Bayesian Analysis of Multivariate Smoothing Splines

Dongchu Sun, Department of Statistics, University of Missouri, Columbia, MO 65211, USA

Shawn Ni, Department of Economics, University of Missouri, Columbia, MO 65211, USA

Paul L. Speckman, Department of Statistics, University of Missouri, Columbia, MO 65211, USA

Abstract

A general version of multivariate smoothing splines with correlated errors and correlated curves is proposed. A suitable symmetric smoothing parameter matrix is introduced, and practical priors are developed for the unknown covariance matrix of the errors and the smoothing parameter matrix. An efficient algorithm for computing the multivariate smoothing spline is derived, which leads to an efficient Markov chain Monte Carlo method for Bayesian computation. Key to the computation is a natural decomposition of the estimated curves into components intrinsic to the problem that extend the notion of principal components. These intrinsic principal curves are useful both for computation and for interpreting the data. Numerical simulations show multivariate smoothing splines outperform univariate smoothing splines. The method is illustrated with analysis of a multivariate macroeconomic time series data set.

Keywords multivariate analysis, Bayesian analysis, smoothing splines, intrinsic principal curves

1 Introduction

Consider the problem of estimating latent smooth curves from a multivariate data set. The functional form of the curves and the distribution of the multivariate errors are unknown. In applications, it is quite common that the data-generating curves are co-moving and the errors correlated. Efficient estimation of the curves and the covariance of the errors requires joint estimation of all curves. For instance, to decompose multivariate macroeconomic time series data into unknown co-moving trends in the presence of correlated errors, the data of one variable are useful for estimating the trend of another variable. This study provides, for the first time, a simple Bayesian solution to this problem.

Formally, suppose multivariate observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$ are taken at points $\boldsymbol{t} = \{t_1 < \cdots < t_n\}$, where $-\infty < a \leq t_1$ and $t_n \leq b < \infty$. Without loss of generality, we can assume $a = 0$ and $b = 1$. In the corresponding spline smoothing problem, a vector-valued unknown function $\boldsymbol{g}(s) = (g_1(s), \ldots, g_p(s))$ is chosen to minimize the loss function with a penalty on roughness,

$$\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{g}(t_i))\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{y}_i - \boldsymbol{g}(t_i))' + \int_0^1 \boldsymbol{g}^{(k)}(s)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{g}^{(k)}(s))'ds, \tag{1}$$

where $\boldsymbol{g}^{(k)}(s) = (g_1^{(k)}(s), \ldots, g_p^{(k)}(s))$ is a vector of kth derivatives. In (1), $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are positive definite $p \times p$ penalty matrices on the approximation error and the "roughness" of $\boldsymbol{g}(t)$, respectively. Throughout the paper we also refer to them as covariance matrices. Using "tr$(\cdot)$" for trace, the loss function (1) can be rewritten as

$$\mathrm{tr}\left\{\boldsymbol{\Sigma}_0^{-1}\left[\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{g}(t_i))'(\boldsymbol{y}_i - \boldsymbol{g}(t_i)) + \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1^{-1}\int_0^1 (\boldsymbol{g}^{(k)}(s))'\boldsymbol{g}^{(k)}(s)ds\right]\right\}. \tag{2}$$

When $p = 1$, the multivariate spline becomes a univariate smoothing spline, where the smooth component $g(t)$ is chosen to minimize the loss function

$$\frac{1}{\sigma_0^2}\left\{\sum_{i=1}^{n}(y_i - g(t_i))^2 + \frac{\sigma_0^2}{\sigma_1^2}\int_0^1 [g^{(k)}(s)]^2 ds\right\}. \tag{3}$$

The noise-to-signal ratio $\eta = \sigma_0^2/\sigma_1^2$ is called the *smoothing parameter* and controls the balance between fidelity to the data and smoothness of the fitted function.

The problem of spline smoothing has been thoroughly studied for univariate models. See, for example, Wahba (1990), Green & Silverman (1994) or Eubank (1999). One intriguing property of smoothing splines is the fact that they can be interpreted as Bayes estimates with a suitable extended Gaussian process prior for fixed σ_0^2 and σ_1^2 (Kimeldorf & Wahba 1970). Moreover, Wahba (1985) and Wecker & Ansley (1983) showed that a univariate smoothing spline corresponds to a Bayesian linear mixed model and a state space model, respectively. These properties make a fully Bayesian approach to spline smoothing quite natural.

Several authors (e.g., Fessler (1991), Yee & Wild (1996), Wang et al. (2000)) have considered restricted versions of multivariate smoothing splines with multivariate dependent variables. These authors allowed the penalty matrix $\mathbf{\Sigma}_0$ to be treated as either known (including the case where $\mathbf{\Sigma}_0$ depends on i) or estimated as the covariance of residuals of univariate splines iteratively, but they restricted $\mathbf{\Sigma}_1$ to be diagonal. To our knowledge, the multivariate smoothing spline has not been treated with general $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$.

In this paper, we propose a fully Bayesian approach to fitting multivariate smoothing splines with general $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. To that end, we need priors on $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. Because it can be quite difficult to elicit informative priors, especially for $\mathbf{\Sigma}_1$, we propose a matrix version of the smoothing parameter, to be denoted by $\mathbf{\Xi}$, an objective noninformative prior on $\mathbf{\Sigma}_0$ and an informative prior on $\mathbf{\Xi}$.

We present the following results: (i) given $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ (or $\mathbf{\Sigma}_0$ and $\mathbf{\Xi}$), the minimizer of (1) exists and is a vector of natural spline functions, generalizing the univariate result; (ii) there are computationally efficient algorithms so that computing the solution to (1) is essentially only p times more costly than computing a univariate solution;

(iii) under the proposed priors on $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Xi}$, we develop a fully Bayesian procedure that can be estimated efficiently with MCMC; and (iv) we introduce a version of principle components based on decomposition of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Xi}$ that provides a basis for interpreting the fitted curves.

In Section 2, we treat the multivariate smoothing problem for fixed $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. We demonstrate the existence of a unique solution to (1) in Section 2.1, and we relate that solution to univariate spline smoothing in 2.2. We also develop a Bayesian linear model in which the latent curves are assigned correlated partially informative Gaussian priors in Section 2.4. With this model, we show in Section 2.3 the solution to (1) is exactly the posterior mean, generalizing the result of Kimeldorf & Wahba (1970). Finally, we introduce the concept of *intrinsic principle curves*, a functional basis of p smooth curves orthogonal with respect to an inner product defined by the problem, that decomposes the fitted curves in the manner of principle components in multivariate analysis. This decomposition is closely related to but differs from principal curves (Hastie & Stuetzle 1989) and the version of principal components developed in functional data analysis (e. g., Ramsay & Silverman 1997).

The Bayesian model specification presented here includes improper or partially improper priors. As a limit of proper priors, the Gaussian process prior on $\boldsymbol{g}(t)$ is partially improper. For full Bayesian analysis, we introduce priors in Section 3. The prior we advocate for $\boldsymbol{\Sigma}_0$ is a right Haar prior, which is noninformative and improper. A proof that the posterior is proper is will appear elsewhere (Sun et al. 2014). Section 4 is devoted to our algorithms for Bayesian computation. Some results from an extensive simulation study are presented in Section 5, showing situations in which multivariate smoothing can dominate univariate smoothing and also demonstrating that there may be little loss in efficiency using multivariate smoothing when univariate smoothing is appropriate. Finally, the method is demonstrated through analysis of an econometric data set analyzing and comparing trends in economic policy uncertainty in Section 6.

2 Multivariate Spline Smoothing

2.1 Existence and solution

It is well known that the minimizer of (3) lies in an n-dimensional space of natural spline functions (Schoenberg 1964). To implement the multivariate version, it's necessary to generalize this result to the multivariate case. To be precise, let $\mathcal{W}^{2,k}[0,1]$ denote the Sobolev space of functions $\{g \in L_2[0,1] : g, g', \ldots, g^{(k-1)}$ are absolutely continuous and $g^{(k)} \in L_2[0,1]\}$, so the minimizer of (2) is taken over the product space $\mathcal{W}_p^{2,k}[0,1] \equiv \mathcal{W}[0,1]^{2,k} \times \cdots \times \mathcal{W}^{2,k}[0,1]$. In addition, let $\mathcal{NS}^{2k}(\boldsymbol{t})$ denote the space of *natural smoothing splines* of order $2k$ with knot set $\boldsymbol{t} = \{t_1 < \cdots < t_n\}$. This space consists of all functions f such that (i) $f \in C^{2k-2}(\mathbb{R})$, (ii) $f^{(2k-1)}(s)$ and $f^{(2k)}(s)$ exist for all $s \notin \boldsymbol{t}$, (iii) $f^{(2k)}(s) = 0$ for all $s \notin \boldsymbol{t}$, and (iv) $f^{(k+j)}(t_1-) = f^{(k+j)}(t_n+) = 0$, $j = 0, \ldots, k-1$. In words, f is a natural spline if it is a polynomial of degree $2k-1$ between knots, $f^{(2k-2)}$ is a continuous, piecewise linear function, and f is a polynomial of degree $k-1$ for $s < t_1$ or $s > t_n$. Let $\mathcal{NS}_p^{2k}(\boldsymbol{t}) = \mathcal{NS}^{2k}(\boldsymbol{t}) \times \cdots \times \mathcal{NS}^{2k}(\boldsymbol{t})$. The next lemma, proved in the Appendix, extends a classical result for univariate smoothing splines.

Lemma 1 *The minimizer of (2) exists and lies in* $\mathcal{NS}_p^{2k}(\boldsymbol{t})$.

Now let $b_1(t), \ldots, b_n(t)$ be a basis of B-spline functions for $\mathcal{NS}_p^{2k}(\boldsymbol{t})$. In (1), the jth component of $\boldsymbol{g}$ can be written in terms of unknown parameters c_{ij} ($i = 1, \ldots, n; j = 1, \ldots, p$) as $g_j(t) = \sum_{i=1}^{n} c_{ij} b_i(t)$. For $h, l = 1, \ldots, n$, define $\kappa_{hl} = \int_0^1 b_h^{(k)}(s) b_l^{(k)}(s)\, ds$. Denote the $1 \times n$ row-vector of basis functions as $\boldsymbol{b}(t) = (b_1(t), \ldots, b_n(t))$, and define the matrices $\boldsymbol{C} = [c_{ij}]_{n\times p}$ and $\boldsymbol{K} = [\kappa_{hl}]_{n\times n}$. Then we can write $\boldsymbol{g}(t) = \boldsymbol{b}(t)\boldsymbol{C}$, $\boldsymbol{g}^{(k)}(t) = \boldsymbol{b}^{(k)}(t)\boldsymbol{C}$, $\int_0^1 \boldsymbol{g}^{(k)}(s)'\boldsymbol{g}^{(k)}(s)\,ds = \boldsymbol{C}'\boldsymbol{K}\boldsymbol{C}$.

The rank of matrix $\boldsymbol{K}$ is $n-k$. Let

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}_{n\times p}, \quad \boldsymbol{B} = \begin{pmatrix} b_1(t_1) & b_2(t_1) & \cdots & b_n(t_1) \\ b_1(t_2) & b_2(t_2) & \cdots & b_n(t_2) \\ \vdots & \vdots & \vdots & \vdots \\ b_1(t_n) & b_2(t_n) & \cdots & b_n(t_n) \end{pmatrix}_{n\times n}.$$

Then (2) is equivalent to

$$\min_{\boldsymbol{C}} \operatorname{tr}\left\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{C})'(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{C}) + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{C}'\boldsymbol{K}\boldsymbol{C}\right\}. \tag{4}$$

If we define

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{C} \text{ and } \boldsymbol{Q} = (\boldsymbol{B}^{-1})'\boldsymbol{K}\boldsymbol{B}^{-1}, \tag{5}$$

then (4) can be written as

$$\min_{\boldsymbol{Z}} \operatorname{tr}\left\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{Y}-\boldsymbol{Z})'(\boldsymbol{Y}-\boldsymbol{Z}) + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}\right\}. \tag{6}$$

Now let $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$ and $\boldsymbol{z} = \text{vec}(\boldsymbol{Z})$. Using the fact that

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}) = \text{vec}'(\boldsymbol{D})(\boldsymbol{A}\otimes\boldsymbol{C}')\text{vec}(\boldsymbol{B}') \tag{7}$$

for any conforming matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, (6) is equivalent to

$$\min_{\boldsymbol{z}} \left\{(\boldsymbol{y}-\boldsymbol{z})'(\boldsymbol{\Sigma}_0^{-1}\otimes\boldsymbol{I}_n)(\boldsymbol{y}-\boldsymbol{z}) + \boldsymbol{z}'(\boldsymbol{\Sigma}_1^{-1}\otimes\boldsymbol{Q})\boldsymbol{z}\right\}. \tag{8}$$

The solution to (8) is

$$\hat{\boldsymbol{z}} = (\boldsymbol{I}_{np} + \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1^{-1}\otimes\boldsymbol{Q})^{-1}\boldsymbol{y}. \tag{9}$$

The matrix $\boldsymbol{Q}$ in (5) is well known from the univariate smoothing spline literature, often in different notation. For example, it is denoted as $\boldsymbol{K}$ in Green & Silverman

(1994). When $k = 2$, for univariate cubic natural smoothing splines with equal spaced knots at $t = 1, 2, \ldots, n$, Shiller (1984) showed that $\boldsymbol{Q} = \boldsymbol{F}_0' \boldsymbol{F}_1^{-1} \boldsymbol{F}_0$, where

$$\boldsymbol{F}_0 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}_{(n-2)\times n}, \boldsymbol{F}_1 = \frac{1}{6} \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 4 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & 1 \\ 0 & 0 & 0 & \cdots & 1 & 4 \end{pmatrix}_{(n-2)\times(n-2)}.$$

A general formula for arbitrary $t_1 < \cdots < t_n$ is given in Green & Silverman (1994).

Alternatively, it's possible to use a discrete approximation to obtain an approximate solution using a band matrix $\boldsymbol{Q}$. For equally spaced points $t_1 < \cdots < t_n$, one can use $\boldsymbol{Q} = \boldsymbol{F}_0' \boldsymbol{F}_0$ for a cubic spline (Rue & Held 2005, p. 110).

The smoothing spline of order k has an important connection with linear polynomial regression of degree $k-1$. Consider the univariate case with smoothing parameter η,

$$\hat{\boldsymbol{z}} = (\boldsymbol{I}_n + \eta \boldsymbol{Q})^{-1} \boldsymbol{y}. \tag{10}$$

The matrix $\boldsymbol{Q}$ is known to have rank $n - k$ with null space spanned by $\{1, \ldots, t^{k-1}\}$. Thus there exists an orthogonal matrix $\boldsymbol{\Gamma} = [\boldsymbol{X}_0, \boldsymbol{X}_1]$ such that

$$\boldsymbol{Q} = \boldsymbol{\Gamma} \widetilde{\boldsymbol{\Lambda}} \boldsymbol{\Gamma}' = \boldsymbol{X}_1 \boldsymbol{\Lambda} \boldsymbol{X}_1', \tag{11}$$

where $\widetilde{\boldsymbol{\Lambda}} = \text{diag}(\boldsymbol{0}_{k\times k}, \boldsymbol{\Lambda})$ and $\boldsymbol{\Lambda}$ is diagonal. Clearly,

$$\boldsymbol{X}_0' \boldsymbol{X}_0 = \boldsymbol{I}_k, \;\; \boldsymbol{X}_1' \boldsymbol{X}_1 = \boldsymbol{I}_{n-k}, \;\; \boldsymbol{X}_0' \boldsymbol{X}_1 = \boldsymbol{0}_{k\times(n-k)}. \tag{12}$$

Also, $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ are $n \times k$ and $n \times (n-k)$ matrices corresponding to the k zero eigenvalues and $n - k$ positive eigenvalues of $\boldsymbol{Q}$, respectively. Then

$$\hat{\boldsymbol{z}} = \boldsymbol{\Gamma}(\boldsymbol{I}_n + \eta \widetilde{\boldsymbol{\Lambda}})^{-1} \boldsymbol{\Gamma}' \boldsymbol{y} = \boldsymbol{P}_0 \boldsymbol{y} + \boldsymbol{X}_1 (\boldsymbol{I}_{n-k} + \eta \boldsymbol{\Lambda})^{-1} \boldsymbol{X}_1' \boldsymbol{y}, \tag{13}$$

where $\boldsymbol{P}_0 = \boldsymbol{X}_0 \boldsymbol{X}_0'$. The first term on the right is the least squares polynomial fit of degree $k - 1$. The second term reflects the amount of smoothing and is controlled by

η. In the case $k = 2$, the cubic spline can be decomposed as the least squares line plus a smooth term. We will see that this property carries over to the multivariate case.

2.2 Connection with univariate spline smoothing

One central issue in defining the multivariate smoothing spline is to generalize the smoothing parameter η when $p = 1$ in (3) to the general case, where the analog is the matrix $\mathbf{\Sigma}_0\mathbf{\Sigma}_1^{-1}$ in (2). However, $\mathbf{\Sigma}_0\mathbf{\Sigma}_1^{-1}$ is not an ideal smoothing parameter matrix because it is not symmetric and it is overparameterized with p^2 parameters. A matrix version of the smoothing parameter should be symmetric with $p(p+1)/2$ free parameters. We reparameterize $(\mathbf{\Sigma}_0, \mathbf{\Sigma}_1)$ as follows. Suppose

$$\mathbf{\Sigma}_0^{-1} = \mathbf{\Psi}'\mathbf{\Psi}, \tag{14}$$

$$\mathbf{\Sigma}_1^{-1} = \mathbf{\Psi}'\mathbf{\Xi}\mathbf{\Psi}, \tag{15}$$

where $\mathbf{\Psi}$ is a $p \times p$ invertible matrix (perhaps with $p(p+1)/2$ free parameters) and $\mathbf{\Xi}$ is symmetric. The $p \times p$ positive definite matrix $\mathbf{\Xi}$ is a matrix version of the noise-to-signal ratio or smoothing parameter with $p(p+1)/2$ free parameters. When $p = 1$, $\mathbf{\Xi}$ is exactly the smoothing parameter σ_0^2/σ_1^2. For $p > 1$, decompositions (14) and (15) imply $\mathbf{\Xi} = \mathbf{\Psi}^{-T}\mathbf{\Sigma}_1^{-1}\mathbf{\Psi}^{-1}$, where $\mathbf{\Psi}^{-T} = (\mathbf{\Psi}')^{-1}$, and $\mathbf{\Sigma}_0\mathbf{\Sigma}_1^{-1} = \mathbf{\Psi}^{-1}\mathbf{\Xi}\mathbf{\Psi}$. With this definition, solution (9) becomes

$$\hat{\boldsymbol{z}} = (\mathbf{\Psi}^{-1} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np} + \mathbf{\Xi} \otimes \boldsymbol{Q})^{-1}(\mathbf{\Psi} \otimes \boldsymbol{I}_n)\boldsymbol{y}.$$

Suppose

$$\mathbf{\Xi} = \boldsymbol{OHO}', \tag{16}$$

where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{H} = \mathrm{diag}(\eta_1, \ldots, \eta_p)$. Define

$$\mathbf{\Delta} = \boldsymbol{O}'\mathbf{\Psi}. \tag{17}$$

Then (14) and (15) imply

$$\begin{aligned} \boldsymbol{\Sigma}_0^{-1} &= \boldsymbol{\Delta}'\boldsymbol{\Delta}, \qquad (18) \\ \boldsymbol{\Sigma}_1^{-1} &= \boldsymbol{\Delta}'\boldsymbol{H}\boldsymbol{\Delta}, \qquad (19) \end{aligned}$$

hence (16) becomes

$$\widehat{\boldsymbol{z}} = (\boldsymbol{\Delta}^{-1} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1}(\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)\boldsymbol{y}. \qquad (20)$$

For the rest of the paper, it's important to differentiate between the rows and columns of matrices like $\boldsymbol{Y}$ and $\boldsymbol{Z}$. As customary with multivariate analysis, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ denote row vectors as in (1). On the other hand, it's also important to label the columns of $\boldsymbol{Y}$ as they represent data associated with the p separate smooth curves. We will denote such column vectors as $\boldsymbol{y}_j^*$, $\boldsymbol{z}_j^*$, etc. Thus $\boldsymbol{Y} = [\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_p^*]$, $\boldsymbol{y} = \text{vec}([\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_p^*])$, $\boldsymbol{Z} = [\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_p^*]$, etc. (Note that $\boldsymbol{y}$ and $\boldsymbol{z}$ with no subscript denote vectors of length np.)

The fact that $\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q}$ is block diagonal allows us to interpret (20) in terms of p univariate smoothing splines. Let

$$\boldsymbol{u} = (\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)\boldsymbol{y}, \boldsymbol{v} = (\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)\boldsymbol{z}. \qquad (21)$$

Using the fact that $\text{vec}(\boldsymbol{AXB}) = (\boldsymbol{B}' \otimes \boldsymbol{A})\text{vec}(\boldsymbol{X})$, we have $(\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)\boldsymbol{y} = \text{vec}(\boldsymbol{Y}\boldsymbol{\Delta}')$. Define

$$\boldsymbol{U} = [\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_p^*] = \boldsymbol{Y}\boldsymbol{\Delta}', \quad \boldsymbol{V} = [\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_p^*] = \boldsymbol{Z}\boldsymbol{\Delta}'. \qquad (22)$$

If we let $\boldsymbol{u} = \text{vec}(\boldsymbol{U})$ and $\widehat{\boldsymbol{v}} = (\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1}\boldsymbol{u} = \text{vec}([\widehat{\boldsymbol{v}}_1^*, \ldots, \widehat{\boldsymbol{v}}_p^*])$, then (20) implies

$$\widehat{\boldsymbol{v}}_j^* = (\boldsymbol{I}_n + \eta_j\boldsymbol{Q})^{-1}\boldsymbol{u}_j^*, \quad j = 1, \ldots, p. \qquad (23)$$

Finally, let $\widehat{\boldsymbol{z}} = \text{vec}(\widehat{\boldsymbol{Z}})$. Using (20) again,

$$\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{V}}\boldsymbol{\Delta}^{-T}. \qquad (24)$$

Thus the multivariate smoothing spline formula (9) is equivalent to solving (22), (23), and (24). Equations (22)-(23) have both computational and practical significance. If $\boldsymbol{\Delta}$ is known, one can transform $\boldsymbol{y}$ by (22), do univariate smoothing on the $\boldsymbol{u}_j^*$, and transform back to get $\widehat{\boldsymbol{z}}$. This avoids inverting the $np \times np$ matrix $(\boldsymbol{I}_{np} + \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q})$ and only requires p solutions of the n-dimensional problem (23). In addition, the $\boldsymbol{v}_j^*$ coordinates may be natural to the problem and suggest an interpretation similar to principal components. Thus the $\boldsymbol{\Delta}'$ transformation is fundamental to multivariate spline smoothing.

Although the construction of $\boldsymbol{\Delta}$ appears to depend on the specific factorization used in (14), it turns out that $\boldsymbol{\Delta}$ is essentially invariant with respect to this factorization. From (18), $\boldsymbol{\Delta}\boldsymbol{\Sigma}_0\boldsymbol{\Delta}' = \boldsymbol{I}$, hence $\boldsymbol{\Sigma}_0 = \boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^{-T}$, and from (19),

$$\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Delta}^{-1} \boldsymbol{H} \boldsymbol{\Delta}. \tag{25}$$

Equivalently, $\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Delta}^{-1} = \boldsymbol{\Delta}^{-1} \boldsymbol{H}$, which implies that the columns of $\boldsymbol{\Delta}^{-1}$ are the eigenvectors of $\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1^{-1}$, and the diagonal elements of the diagonal matrix $\boldsymbol{H}$ are the eigenvalues of $\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1^{-1}$. Since eigenvectors are essentially unique, this proves that $\boldsymbol{\Delta}$ is essentially independent of the specific factorization $\boldsymbol{\Psi}$ in (14). Moreover, (25) provides a direct interpretation linking (9) with (20).

Finally, equation (13) shows the intimate connection between univariate spline smoothing and polynomial regression. To see that this carries over to the multivariate case, consider representation (20). Since $\boldsymbol{H} = \text{diag}(\eta_1, \ldots, \eta_p)$,

$$\begin{aligned}
(\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1} &= \text{diag}(\ldots, (\boldsymbol{I}_n + \eta_j \boldsymbol{Q})^{-1}, \ldots) \\
&= \text{diag}(\ldots, \boldsymbol{P}_0 + \boldsymbol{X}_1 (\boldsymbol{I}_{n-k} + \eta_j \boldsymbol{\Lambda})^{-1} \boldsymbol{X}_1', \ldots) \\
&= \boldsymbol{I}_p \otimes \boldsymbol{P}_0 + (\boldsymbol{I}_p \otimes \boldsymbol{X}_1)(\boldsymbol{I}_{p(n-k)} + \boldsymbol{H} \otimes \Lambda)^{-1} (\boldsymbol{I}_p \otimes \boldsymbol{X}_1').
\end{aligned}$$

Thus from (20),

$$\widehat{\boldsymbol{z}} = (\boldsymbol{I}_p \otimes \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{\Delta}^{-1} \otimes \boldsymbol{X}_1)(\boldsymbol{I}_{p(n-k)} + \boldsymbol{H} \otimes \Lambda)^{-1} (\boldsymbol{\Delta} \otimes \boldsymbol{X}_1')\boldsymbol{y}. \tag{26}$$

The first term on the right is exactly the least squares polynomial fit to each of the p data sets.

2.3 A Bayesian smoothing model for fixed $(\mathbf{\Sigma}_0, \mathbf{\Sigma}_1)$

It is well known that the univariate smoothing spline problem arises naturally in a Bayesian context. Suppose

$$y_i = g(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{27}$$

where the ϵ_i are independent $N(0, \sigma_0^2)$ random variables, and

$$g(t) = \beta_0 + \beta_1 t + \cdots + \beta_{k-1} t^{k-1} + g_0(t),$$

with a flat (improper) prior on the β_j and a suitable Gaussian process prior on g_0. For example, let

$$g_0(t) = \sigma_1 \int_0^1 \frac{(t-u)^{k-1}}{(k-1)!} \, dW(u),$$

where $dW(u)$ is standard Gaussian white noise. Thus, for $k = 1$, the prior on g_0 is scaled Brownian motion, for $k = 2$, the prior is the integral of scaled Brownian motion, etc. After some manipulation, it can be shown that this prior can be represented as follows. Define the *reproducing kernel*

$$R(s,t) = \int_0^1 \frac{(s-u)^{k-1}}{(k-1)!} \frac{(t-u)^{k-1}}{(k-1)!} \, du, \quad 0 \le s, t \le 1,$$

and let $\boldsymbol{R} = [R(t_i, t_j)]_{n \times n}$. Then $\sigma_1^2 \boldsymbol{R}$ is the covariance matrix of the prior on $(g_0(t_1), \ldots g_0(t_n))'$. Let $\boldsymbol{P}_0$ be the projection matrix in $\mathbb{R}^n$ onto the span of $1, t, \ldots, t^{k-1}$. It can be shown that the matrix $\boldsymbol{Q}$ in (5) has the alternate representation

$$\boldsymbol{Q} = (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{R}(\boldsymbol{I} - \boldsymbol{P}_0)$$

(e. g., Wahba 1990). Setting $\boldsymbol{z} = (g(t_1), \ldots, g(t_n))'$, this partially informative Bayes prior can be shown to have the partially improper pdf $p(\boldsymbol{z} \mid \sigma_1) \propto \sigma_1^{-(n-k)} \exp\left(-\frac{1}{2\sigma_1^2} \boldsymbol{z}'\boldsymbol{Q}\boldsymbol{z}\right)$

(see, e. g., Speckman & Sun 2003). Expressing (27) in the vector notation $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{\epsilon}$, the posterior of $\boldsymbol{z}$ satisfies

$$f(\boldsymbol{z} \mid \boldsymbol{y}, \sigma_0, \sigma_1) \propto \sigma_0^{-n} \sigma_1^{-(n-k)} \exp\left(-\frac{1}{2\sigma_0^2}\|\boldsymbol{y} - \boldsymbol{z}\|^2 - \frac{1}{2\sigma_1^2}\boldsymbol{z}'\boldsymbol{Q}\boldsymbol{z}\right).$$

From this expression, it's easy to show that the posterior distribution of $\boldsymbol{z}$ is multivariate normal with mean $\widehat{\boldsymbol{z}} = (\boldsymbol{I} + \eta\boldsymbol{Q})^{-1}\boldsymbol{y}$, where $\eta = \sigma_0^2/\sigma_1^2$. Thus the smoothing spline is a Bayes estimate under a partially improper integrated Brownian motion prior on g.

This argument carries over directly to the multivariate case. Suppose

$$y_{ij} = g_j(t_i) + \epsilon_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, p. \tag{28}$$

For notational simplicity, we write $\boldsymbol{z}_i = (g_1(t_i), \ldots, g_p(t_i))$. With this notation, stacking the row vectors $\boldsymbol{z}_i$ defines $\boldsymbol{Z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_n')'$. The vector form of the observations now can be written as

$$\boldsymbol{y}_i = \boldsymbol{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{29}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})'$ and we assume independent correlated errors $\boldsymbol{\epsilon}_i' \sim N(0, \boldsymbol{\Sigma}_0)$. The density (likelihood) of $\boldsymbol{y}$ given $\boldsymbol{z}$ and $\boldsymbol{\Sigma}_0$ based on model (29) is

$$f(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\Sigma}_0) = (2\pi)^{-\frac{np}{2}}|\boldsymbol{\Sigma}_0|^{-\frac{n}{2}} \exp\Big\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{z})'(\boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{I}_n)(\boldsymbol{y} - \boldsymbol{z})\Big\}. \tag{30}$$

Analogous to the one-dimensional case, suppose $g_j(s) = \sum_{\ell=0}^{k-1} \beta_{j\ell} s^\ell + g_{j0}(s)$, $j = 1, \ldots, p$, where

$$\boldsymbol{g}_0(s) = \begin{pmatrix} g_{10}(s) \\ \vdots \\ g_{1p}(s) \end{pmatrix} = \boldsymbol{\Sigma}_1^{1/2} \begin{pmatrix} \tilde{g}_{10}(s) \\ \vdots \\ \tilde{g}_{p0}(s) \end{pmatrix} \tag{31}$$

and

$$\tilde{g}_{j0}(t) = \int_0^1 \frac{(t-u)^{k-1}}{(k-1)!}\, dW_j(u), \quad 0 \le t \le 1, \tag{32}$$

for independent Gaussian white noise processes $dW_j(u), j = 1, \dots, p$. Again, assuming flat priors $[\beta_{j\ell}] \propto 1$ and following the arguments in Speckman & Sun (2003), it can be shown that this partially improper prior on the multivariate function $\boldsymbol{g}(t)$ induces a partially improper distribution on the stacked state vector of length np, $\boldsymbol{z}' = (g_1(t_1), \dots, g_p(t_n))'$, at the points $t_1 < \cdots < t_n$ with density of the form

$$f(\boldsymbol{z} \mid \boldsymbol{\Sigma}_1) \propto \left|\left(\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q}\right)\right|_+^{1/2} \exp\Big\{-\frac{1}{2}\boldsymbol{z}'(\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q})\boldsymbol{z}\Big\}, \tag{33}$$

where $|\boldsymbol{A}|_+$ is the product of positive eigenvalues of a nonnegative definite matrix $\boldsymbol{A}$. Theorem 1 is the multivariate version of the Kimeldorf-Wahba theorem (Kimeldorf & Wahba 1970). For fixed $(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$, the solution of smoothing spline (9) coincides with the posterior mean of $\boldsymbol{z}$ under the prior (33). The routine proof of the resulting theorem is omitted.

Theorem 1 *Consider model (28) or (30) with prior* (33). *For fixed* $(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$, *the conditional posterior distribution of* $\boldsymbol{z}$ *given* $\boldsymbol{y}$ *is*

$$(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) \quad \sim \quad N_{pn}(\widehat{\boldsymbol{z}}, \boldsymbol{\Omega}^{-1}), \tag{34}$$

where $\widehat{\boldsymbol{z}}$ *is given by (9) and* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{I}_n + \boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q}$.

2.4 A formal Bayesian linear mixed model

We denote the positive eigenvalues of the nonnegative definite matrix $\boldsymbol{Q}$ as $0 < \lambda_1 < \cdots < \lambda_{n-k}$. So $|\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q}|_+ = |\boldsymbol{\Sigma}_1|^{-(n-k)}|\boldsymbol{\Lambda}|^p$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_{n-k})$. Define

$$c_0 = (2\pi)^{-\frac{(n-k)p}{2}}|\boldsymbol{\Lambda}|^{\frac{p}{2}}. \tag{35}$$

Then (33) becomes

$$f(\boldsymbol{z} \mid \boldsymbol{\Sigma}_1) \propto c_0|\boldsymbol{\Sigma}_1|^{-\frac{n-k}{2}} \exp\Big\{-\frac{1}{2}\boldsymbol{z}'(\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q})\boldsymbol{z}\Big\}. \tag{36}$$

Using the definition of $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ after (11), we have the following.

Lemma 2 *Let* $\boldsymbol{\Theta}$ *and* $\boldsymbol{W}$ *denote* $k \times p$ *and* $(n-k) \times p$ *random matrices, respectively. Write* $\boldsymbol{\theta} = vec(\boldsymbol{\Theta})$ *and* $\boldsymbol{w} = vec(\boldsymbol{W})$*. Assume that*

$$p(\boldsymbol{\theta}) \propto 1 \text{ and } (\boldsymbol{w} \mid \boldsymbol{\Sigma}_1) \sim N_{(n-k)p}(\mathbf{0}, \boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Lambda}^{-1}), \text{ and define} \tag{37}$$

$$\boldsymbol{Z} = \boldsymbol{X}_0\boldsymbol{\Theta} + \boldsymbol{X}_1\boldsymbol{W} = (\boldsymbol{X}_0, \boldsymbol{X}_1)\begin{pmatrix} \boldsymbol{\Theta} \\ \boldsymbol{W} \end{pmatrix}. \tag{38}$$

Then the improper prior density of $\boldsymbol{z} = vec(\boldsymbol{Z})$ *has the form (36).*

Proof. It follows from the fact that $\boldsymbol{z}'(\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{Q})\boldsymbol{z} = \boldsymbol{w}'(\boldsymbol{\Sigma}_1^{-1} \otimes \boldsymbol{\Lambda})\boldsymbol{w}$. □

2.5 Intrinsic principal curves for multivariate smoothing

With the prior of Lemma 2, the decomposition $\boldsymbol{V} = [\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_p^*] = \boldsymbol{Z}\boldsymbol{\Delta}'$ has a natural interpretation. Heuristically, since $\boldsymbol{g}'\boldsymbol{Q}\boldsymbol{g} = \int [g^{(k)}(t)]^2 \, dt$ for any natural spline $\boldsymbol{g} = (g(t_1), \ldots, g(t_n))'$, one would expect that the prior specification (31-32) implies

$$E\left[\int g_i^{(k)}(t) g_j^{(k)}(s)\, ds\right] \propto \sigma_{1ij}, \quad 1 \leq i, j \leq p,$$

where $\boldsymbol{\Sigma}_1 = [\sigma_{1ij}]_{p \times p}$. This argument is made rigorous in the following theorem, which also shows that the $\boldsymbol{v}_j^*$ have a natural orthogonality property. Thus $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Delta}^{-T}$ is a kind of principle components decomposition of the signal $\boldsymbol{Z}$. We term the columns of $\boldsymbol{V}$ as *intrinsic principal curves.*

Theorem 2 *If* $\boldsymbol{Z}$ *has prior (33), then*

$$E[\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}] = (n-k)\boldsymbol{\Sigma}_1. \tag{39}$$

Moreover, if $\boldsymbol{\Delta}$ *satisfies (18)-(19), then*

$$E[\boldsymbol{V}'\boldsymbol{Q}\boldsymbol{V}] = (n-k)\boldsymbol{H}^{-1}. \tag{40}$$

Proof. Lemma 2 implies that $\boldsymbol{W}$ follows the matrix normal distribution $N_{(n-k)\times p}(\mathbf{0}, \boldsymbol{\Lambda}^{-1}, \boldsymbol{\Sigma}_1)$ if $\boldsymbol{Z}$ has prior (33). Using a property of matrix normal distributions (e. g., Gupta & Nagar 2000), we have $E[\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}'] = [tr(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}')]\boldsymbol{\Sigma}_1 = (n-k)\boldsymbol{\Sigma}_1$. Lastly, (12) and (38) imply $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}'$. Thus (39) holds. Moreover, (39) implies $E[\boldsymbol{V}'\boldsymbol{Q}\boldsymbol{V}] = \boldsymbol{\Delta}E[\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}]\boldsymbol{\Delta}' = (n-k)\boldsymbol{\Delta}\boldsymbol{\Sigma}_1\boldsymbol{\Delta}'$. But from (19), $\boldsymbol{\Delta}\boldsymbol{\Sigma}_1\boldsymbol{\Delta}' = \boldsymbol{H}^{-1}$, proving (40). □

In principle, one could attempt two kinds of principle components analysis on the data matrix $\boldsymbol{Y}$. Traditional PCA treats the rows $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ as a random sample of vectors, while functional data analysis treats the columns $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_p^*$ as a random sample of *functional data* of size p. Since both the rows and columns of $\boldsymbol{Y}$ are correlated, neither approach is appropriate. However, intrinsic principle curves are closely related to one approach to functional data analysis (e. g., Ramsay & Silverman 1997). A covariance matrix $\boldsymbol{R}_{n\times n}$ for the columns of $\boldsymbol{Y}$ is estimated. Since the problem is typically quite ill-posed (often with $p < n$), some form of regularization is needed. The functional data are projected onto "smoothed principal components" of $\boldsymbol{R}$ for data reduction. In this way, high-dimensional functional data can be reduced to a few coefficients. Although our analysis with intrinsic principle curves can produce similar results, the method is fundamentally different because we assume the columns of $\boldsymbol{Y}$ are correlated via the covariance matrix $\boldsymbol{\Sigma}_1$. Intrinsic principal curves implicitly make use of the estimated correlations among the curves.

Another related technique is the method of principal curves introduced by Hastie & Stuetzle (1989). They proposed a technique for passing a smooth curve through p-dimensional data. Their method is purely descriptive and tacitly assumes $\boldsymbol{\Sigma}_0$ is diagonal.

There is a close connection between multivariate smoothing splines and spatio-temporal models (see Cressie & Wikle 2011). These models pertain to dependent sets of time series or stochastic processes observed at different geographical locations.

The setup is similar to the model here, but spatio-temporal models assume a spatial correlation model for each data vector $\boldsymbol{y}_i$, and the error variance $\boldsymbol{\Sigma}_0$ is generally taken to be diagonal. In our models, there is no possible geographic structure that can be used to simplify $\boldsymbol{\Sigma}_1$.

3 Fully Bayesian Analysis: a Prior for $(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$

3.1 A noninformative prior on $\boldsymbol{\Sigma}_0$

One way to choose a prior for $(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$ is with independent (perhaps inverse-Wishart) priors. The inverse-Wishart distribution for a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$, denoted by $\mathrm{IW}_p(m, \boldsymbol{A})$, has density

$$\pi(\boldsymbol{\Sigma} \mid m, \boldsymbol{A}) \propto |\boldsymbol{\Sigma}|^{-\frac{m+p+1}{2}} \mathrm{etr}\Big(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\Big),$$

where $\mathrm{etr}(\cdot)$ stands for $\exp[\mathrm{tr}(\cdot)]$. In this formulation, m is often interpreted as degrees of freedom and $\boldsymbol{A}$ is a known nonnegative definite matrix. If $m > p-1$ and $\boldsymbol{A}$ is positive definite, the prior distribution of $\boldsymbol{\Sigma}$ is proper.

Suppose $\boldsymbol{\Sigma}_0$ has an $\mathrm{IW}_p(m_0, \boldsymbol{Q}_0)$ prior. If $\boldsymbol{\Psi}$ satisfies (14) and $\boldsymbol{\Psi}$ is lower triangular, $\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}^{-T}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_0$. The corresponding prior on $\boldsymbol{\Psi}$ is

$$\pi(\boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}'\boldsymbol{\Psi}|^{\frac{m_0-p-1}{2}} \mathrm{etr}(-\frac{1}{2}\boldsymbol{Q}_0^{-1}\boldsymbol{\Psi}'\boldsymbol{\Psi}) \prod_{j=1}^{p} \psi_{jj}^{-j} = \prod_{j=1}^{p} \psi_{jj}^{m_0-p-1-j} \mathrm{etr}(-\frac{1}{2}\boldsymbol{Q}_0^{-1}\boldsymbol{\Psi}'\boldsymbol{\Psi}).$$

If $m_0 = p+1$ and $\boldsymbol{Q}_0^{-1} \to \boldsymbol{0}$, the prior for $\boldsymbol{\Sigma}_0$ approaches the right Haar prior

$$\boldsymbol{\Psi} \text{ is lower triangular and } \pi_{RH}(\boldsymbol{\Psi}) \;\propto\; \prod_{j=1}^{p} \frac{1}{\psi_{jj}^{j}}, \tag{41}$$

where ψ_{jj} is the jth diagonal element of $\boldsymbol{\Psi}$. For an i. i. d. $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ population, Berger & Sun (2008) showed that this right Haar prior is a matching prior. We propose the independent RH prior (41) for $\boldsymbol{\Sigma}_0$. Note that in the case of the univariate model

$p = 1$, (41) is equivalent to $\pi(\sigma_0^2) \propto 1/\sigma_0^2$, which is also the Jeffreys prior for the univariate case.

3.2 A generalized Pareto prior on Ξ

It's becoming increasingly popular to use a Pareto prior in the context of Zellner's g-prior (e. g., Liang et al. 2008). The parameter g is analogous to the smoothing parameter Ξ here. Given a scale parameter $b > 0$, the Pareto prior has the density

$$\pi(\eta \mid b) = b/(\eta + b)^2, \ \eta > 0. \tag{42}$$

We propose a proper multivariate analogue of the form

$$\pi(\Xi \mid b) \ = \ \frac{b^{\frac{(p+1)p}{2}}\Gamma_p(p+1)}{\left(\Gamma_p(\frac{p+1}{2})\right)^2}|\Xi + b\boldsymbol{I}_p|^{-(p+1)}, \ \Xi > 0, \tag{43}$$

where again $b > 0$ is a scale parameter and $\Gamma_p(\frac{n}{2}) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma(\frac{n}{2} - \frac{j-1}{2})$ for any $n > p$. This distribution has several attractive properties as a prior on Ξ. It is heavy-tailed so that the posterior distribution is not overly influenced by the prior. This is especially important for components where η_j is large, corresponding to almost linear fits. Moreover, there is a simple hierarchical model for this distribution, making it convenient for Bayesian computation.

It is well known that the Pareto distribution is the distribution of U/V, where U and V are independent exponential random variables with $[u] = e^{-u}$, $u > 0$ and $[v] = be^{-bv}$, $v > 0$ (here the scalar random variables U and V are not to be confused with matrices in bold letters in other sections.) A special case of the multivariate Feller-Pareto distribution is obtained by taking independent gamma(1) variables U_j, $j = 1, \ldots, p$ and independent $V \sim$ gamma(b). Then $(U_1/V, \ldots, U_p/V)$ has a multivariate Feller-Pareto distribution (e. g. Arnold 1983). The next lemma shows that $\pi(\Xi \mid b)$ has a similar hierarchical derivation, hence it is a proper distribution and is a matrix extension of the Pareto distribution. Moreover, it has a useful conditional property.

Lemma 3 *Assume* $(\Xi \mid \mathbf{\Phi}) \sim \text{Wishart}_p(p+1, \mathbf{\Phi}^{-1})$ *and* $\mathbf{\Phi} \sim \text{Wishart}_p(p+1, b^{-1}\boldsymbol{I}_p)$.

(a) The conditional distribution of $(\mathbf{\Phi} \mid \Xi)$ *is* $\text{Wishart}_p(2(p+1), (\Xi + b\boldsymbol{I}_p)^{-1})$.

(b) The marginal density of Ξ *has the form (43).*

The proof of the lemma is in the Appendix.

Care must be taken with improper priors to ensure that the posterior is proper. The problem is well-studied in univariate mixed linear models (e g., Hill 1965, Hobert & Casella 1996). The authors have extended results of Sun et al. (1999) and Sun & Speckman (2008) to the present case. Under model (28) or (30) with prior (37) and parametrization $(\mathbf{\Psi}, \Xi)$ given by (14)-(15) with right-Haar prior (41) on $\mathbf{\Psi}$ and $n > p + 1$, the posterior is proper if and only if the prior on Ξ is proper. Hence the posterior $(\boldsymbol{Z}, \Xi, \mathbf{\Psi} \mid \boldsymbol{Y})$ is proper under the generalized Pareto prior (43) for our multivariate smoothing spline model. Details will appear elsewhere (Sun et al. 2014).

3.3 Eliciting the hyperparameter b

The solution adopted by White (2006) and Cheng & Speckman (2013) for eliciting the prior for the univariate smoothing parameter η is based on the effective degrees of freedom of the smoother. From (23), the smoother matrix for the univariate smoothing spline is $\boldsymbol{S}_\eta = (\boldsymbol{I}_n + \eta\boldsymbol{Q})^{-1}$. Hastie & Tibshirani (1999) defined the effective degrees of freedom for a nonparametric linear smoother of the form $\boldsymbol{S}\boldsymbol{y}$ as $\text{tr}(\boldsymbol{S})$, extending the definition of degrees of freedom of the fit in a linear model. White (2006) argued that considering the prior distribution of effective degrees of freedom is a meaningful way to elicit prior information about η. In particular, let

$$\text{edf}(\eta) = \text{tr}(\boldsymbol{S}_\eta) = \sum_{i=1}^{n} \frac{1}{1 + \eta\lambda_i}, \tag{44}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\boldsymbol{Q}$. Since the trace is a monotonic function of η, the median of the distribution of $\text{edf}(\eta)$ is $\text{edf}(\tilde{\eta})$, where $\tilde{\eta}$ is the median of the prior distribution of η. Thus one can choose the parameters of a prior distribution on η to have median prior effective degrees of freedom consistent with prior belief in the complexity of the curves to be fitted. Complexity can be envisioned as the number of parametric terms needed to fit the curve in a regression model. Of course, the complexity depends on the amount of noise in the data as well as the number of observations. Less noise or more observations will admit a more complex fit.

For multivariate smoothing, our experience suggests that the best way to apply prior information is in terms of the smallest of the η_j, say η_1. This corresponds to the most complex component in the fitted spline. For general p, the marginal prior of η_1 under the multivariate (43) doesn't appear to be tractable, but it can be estimated easily by Monte Carlo simulation using the hierarchical scheme of Lemma 3(b). Since b is a scale parameter, we use simulation to estimate the median of the distribution of η_1, say $\tilde{\eta}_1$, for $b = 1$ and use (44) to solve $\text{edf}(b\tilde{\eta}_1) = \tilde{d}$ for b, where $\tilde{d}$ is the desired prior median degrees of freedom. For large p, there is some evidence that this choice of prior may oversmooth in some cases, and it may be preferable to elicit prior information on several other components from (23), for example η_1 and η_2. Future research will shed light on the problem. In the applications considered here, specifying the prior on η_1 alone appears to be satisfactory.

4 Bayesian Computation

Under the proposed priors, the joint posterior $(\boldsymbol{Z}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \mid \boldsymbol{Y})$ is not a standard distribution, but we can use MCMC simulation (see Gelfand & Smith 1990).

4.1 Efficient full conditional for $\boldsymbol{z}$

The following algorithm efficiently computes the multivariate spline. Recall from Theorem 1 that the full conditional distribution of $\boldsymbol{z}$ is $(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) \sim N_{pn}(\widehat{\boldsymbol{z}}, \boldsymbol{\Omega}^{-1})$, where $\widehat{\boldsymbol{z}}$ is given by (20) and $\boldsymbol{\Omega} = (\boldsymbol{\Delta}' \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1}(\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)$. From (21) we have

$$(\boldsymbol{v} \mid \boldsymbol{y}, \boldsymbol{\Delta}, \boldsymbol{H}) \sim N_{np}(\hat{\boldsymbol{v}}, (\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1}), \tag{45}$$

where

$$\hat{\boldsymbol{v}} = (\boldsymbol{I}_{np} + \boldsymbol{H} \otimes \boldsymbol{Q})^{-1}\boldsymbol{u} = (\boldsymbol{\Delta} \otimes \boldsymbol{I}_n)\widehat{\boldsymbol{z}}. \tag{46}$$

Thus $\boldsymbol{z} = (\boldsymbol{\Delta}^{-1} \otimes \boldsymbol{I}_n)\boldsymbol{v}$ has the posterior distribution (34). But $\boldsymbol{v}$ can be calculated efficiently since the covariance matrix $(\boldsymbol{I}_{np}+\boldsymbol{H}\otimes\boldsymbol{Q})^{-1}$ is block diagonal. In particular, (46) implies that (45) is equivalent to

$$\boldsymbol{v}_j^* \overset{ind}{\sim} N_n(\widehat{\boldsymbol{v}}_j^*, (\boldsymbol{I}_n + \eta_j \boldsymbol{Q})^{-1}), \quad j = 1, \ldots, p,$$

where from (46), $\hat{\boldsymbol{v}}_j^* = (\boldsymbol{I}_n + \eta_j \boldsymbol{Q})^{-1}\boldsymbol{u}_j^*$, $j = 1, \ldots, p$. Now assume a factorization of the form $(\boldsymbol{I}_n + \eta_j \boldsymbol{Q}) = \boldsymbol{R}_j' \boldsymbol{R}_j$. If $\boldsymbol{Q}$ is banded or sparse, one can use the Cholesky decomposition for $\boldsymbol{R}_j$, which can be computed efficiently. Otherwise, take the spectral decomposition for $\boldsymbol{Q}$ in (11). Then $\boldsymbol{R}_j = (\boldsymbol{I}+\eta_j\tilde{\boldsymbol{\Lambda}})^{1/2}\boldsymbol{\Gamma}'$ will work. (Note that $\boldsymbol{I}+\eta_j\tilde{\boldsymbol{\Lambda}}$ is a diagonal matrix with diagonal elements $(1 + \eta_j\lambda_i)$, where λ_i is the ith diagonal element of $\tilde{\boldsymbol{\Lambda}}$.)

With this notation, the following algorithm generates a single sample from (34).

1. Compute $\boldsymbol{U} = [\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_p^*] = \boldsymbol{Y}\boldsymbol{\Delta}'$.

2. Compute $\boldsymbol{v}_j^{(1)} = \boldsymbol{R}_j^{-T}\boldsymbol{u}_j^*$, $j = 1, \ldots, p$.

3. Compute $\boldsymbol{v}_j^* = \boldsymbol{R}_j^{-1}(\boldsymbol{v}_j^{(1)} + \boldsymbol{\varepsilon}_j)$, $j = 1, \ldots, p$, where $\boldsymbol{\varepsilon}_j \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, and set $\boldsymbol{V} = [\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_p^*]$.

4. Compute $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Delta}^{-T}$, and set $\boldsymbol{z} = \text{vec}(\boldsymbol{Z})$.

Remark 1 *It takes $O(np^2)$ operations (i. e., a floating point multiply followed by an addition) to calculate $\boldsymbol{U}$ and $\boldsymbol{Z}$ in steps 1 and 4. If $\boldsymbol{Q}$ is banded with bandwidth ℓ, then the Cholesky factorization $\boldsymbol{R}_j$ can be computed with $O(n\ell^2)$ operations, and steps (2) and (3) each take $O(n\ell)$ operations. Thus it takes $O(n\ell^2 p)$ operations to update $\boldsymbol{Z}$. In the general case where $\boldsymbol{Q}$ is not banded or sparse, the factorization $\boldsymbol{Q} = \boldsymbol{\Gamma}\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}'$ need only be calculated once. Since $\boldsymbol{R}_j^{-T}\boldsymbol{u}_j^* = (\boldsymbol{I} + \eta_j\tilde{\boldsymbol{\Lambda}})^{-1/2}\boldsymbol{\Gamma}'\boldsymbol{u}_j^*$, the calculation in step (2) can be calculated by $\boldsymbol{v}_j^{(2)} = \boldsymbol{\Gamma}'\boldsymbol{u}_j^*$ and $\boldsymbol{v}_j^{(1)} = (\boldsymbol{I} + \eta_j\tilde{\boldsymbol{\Lambda}})^{-1/2}\boldsymbol{v}_j^{(2)}$. The former calculation takes $O(n^2)$ operations and the latter only $O(n)$ operations. Similar calculations hold for Step (3), so one cycle to update $\boldsymbol{Z}$ takes $O(n^2p)$ operations. In either case, there is a dramatic computational savings over the naive computation of $\hat{\boldsymbol{z}} = (\boldsymbol{I}_{np} + \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1 \otimes \boldsymbol{Q})^{-1}\boldsymbol{y}$, which requires $O(n^3p^3)$ operations.*

4.2 Bayesian Computation of Variance Parameters

The conditional posteriors of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ can be computed from those of $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}$ based on the following proposition. Throughout the discussion, we use the Bayesian convention "$[\cdot \mid \cdot]$" to denote a conditional density.

Proposition 1 *Consider decomposition (14), with a lower triangular matrix $\boldsymbol{\Psi}$. The priors for $\boldsymbol{\Psi}$ and $\boldsymbol{\Xi}$ are given by (41) and (43). Let ψ_{ij} $(1 \le j \le i \le p)$ be the elements of $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi}_{-ij}$ be the elements of $\boldsymbol{\Psi}$ excluding ψ_{ij}. Let a_{ij} be the (i,j)th element of the $p^2 \times p^2$ positive definite matrix $\boldsymbol{A} = ((\boldsymbol{Y}-\boldsymbol{Z})'(\boldsymbol{Y}-\boldsymbol{Z}))\otimes\boldsymbol{I}_p + (\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})\otimes\boldsymbol{\Xi}$, and let $\boldsymbol{\psi} = vec(\boldsymbol{\Psi})$. Then*

(a) $(\boldsymbol{\Xi} \mid \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) \propto$ *Wishart*$_p(n+p-1, [\boldsymbol{\Psi}(\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})\boldsymbol{\Psi}' + \boldsymbol{\Phi}]^{-1})$, *and* $(\boldsymbol{\Phi} \mid \boldsymbol{\Xi}) \propto$ *Wishart*$_p(2(p+1), (\boldsymbol{\Xi} + b\boldsymbol{I})^{-1})$.

(b) $[\boldsymbol{\psi} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Xi}] \propto \prod_{i=1}^{p} \psi_{ii}^{2n-2-i} \exp\{-\frac{1}{2}\boldsymbol{\psi}'\boldsymbol{A}\boldsymbol{\psi}\}$.

(c) For $j < i$, $(\psi_{ij} \mid \boldsymbol{Y}, \boldsymbol{Z}, \Xi, \boldsymbol{\Psi}_{-ij}) \propto N(-\frac{q_{ij}}{2r_{ij}}, r_{ij}^{-1})$, where

$$r_{ij} = g_{i+(j-1)p,i+(j-1)p} \text{ and } q_{ij} = \sum_{i' \geq j', (i',j') \neq (i,j)} \psi_{i'j'} g_{i+(j-1)p,i'+(j'-1)p}.$$

(d) $[\psi_{ii} \mid \boldsymbol{Y}, \boldsymbol{Z}, \Xi, \boldsymbol{\Psi}_{-ii}] \propto \psi_{ii}^{2n-2-i} \exp\{-\frac{1}{2}[r_{ii}\psi_{ii}^2 + q_{ii}\psi_{ii}]\}$, where

$$r_{ii} = g_{i+(i-1)p,i+(i-1)p} \text{ and } q_{ii} = \sum_{i' \geq j', i' \neq i} \psi_{i'j'} w_{i+(i-1)p,i'+(j'-1)p}.$$

Proof. To prove Part (a), note that

$$[\boldsymbol{\Xi} \mid \boldsymbol{Z}, \boldsymbol{\Psi}, \boldsymbol{\Phi}] \propto |\boldsymbol{\Xi}|^{\frac{n-2}{2}} \mathrm{etr}\Big\{-\frac{1}{2}[\boldsymbol{\Psi}(\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})\boldsymbol{\Psi}' + \boldsymbol{\Phi}]\boldsymbol{\Xi}\Big\}$$

and that $[\boldsymbol{\Phi} \mid \boldsymbol{\Xi}]$ is given by (50). To prove Part (b), note

$$\begin{aligned} [\boldsymbol{\Psi} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Xi}] &\propto \prod_{i=1}^{p} \psi_{ii}^{2n-2+m_0-(p+1)-i} \mathrm{etr}\Big\{-\frac{1}{2}[((\boldsymbol{Y}-\boldsymbol{Z})'(\boldsymbol{Y}-\boldsymbol{Z}))\boldsymbol{\Psi}'\boldsymbol{\Psi} + (\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})\boldsymbol{\Psi}'\boldsymbol{\Xi}\boldsymbol{\Psi}]\Big\} \\ &= \prod_{i=1}^{p} \psi_{ii}^{2n-2-i} \mathrm{etr}\Big\{-\frac{1}{2}\boldsymbol{\psi}'\boldsymbol{W}\boldsymbol{\psi}\Big\}. \end{aligned}$$

Denote the kth element of the vector $\boldsymbol{\psi}$ by $\check{\psi}_k$, so $\check{\psi}_{i+(j-1)p}$ corresponds to ψ_{ij} in the matrix $\boldsymbol{\Psi}$. Because $\psi_{ij} = 0$ if $j > i$, we can express $\boldsymbol{\psi}'\boldsymbol{W}\boldsymbol{\psi}$ as

$$\sum_{k_1=1}^{p^2} \sum_{k_2=1}^{p^2} \check{\psi}_{k_1}\check{\psi}_{k_2} w_{k_1k_2} = \sum_{i \geq j} \sum_{i' \geq j'} \psi_{ij}\psi_{i'j'} w_{i+(j-1)p,i'+(j'-1)p}.$$

Straightforward algebra yields $(\psi_{ij} | \boldsymbol{\Psi}_{-ij}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Xi}) \propto \exp\{-\frac{1}{2}[r_{ij}\psi_{ij}^2 + q_{ij}\psi_{ij}]\}$, which proves Part (c). Part (d) follows similarly. □

Note that r_{ij} is positive because it is the $i + (j-1)p$th diagonal element of the positive definite $\boldsymbol{A}$. The conditional posterior of diagonal element ψ_{ii} of $\boldsymbol{\Psi}$ is nonstandard but is log-concave. The log-concavity of the conditional posterior of ψ_{ii} permits efficient simulation using the adaptive rejection sampling algorithm of Gilks & Wild (1992). In the numerical examples and empirical applications below, we will draw the conditional posterior of ψ_{ii} via the adoptive rejective method.

Remark 2 *The off-diagonal elements ψ_{ij} in (c) above can be sampled as a block since the full conditional is multivariate normal.*

4.3 Estimating intrinsic principal curves

Some care is needed in calculating the estimated intrinsic principal curves, namely the columns of $\widehat{\boldsymbol{V}}$. One can save the sampled $\boldsymbol{V}$ at each MCMC cycle and report the average, but this procedure is not advisable because the calculation of the columns $\boldsymbol{O}$ in (16) is not unique, rendering MCMC averages meaningless. For the same reason, the ordinary MCMC estimate of $\boldsymbol{\Delta}$ is not appropriate. Instead, we first compute the MCMC estimates $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\Xi}}$ and then compute $\widehat{\boldsymbol{\Delta}}$ using (16) and (17). The estimated intrinsic principal curves can be estimated as the columns of $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Delta}}'$.

A complication in interpreting the columns of $\widehat{\boldsymbol{V}}$ is the presence of irrelevant linear trends. Using (26), one can see that the columns of $\widehat{\boldsymbol{V}}$ contain the least squares terms generated by $(\boldsymbol{\Delta} \otimes \boldsymbol{P}_0)\boldsymbol{y}$. These terms are essentially arbitrary and distract from the interpretation of the intrinsic principal curves as defined in Section 2.5. From (26), one can show that $\widehat{\boldsymbol{Z}} = \boldsymbol{P}_0\boldsymbol{Y} + \widehat{\boldsymbol{V}}_*\widehat{\boldsymbol{\Delta}}^{-T}$, where $\widehat{\boldsymbol{V}}_* = (\boldsymbol{I}_n - \boldsymbol{P}_0)\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Delta}}'$. The columns of $\widehat{\boldsymbol{V}}_*$ are now free of distracting linear trends and reflect the level of smoothing controlled by $(\widehat{\eta}_1, \ldots, \widehat{\eta}_p)$. Let $\widehat{\boldsymbol{Z}}_* = \widehat{\boldsymbol{V}}_*\widehat{\boldsymbol{\Delta}}^{-T}$. Then $\widehat{\boldsymbol{Z}}_* = (\boldsymbol{I}_n - \boldsymbol{P}_0)\widehat{\boldsymbol{Z}}$, so the columns of $\widehat{\boldsymbol{Z}}_*$ are precisely the shrinkage part of the multivariate smoothing spline fit. Intrinsic principle curves describe the relationships among the columns of $\widehat{\boldsymbol{Z}}_*$.

In analogy to ordinary principal components, the columns of $\widehat{\boldsymbol{\Delta}}^{-T}$ can be interpreted as "factor loadings" for the intrinsic principal curves. These "factor loadings" may also be difficult to interpret because the columns of $\widehat{\boldsymbol{V}}_*$ are not orthogonal or normalized. However, they can be used for dimension reduction, and analysis of reduced dimension fits can lead to insight into the original data. Denote the factor loadings by $\boldsymbol{L} = \widehat{\boldsymbol{\Delta}}^{-T} = [\ell_{ij}]_{p\times p}$, so $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{V}}\boldsymbol{L}$. As before, denote the fitted curves and intrinsic principal curves respectively by $\widehat{\boldsymbol{Z}} = [\widehat{\boldsymbol{z}}_1^*, \ldots, \widehat{\boldsymbol{z}}_p^*]_{n\times p}$ and $\widehat{\boldsymbol{V}} = [\widehat{\boldsymbol{v}}_1^*, \ldots, \widehat{\boldsymbol{v}}_p^*]_{n\times p}$. With this notation, the jth fitted curve can be expressed in terms of intrinsic principal curves as

$$\widehat{\boldsymbol{z}}_j^* = \sum_{i=1}^{p} \ell_{ij} \widehat{\boldsymbol{v}}_i^*. \tag{47}$$

One can reduce the dimension of this representation by using only the first m terms above, which we can denote by $\widehat{\boldsymbol{z}}_j^{(m)} = \sum_{i=1}^{m} \ell_{ij} \widehat{\boldsymbol{v}}_i^*$. If we take out the least squares trend component by projection, the goodness of fit with this reduced dimension estimate is measured by a pseudo coefficient of determination,

$$R_{jm}^2 = 100 \times \left(1 - \frac{\|\widehat{\boldsymbol{z}}_j^* - \widehat{\boldsymbol{z}}_j^{(m)}\|^2}{\|(\boldsymbol{I}_n - \boldsymbol{P}_0)\widehat{\boldsymbol{z}}_j^*\|^2}\right), \tag{48}$$

the percent of variation of the jth curve explained by the first m intrinsic principal curves.

5 Simulation Study

We generated data sets from the two-equation model for $t = 1, \ldots, n$,

$$y_{1t} = g_1(t) + \epsilon_{1t}, \quad y_{2t} = g_2(t) + \epsilon_{2t},$$

where the ϵ_{jt} are independent $N(0, \sigma_{0j}^2)$, $j = 1, 2$, with $\mathrm{Corr}(\epsilon_{1t}, \epsilon_{2t}) = \rho$, $t = 1, \ldots, n$. For each example, we generated $N = 200$ samples of data, each with sample size $n = 100$. We compared the estimation errors of the functions for the same generated data $\boldsymbol{Y}$ using the multivariate smoothing spline (1) with $p = 2$ and $k = 2$ or separate univariate smoothing splines, treating the smoothing parameters as unknown.

For multivariate splines, we used the right Haar prior (41) for $\boldsymbol{\Psi}$ and prior (43) for $\boldsymbol{\Xi}$, and for univariate splines, we used the priors $\pi(\sigma_0^2) \propto 1/\sigma_0^2$ and (42). We used $b = 2000$ for the univariate splines (which corresponds to a median edf of 6.3) and 8000 for the multivariate splines (so that the smallest eigenvalue of the $\boldsymbol{\Xi}$ corresponds to a median edf of 6.2). For each data set, we ran 20,000 MCMC cycles after 1,000 burn-in cycles. We chose initial values $\boldsymbol{\Sigma}_0 = 0.1\boldsymbol{I}$ and $\boldsymbol{\Xi} = \boldsymbol{I}$.

We present three cases. For each case, we used $\sigma_{01} = \sigma_{02} = 0.1$ and generated data with three different correlations, $\rho = -0.8, 0, 0.8$. To define the cases, let $f_1(t) = \sin(4t\pi/n)$, $f_2(t) = \sin(4t\pi/n + \pi/2)$, $f_3(t) = \sin(t\pi/n)$, and $f_4(t) = \sin(2t\pi/n)$. The two data-generating curves for each are

Case 1: $g_1(t) = f_1(t)$, $g_2(t) = f_1(t)$;

Case 2: $g_1(t) = f_1(t)$, $g_2(t) = f_2(t)$;

Case 3: $g_1(t) = (f_1(t) + f_3(t))/2$, $g_2(t) = (f_1(t) + f_4(t))/2$.

In Case 1, the data-generating curves are the same. The singular data-generating $\boldsymbol{\Sigma}_1$ violates our model assumption but serves as a good test of our algorithm when the posterior of $\boldsymbol{\Sigma}_1$ is near singular. In Case 2, the two curves are orthogonal. In Case 3, the two curves share a common complex component $f_1(t)$. The second derivatives of the two curves are perfectly correlated in Case 1, perfectly uncorrelated in Case 2, and positively correlated in Case 3.

Table 1: Average mean square error results from simulations for Cases 1-3

	function	univariate	multivariate		
ρ			-0.8	0.0	0.8
Case 1	g_1	1.042	0.507	0.748	0.998
		(0.029)	(0.021)	(0.021)	(0.028)
Case 2	g_1	1.059	1.027	1.062	1.013
		(0.032)	(0.031)	(0.032)	(0.032)
	g_2	1.001	1.000	1.003	0.993
		(0.030)	(0.031)	(0.031)	(0.029)
Case 3	g_1	1.060	0.786	0.991	1.052
		(0.031)	(0.027)	(0.030)	(0.031)
	g_2	0.902	0.690	0.746	0.798
		(0.030)	(0.028)	(0.030)	(0.029)

Table 1 summarizes the average mean square estimation error (AMSE) over the 200 samples (times 100). The standard errors of the 200 MSEs for each simulation are included in parentheses. From the table, we conclude that multivariate splines produce estimates with better accuracy than univariate splines in Cases 1 and 3 and about the same accuracy in Case 2. Within each case, the difference in univariate splines stems from sampling error only. The average mean square error of the multivariate estimates clearly depends on the error correlation, ρ.

For all correlations in Case 1, we found that the estimated $\boldsymbol{\Delta}$ was quite similar, and the transformed data vectors $[\boldsymbol{u}_1^*, \boldsymbol{u}_2^*] = [\boldsymbol{y}_1^*, \boldsymbol{y}_2^*]\widehat{\boldsymbol{\Delta}'}$ approximately satisfied $\boldsymbol{u}_1^* \propto \boldsymbol{y}_1^* + \boldsymbol{y}_2^*$ and $\boldsymbol{u}_2^* \propto \boldsymbol{y}_1^* - \boldsymbol{y}_2^*$. Thus $\boldsymbol{u}_1^*$ contains the common signal f_1, and $\boldsymbol{u}_2^*$ is essentially all noise. The data from one simulation (sample number 1 out of 200) with $\rho = -0.8$ is displayed in Figure 1. The top row shows the true curves and data for data sets 1 and 2. The bottom row shows the detrended and transformed data $(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}_1^*$ and $(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}_2^*$ along with the detrended smoothed data, $(\boldsymbol{I} - \boldsymbol{P}_0)\hat{\boldsymbol{v}}_1^*$ and $(\boldsymbol{I} - \boldsymbol{P}_0)\hat{\boldsymbol{v}}_2^*$. The cross sample averages of the estimated smoothing parameters were $\hat{\eta}_1 = 12.1$ and $\hat{\eta}_2 = 3.93 \times 10^5$, showing again how the first component contains the signal and the second component is mostly noise. In contrast, the cross-sample averages of the univariate smoothing parameters were $(86.73, 88.11)$, nearly equal as expected. For $\rho < 0$, the variance of the components of $\boldsymbol{u}_1^*$ is small, and smoothing $\boldsymbol{u}_1^*$ is very efficient as seen in the figure. Thus this case shows the most reduction in AMSE. When $\rho = 0.8$, the high correlation between data sets suggests that there is little additional information to be gained by multivariate smoothing, a fact confirmed in the simulation. The intermediate case $\rho = 0.0$ also shows intermediate gain in AMSE.

In Case 2, the data generating curves are orthogonal. Unsurprisingly, there is no apparent gain in estimating the curves using the multivariate smoothing spline. On the other hand, there is no apparent loss in using multivariate smoothing despite the fact that more parameters must be estimated.

Case 3, with positively correlated signals, is an intermediate case between the

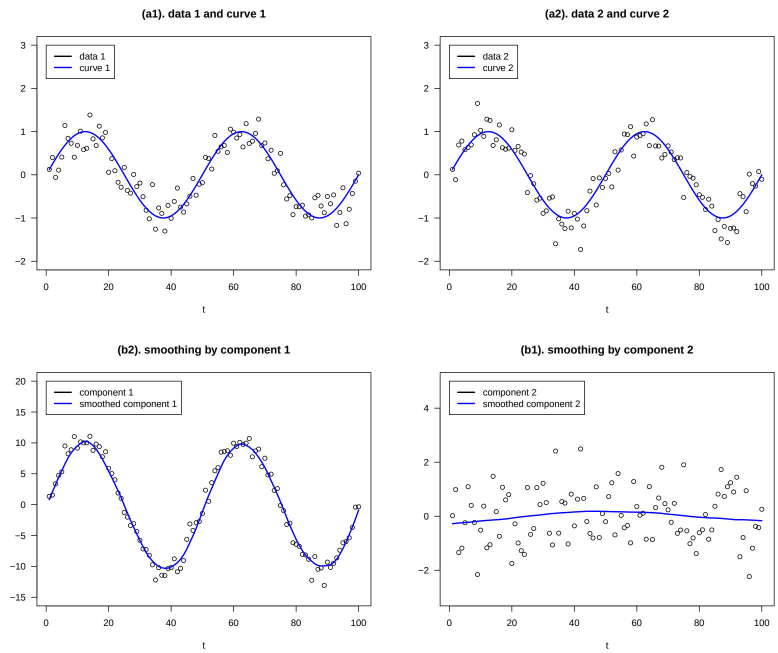

Figure 1: Data generating curves and decomposed components in Case 1. $\rho = -0.8$. Top row: one data sample $\boldsymbol{y}_j^*$ with data-generating curves $\boldsymbol{g}_j^*$, $j = 1, 2$. Bottom row: the detrended intrinsic principal curves $(\boldsymbol{I} - \boldsymbol{P}_0)\hat{\boldsymbol{v}}_j^*$ (solid line) and the corresponding detrended pseudo data $(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}_j^*$ (circles), $j = 1, 2$.

perfectly correlated signals of Case 1 and the uncorrelated signals of Case 2. The simulation showed substantially smaller AMSE using multivariate smoothing for all cases except estimating g_1 with $\rho = 0.8$. In all cases, the effect of the data transformation $\boldsymbol{\Delta}'$ was approximately $\boldsymbol{u}_1^* \propto \boldsymbol{y}_1^* + \boldsymbol{y}_2^*$ and $\boldsymbol{u}_2^* \propto \boldsymbol{y}_1^* - 2\boldsymbol{y}_2^*$. When $\rho = -0.8$, the cross-sample averages of the smoothing parameters were $\hat{\eta}_1 = 19.3$ and $\hat{\eta}_2 = 2.11 \times 10^4$.

Space considerations preclude reporting further results, but more simulations not reported here of bivariate and some high dimensional cases showed that the gain from multivariate smoothing is quite robust. In cases where gain is possible, the multivariate approach is better. In other cases with reasonable sample sizes, there appears to be little if any loss in multivariate smoothing compared to univariate smoothing.

6 Application: Estimate the Trends in Economic Policy Uncertainty

Bloom (2009) showed that shocks in macroeconomic uncertainty produces fluctuations in aggregate output, employment and productivity. Economic policy is an important source of macroeconomic uncertainty. Understanding the nature of policy-induced uncertainty is useful for better policy making.

In a recent study, Baker et al. (2013) quantified the U.S. economic policy uncertainty (EPU) by aggregating three measures, the frequency of references to economic uncertainty and policy in ten leading newspapers, the number of federal tax code provisions set to expire in future years, and the extent of disagreement among economic forecasters over government purchases and CPI. For European countries, they constructed the EPU without using the tax code expiration provisions and only used the newspaper keywords counts and disagreements among economic forecasters.

The EPU data (obtained at www.policyuncertainty.com) are quite noisy, even plotted on a log scale. One possible source of the noise is the randomness in the construction of the indexes. For example, keywords in newspaper coverage may be influenced by the opinions of reporters and editors, and forecast disagreement may be due to subjective errors of some forecasters. Smoothing the EPU is useful for a number of reasons. First, the trend in EPU is likely a better measure of economic policy uncertainty. Second, economic decisions of long-term consequences (such as decisions on investment) depend on the trend in EPU. Third, the trend in EPU is more useful than the noisy data for evaluating the performance of policy makers. Fourth, the trends in EPU may depict a clearer picture of the shifts in EPU over time and across countries.

Availability of monthly EPU data dictates our focus on seven countries: US, China, Canada, France, Germany, Italy, and UK. The sample is from Jan 1997 to Mar 2013. We set the prior parameter b to 750,000 for the multivariate spline, which makes the posterior mean of the edf with $p = 7$ and $n = 195$ close to 7 (using the smallest

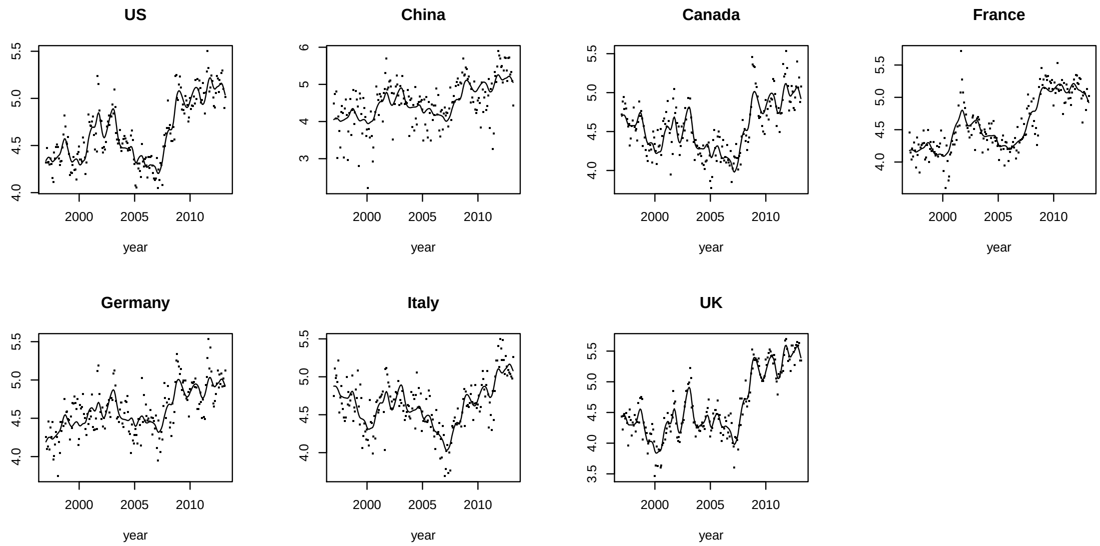

Figure 2: Plots of policy uncertainty data by country. Each panel displays the log EPU data by year and the fitted trend.

eigenvalue of Ξ). We used 900,000 MCMC cycles following 100,000 burn in cycles, saving every 100th cycle. The MCMC estimates of the smoothing parameters $(\widehat{\eta}_1, \ldots, \widehat{\eta}_7)$ were approximately $(9.92, 822, 1.57 \times 10^4, 9.64 \times 10^4, 6.27 \times 10^5, 3.17 \times 10^6, 1.20 \times 10^8)$. The raw data (the logarithm of the EPU indexes) along with the multivariate smoothing spline estimates are shown in Figure 2. The detrended estimated intrinsic principal curves are shown in Figure 3 along with the corresponding detrended pseudo-data. Table 2 shows that the correlation in EPU trend is much stronger than the correlation in the errors.

While there is one clear dominant component, this analysis suggests more complicated relationships among the data. The percent of unexplained variation using m intrinsic principal curves (48) in Table 3 shows that the dominant component is essentially the UK series. This component is also strongly associated with Canada. Component 2 accounts for most of the remaining variability for the US and France. The third component is mainly associated with China and Italy, and the fourth component is associated with Germany.

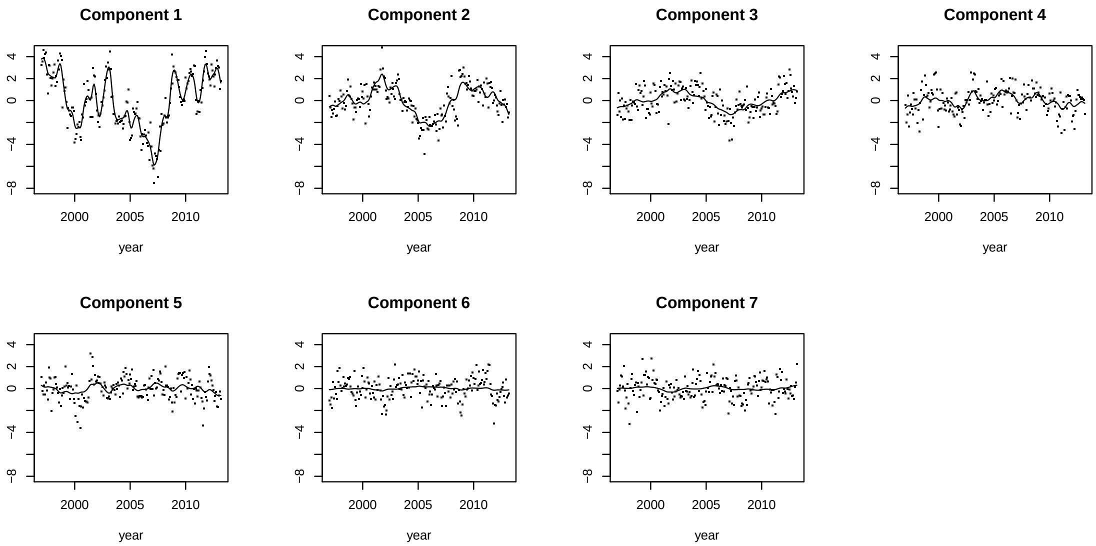

Figure 3: Plots of intrinsic principal curves (columns of $(\boldsymbol{I}_n - \boldsymbol{P}_0)\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\Delta}'}$) for the policy uncertainty data. Each panel displays an estimated detrended intrinsic principal curve $(\boldsymbol{I} - \boldsymbol{P}_0)\hat{\boldsymbol{v}}_j^*$ (solid line,) together with the pseudo data $(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}_j^*$ (dots).

Table 2: Estimated correlation matrices for ($\widehat{\boldsymbol{\Sigma}}_0$ and $\boldsymbol{\Sigma}_1$).

	US	China	Canada	France	Germany	Italy	UK
US	1.000	0.193	0.423	0.178	0.370	0.125	0.210
China	0.966	1.000	0.164	0.210	0.109	0.170	0.167
Canada	0.968	0.911	1.000	0.112	0.327	0.111	0.185
France	0.847	0.883	0.785	1.000	0.222	0.160	0.344
Germany	0.982	0.944	0.977	0.825	1.000	0.133	0.211
Italy	0.946	0.948	0.917	0.892	0.941	1.000	0.147
UK	0.985	0.958	0.972	0.877	0.986	0.949	1.000

The upper-triangular part reports the correlations of $\boldsymbol{\Sigma}_0$ and the lower-triangular part those of $\boldsymbol{\Sigma}_1$.

Table 3: Percent of variation unexplained by the first m intrinsic principal curves for the Policy Uncertainty data.

	Number of components					
	1	2	3	4	5	6
US	42.073	4.360	0.569	0.536	0.044	0.046
China	40.533	21.047	6.210	8.336	5.331	2.820
Canada	7.295	1.720	2.054	1.834	0.184	0.001
France	65.081	4.061	3.630	2.174	0.042	0.033
Germany	80.437	19.386	15.837	3.883	1.716	1.780
Italy	10.904	14.438	1.599	0.109	0.124	0.121
UK	1.438	1.314	0.724	0.206	0.007	0.001

7 Concluding Remarks

The multivariate spline is applicable to smoothing spatial or time series data that contain potentially correlated errors and co-moving curves. In this paper, we lay out an algorithm for joint estimation of the curves and smoothing parameter matrices in a Bayesian setting, where the error covariance matrix has a noninformative prior and the smoothing parameter matrix has a proper prior. Our experience shows that the algorithm is quite efficient and applicable to a wide variety of problems. Consider the problem of measuring business cycles. The commonly used detrending methods are univariate. A stochastic growth model commonly used for business cycle analysis imposes restrictions on the short-run component variance $\mathbf{\Sigma}_0$ and the long run component variance $\mathbf{\Sigma}_1$. Univariate detrending is equivalent to imposing diagonal restrictions to these variances, which violates an essential assumption of all schools of theories, that the detrended series are correlated. The empirical results of univariate detrending are likely biased measurement of business cycles and misleading tests of economic theories. The method on multivariate splines may be employed for better estimates of time series trends, as in the empirical application in this study.

References

Arnold, B. C. (1983), *Pareto Distributions*, International Co-operative Publishing House. Statistical Ecology Series.

Baker, S. R., Bloom, N. & Davis, S. J. (2013), Measuring economic policy uncertainty. Working paper.

Berger, J. O. & Sun, D. (2008), 'Objective priors for the bivariate normal model', *The Annals of Statistics* **36**(2), 963–982.

Bloom, N. (2009), 'The impact of uncertainty shocks', *Econometrica* **77**, 623–685.

Cheng, C.-I. & Speckman, P. (2013), 'Bayesian smoothing spline analysis of variance', *Computational Statist. & Data Anal.* **56**(12), 3945–3958.

Cressie, N. & Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Wiley.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker Inc.

Fessler, J. A. (1991), 'Nonparametric fixed-interval smoothing with vector splines', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **39**, 852–859.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**, 398–409.

Gilks, W. R. & Wild, P. (1992), 'Adaptive rejection sampling for Gibbs sampling', *Applied Statistics* **41**, 337–348.

Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, Chapman & Hall Ltd.

Gupta, A. K. & Nagar, D. K. (2000), *Matrix Variate Distributions*, Chapman & Hall Ltd.

Hastie, T. & Stuetzle, W. (1989), 'Principal curves', *Journal of the American Statistical Association* **84**, 502–516.

Hastie, T. & Tibshirani, R. (1999), *Generalized Additive Models*, Chapman & Hall Ltd.

Hill, B. M. (1965), 'Inference about variance components in the one-way model', *Journal of the American Statistical Association* **60**, 806–825.

Hobert, J. & Casella, G. (1996), 'The effect of improper priors on Gibbs sampling in hierarchicallinear mixed models', *Journal of the American Statistical Association* **91**, 1461–1473.

Kimeldorf, G. S. & Wahba, G. (1970), 'A correspondence between Bayesian estimation on stochastic processes and smoothing by splines', *The Annals of Mathematical Statistics* **41**, 495–502.

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**(481), 410–423.

Ramsay, J. O. & Silverman, B. W. (1997), *Functional Data Analysis*, Springer-Verlag Inc.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC.

Schoenberg, I. J. (1964), 'Spline functions and the problem of graduation', *Proc. Nat. Acad. Sci. USA* **52**, 947–950.

Shiller, R. J. (1984), 'Smoothness priors and nonlinear regression', *Journal of the American Statistical Association* **79**, 609–615.

Speckman, P. L. & Sun, D. (2003), 'Fully Bayesian spline smoothing and intrinsic autoregressive priors', *Biometrika* **90**(2), 289–302.

Sun, D., Ni, S. & Speckman, P. (2014), Bayesian analysis of multivariate smoothing splines ii, Ms., University of Missouri, Departments of Statistics and Economics.

Sun, D. & Speckman, P. L. (2008), 'Bayesian hierarchical linear mixed models for additive smoothing splines', *Annals of the Institute of Statistical Mathematics* **60**(3), 499–517.

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

Wahba, G. (1985), 'A comparison of Gcv and Gml for choosing the smoothing parameter in the generalized spline smoothing problem', *The Annals of Statistics* **13**, 1378–1402.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

Wang, Y., Guo, W. & Brown, M. B. (2000), 'Spline smoothing for bivariate data with applications to association between hormones', *Statistica Sinica* **10**(2), 377–397.

Wecker, W. E. & Ansley, C. F. (1983), 'The signal extraction approach to nonlinear regression and spline smoothing', *Journal of the American Statistical Association* **78**, 81–89.

White, G. A. (2006), Bayesian semiparametric spatial and joint spatial temporal smoothing, Ph.D. dissertation, University of Missouri Columbia, Department of Statistics.

Yee, T. W. & Wild, C. J. (1996), 'Vector generalized additive models', *Journal of the Royal Statistical Society, Series B: Methodological* **58**, 481–493.

Appendix

Proof of Lemma 1:

The space of natural splines $\mathcal{NS}^{2k}(\boldsymbol{t})$ has the following properties (e.g., Eubank, 1999, Chapter 5.8):

1. To each set of reals $(c_1, \ldots, c_n) \in \mathbb{R}^n$, there exists a unique natural spline $q \in \mathcal{NS}^{2k}(\boldsymbol{t})$ such that $q(t_i) = c_i$, $i = 1, \ldots, n$.

2. If $f \in \mathcal{W}^{2,k}$ and $f(t_i) = 0$, $i = 1 \ldots, n$, then

$$\int_0^1 q^{(k)}(s) f^{(k)}(s)\, ds = 0 \text{ for all } q \in \mathcal{NS}^{2k}(\boldsymbol{t}).$$

Now let $\boldsymbol{g}$ in (2) be an arbitrary member of $\mathcal{W}_p^{2,k}$, and let $q_j \in \mathcal{NS}^{2k}(\boldsymbol{t})$ be the unique interpolating natural spline satisfying $g_j(t_i) = q_j(t_i)$, $i = 1, \ldots, q$. With $f_j = g_j - q_j$, property 2 and the fact that $\boldsymbol{g} = \boldsymbol{q} + \boldsymbol{f}$ imply

$$\int_0^1 q_j^{(k)}(s)(g_\ell^{(k)}(s) - q_\ell^{(k)}(s))\, ds = 0, \quad 1 \le j, \ell \le p.$$

Thus,

$$\begin{aligned}\int_0^1 \boldsymbol{g}^{(k)}(s)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{g}^{(k)}(s))'\, ds &= \int_0^1 \boldsymbol{q}^{(k)}(s)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{q}^{(k)}(s))'\, ds + \int_0^1 \boldsymbol{f}^{(k)}(s)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{f}^{(k)}(s))'\, ds \\ &\ge \int_0^1 \boldsymbol{q}^{(k)}(s)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{q}^{(k)}(s))'\, ds\end{aligned}$$

since $\boldsymbol{\Sigma}_1$ is positive definite. This shows that the minimizer of (2) lies in $\mathcal{NS}_p^{2k}(\boldsymbol{t})$.

Proof of Lemma 3:

Write $c_p = 2^{p(p+1)/2}\Gamma_p((p+1)/2)$. Then

$$[\boldsymbol{\Xi} \mid \boldsymbol{\Phi}] = \frac{1}{c_p}|\boldsymbol{\Phi}|^{\frac{p+1}{2}}\mathrm{etr}(-\frac{1}{2}\boldsymbol{\Phi}\boldsymbol{\Xi}) \text{ and } [\boldsymbol{\Phi}] = \frac{b^{\frac{p(p+1)}{2}}}{c_p}\mathrm{etr}(-\frac{b}{2}\boldsymbol{\Phi}).$$

Therefore the joint density of $(\boldsymbol{\Xi}, \boldsymbol{\Phi})$ is

$$[\boldsymbol{\Xi}, \boldsymbol{\Phi}] = \frac{b^{\frac{p(p+1)}{2}}}{c_p^2}|\boldsymbol{\Phi}|^{\frac{p+1}{2}}\mathrm{etr}\{-\frac{1}{2}\boldsymbol{\Phi}(\boldsymbol{\Xi} + b\boldsymbol{I}_p)\}. \tag{49}$$

The conditional distribution of $\boldsymbol{\Phi}$ given $\boldsymbol{\Xi}$ is

$$[\boldsymbol{\Phi} \mid \boldsymbol{\Xi}] = \frac{|\boldsymbol{\Xi} + b\boldsymbol{I}_p|^{p+1}}{2^{p(p+1)}\Gamma_p(p+1)}|\boldsymbol{\Phi}|^{\frac{p+1}{2}}\mathrm{etr}\Big\{-\frac{1}{2}\boldsymbol{\Phi}(\boldsymbol{\Xi} + b\boldsymbol{I}_p)\Big\}, \tag{50}$$

which is the pdf of $\mathrm{Wishart}_p(2(p+1), (\boldsymbol{\Xi} + b\boldsymbol{I}_p)^{-1})$. Integrating out $\boldsymbol{\Phi}$ in (49), we have (43).