \) and \( C_2 \) such that
\[
\int_{0}^{\epsilon} h_2(\eta \mid y) \, d\eta \sim C_1 \int_{0}^{\epsilon} \eta^{-\frac{1}{n}} \, d\eta, \quad \int_{\epsilon}^{\infty} h_2(\eta \mid y) \, d\eta \sim C_2 \int_{\epsilon}^{\infty} \frac{1}{\eta^{\frac{1}{n}}} \, d\eta,
\]

which are both finite if and only if \( N - n + 2a_0 < 0 \) and \( N - p + 2a_0 > 0 \). Cases 3 and 4 are similar. The results then follow.

Table 1 gives the results of the theorem for some typical choices of noninformative priors. The behaviour of the posterior can depend on whether or not there is at least one point with multiple observations. The case \( \text{sse} = 0 \) corresponds to \( r_i = 1 \), that is \( N = n \).

**Table 1. Propriety of posterior for some choices of the hyperparameters \((a_0, b_0; a_1, b_1)\) in the priors of \((\delta_0, \delta_1)\)**

<table>
<thead>
<tr>
<th>Case</th>
<th>Prior</th>
<th>Hyperparameters ((a_0, b_0; a_1, b_1))</th>
<th>Posterior (\text{sse} = 0)</th>
<th>Posterior (\text{sse} &gt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invariance priors</td>
<td>(1/\delta_0\delta_1)</td>
<td>((0, 0; 0, 0))</td>
<td>improper</td>
<td>improper</td>
</tr>
<tr>
<td>Constant for ((\delta_0, \delta_1))</td>
<td>1</td>
<td>((-1, 0; -1, 0))</td>
<td>proper</td>
<td>proper</td>
</tr>
<tr>
<td>Constant for ((\delta_0^2, \delta_1^2))</td>
<td>(1/\delta_0\delta_1)</td>
<td>((-\frac{1}{2}, 0; -\frac{1}{2}, 0))</td>
<td>proper</td>
<td>proper</td>
</tr>
<tr>
<td>Prior (18) with (b_0 = 0)</td>
<td>(1/\delta_0)</td>
<td>((0, 0; -1, 0))</td>
<td>improper</td>
<td>proper</td>
</tr>
<tr>
<td>Limiting reference prior (19)</td>
<td>(1/\delta_0\delta_1)</td>
<td>((0, 0; -\frac{1}{2}, 0))</td>
<td>improper</td>
<td>proper</td>
</tr>
</tbody>
</table>

If \( \text{sse} = 0 \) and \( b_0 = 0 \), the marginal density (23) of \( \eta \) near 0 is proportional to \( \eta^{-a_0} \). Thus, if \( a_0 > -1 \), the marginal posterior of \( \eta \) has a mode at 0. This means that the posterior is typically bimodal.

In our experience, the choice of improper prior is not overly critical. Simulation studies not reported here show that the constant prior on \((\delta_0^2, \delta_1^2)\) performs well, at least for smoothing splines on the line. It is common practice to recommend proper but diffuse priors. Clearly a proper prior with very large variance on the boundary conditions in the Bayesian set-up of § 2.1 will work since the posterior for the limiting case is proper. However, one must be careful with proper but diffuse priors on \( \delta_0 \) and \( \delta_1 \), since in some cases the limiting prior leads to an improper posterior.

4. **Bayesian computation**

In this section, we assume that the prior is given by (9) and (20), with parameters \((a_0, b_0; a_1, b_1)\) chosen so that the posterior is proper. The posterior mean of \( z \) can be computed via sample means of independently and identically distributed Monte Carlo realisations.
Algorithm 1 (Monte Carlo)
Step 1. Sample \( \eta \) from the density (23).
Step 2. For given \( \eta \), sample \( \delta_0 \) from \( \mathcal{I}G\{a_0 + a_1 + \frac{1}{2}(N - p), b_0 + \frac{1}{2}(\text{sse} + \bar{y}'H_{\eta}\bar{y})\} \).
Step 3. For given \( (\delta_0, \eta) \), sample \( z \) from \( N\{(W + \eta A)^{-1}W\bar{y}, \delta_0(W + \eta A)^{-1}\} \).

The marginal posterior density of \( \eta \) is not of a standard form. To avoid this distribution, a second algorithm using Gibbs sampling can be used. Note that, for given \( z, \delta_0 \) and \( \delta_1 \) are independent.

Algorithm 2 (Markov chain Monte Carlo)
Step 1. Sample \( (\delta_1 | z, y) \) from \( \mathcal{I}G\{a_1 + \frac{1}{2}(n - p), b_1 + \frac{1}{2}z'Az\} \).
Step 2. Sample \( (\delta_0 | z, y) \) from \( \mathcal{I}G\{a_0 + \frac{1}{2}N, b_0 + \frac{1}{2}\text{sse} + \sum_{i=1}^{n}r_i(\bar{y}_i - z_i)^2\} \).
Step 3. Sample \( (z | \bar{y}, \delta_0, \delta_1) \) from \( N\{(W + \eta A)^{-1}W\bar{y}, \delta_0(W + \eta A)^{-1}\} \).

Clearly the three kinds of conditional distribution are all standard. Note that the conditional mean of \( (z | \delta_0, \eta; y) \) in Step 3 is \( \tilde{g} = (W + \eta A)^{-1}W\bar{y} \), exactly the smoothing spline estimate with unequal variances. In the one-dimensional case, efficient calculation is possible. Carter & Kohn (1994) used state-space methods to compute both \( g \) and \( \bar{y} \) from the posterior efficiently. Their methods also provide efficient computation for Step 1. In the difference prior and some CAR models, the matrix \( (W + \eta A)^{-1} \) can be factored with a band Cholesky decomposition, so that \( O(n) \) operations are all that are needed to generate \( z \). This is the same order as the exact algorithm of Reinsch (1967) and the state-space algorithm of Kohn & Ansley (1987). Similarly, banded matrix methods can be used to draw efficiently from the full conditional distribution of \( z \) in Step 3. We are not aware of any efficient algorithms for Gibbs sampling for higher-dimensional splines such as thin-plate splines. Successful implementation is a topic for future study.

Implementation of Markov chain Monte Carlo requires a ‘burn-in’ period to achieve stationarity and a follow-up stage for estimation. In our experience, convergence is very rapid and robust to the choice of starting values; we see convergence within a few hundred cycles. Several thousand cycles then suffice for function and variance component estimation.

In some cases, it is also possible to sample \( z \) componentwise through the following alternative to Step 3 in Algorithm 2.

Step 3*. Sample \( (z_i | z_j, j \neq i, \delta_0, \delta_1, y) \) from \( N\{w_i(r_i\delta_0^{-1}y_i - 1/\delta_1\sum_{k \neq i}a_{ik}z_k), w_i\} \), where \( w_i = (r_i\delta_0^{-1} + a_{ii}\delta_1^{-1})^{-1} \).

This is attractive when the prior distribution for \( z \) is specified with a sparse precision matrix, such as an intrinsic autoregressive model. However, componentwise updating of \( z \) may not be efficient in running Gibbs sampling. Carter & Kohn (1994) use state-space methods to generate the \( z \) updates efficiently. They report that a Gibbs cycle that updates the entire vector \( z \) exhibits faster convergence than if the \( z_i \) are updated individually.

Acknowledgement

The research was supported by a grant from the National Science Foundation and a grant from the Missouri Department of Conservation. The authors gratefully acknowledge the constructive comments of the editor, an associate editor and an especially helpful referee.
REFERENCES


[Received August 2000. Revised June 2002]