Bayesian Analysis (2010) **5**, Number 2, pp. 319–344

Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection

Surya T. Tokdar[*], Yu M. Zhu[†] and Jayanta K. Ghosh[‡]

Abstract. We develop a novel Bayesian density regression model based on logistic Gaussian processes and subspace projection. Logistic Gaussian processes provide an attractive alternative to the popular stick-breaking processes for modeling a family of conditional densities that vary smoothly in the conditioning variable. Subspace projection offers dimension reduction of predictors through multiple linear combinations, offering an alternative to the zeroing out theme of variable selection. We illustrate that logistic Gaussian processes and subspace projection combine well to produce a computationally tractable and theoretically sound density regression procedure that offers good out of sample prediction, accurate estimation of subspace projection and satisfactory estimation of subspace dimensionality. We also demonstrate that subspace projection may lead to better prediction than variable selection when predictors are well chosen and possibly dependent on each other, each having a moderate influence on the response.

Keywords: Bayesian Inference, Semiparametric Model, Posterior Consistency, Gaussian Process, Markov Chain Monte Carlo, Dimension Reduction

1 Introduction

Density regression offers flexible modeling of a response variable y given a collection of covariates $\mathbf{x} = (x_1, \cdots, x_p)$. Density regression views the entire conditional density $p(y \mid \mathbf{x})$ as a function valued parameter and allows its center, spread, skewness, modality and other such features to vary with $\mathbf{x}$. Such model flexibility necessitates entertaining a huge parameter space. It is thus important to calibrate the parameter space with a penalty or a prior distribution to facilitate meaningful inference from finite data. The latter approach, leading to Bayesian density regression, is the focus of this paper.

Despite its recent introduction to the literature, Bayesian density regression has seen a rapid development in the last few years. Existing methods are almost exclusively based on a stick-breaking representation of the form $p(y \mid \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) g(y \mid \theta_h(\mathbf{x}))$, where $\{g(y \mid \theta) : \theta \in \Theta\}$ is a chosen family of parametric densities, $\pi_h(\mathbf{x})$'s are weights that add up to one at every $\mathbf{x}$ and θ_h's map $\mathbf{x}$ to Θ. Constructing prior distributions on $\{\pi_h(\mathbf{x}), \theta_h(\mathbf{x})\}_{h=1}^{\infty}$ that guarantee large supports for $p(y \mid \mathbf{x})$ and retain computational tractability is the subject of much ongoing research, see for example Griffin and Steel (2006), Dunson, Pillai, and Park (2007), Dunson and Park (2008) and Chung and Dunson (2009). In addition, increasing attention is being devoted to incorporate dimension

[*] Department of Statistical Science, Duke University, Durham, NC, mailto:tokdar@stat.duke.edu

[†] Department of Statistics, Purdue University, West Lafayette, IN, mailto:yuzhu@stat.purdue.edu

[‡] Department of Statistics, Purdue University, West Lafayette, IN, and Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, Kolkata, India, mailto:ghosh@stat.purdue.edu

DOI:10.1214/10-BA605

reduction within these models. Most commonly, this is done through variable selection – where given an inclusion parameter $\boldsymbol{\gamma} \in \{0,1\}^p$, the model on $p(y \mid \mathbf{x})$ involves only those coordinates x_j for which γ_j equals 1 (Chung and Dunson 2009). A Bayesian inference on $p(y \mid \mathbf{x})$ is then carried out with a suitable prior distribution on $\boldsymbol{\gamma}$.

In this paper we introduce and develop an entirely different framework for modeling a family of conditional densities and incorporating dimension reduction. To model conditional densities, we extend the logistic Gaussian process originally studied by Lenk (1988, 1991, 2003) and Tokdar (2007) for modeling a single density. We work with a Gaussian process on the product space of $(\mathbf{x}, y)$ and apply a logistic transformation to each $\mathbf{x}$-slice of the process to produce a density function in y. The smoothness of the original process and that of the logistic transform ensure that the resulting family of densities vary smoothly across $\mathbf{x}$ and this variation is entirely nonparametric. The extent of this smoothness, or in other words, the dependence between $p(\cdot \mid \mathbf{x})$ across $\mathbf{x}$, is easily encoded in the covariance function of the Gaussian process.

Our extension of the logistic Gaussian process (LGP) is reminiscent of MacEachern's (see MacEachern 1999, 2000) extension of the Dirchlet process to the dependent Dirichlet process. However, the LGP extension appears much simpler both conceptually and technically – owing to the smoothness properties of Gaussian processes defined over the product space of $(\mathbf{x}, y)$. As would be demonstrated later, the LGP process has a large support – it accommodates all conditional densities $p(y \mid \mathbf{x})$ with some tail conditions for which the map $(\mathbf{x}, y) \mapsto \log p(y \mid x)$ is continuous (see also Tokdar and Ghosh 2007, for a simpler treatment of the single density case). Apart from leading to posterior consistency properties, this easily identified support of the LGP offers practitioners a transparent characterization of the induced model. A finer calibration of this support and derivation of posterior convergence rates seem quite plausible in view of recent theoretical studies on Gaussian processes (van der Vaart and van Zanten 2008), although this has not been pursued in this paper.

A second novelty of our approach is the use of subspace projection to reduce the dimensionality of $\mathbf{x}$. We assume that

$$p(y \mid \mathbf{x}) = p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x}) \tag{1}$$

for some lower dimensional linear subspace $\mathcal{S} \subset \mathbb{R}^p$ with associated projection operator $\mathbf{P}_{\mathcal{S}}$. Both $\mathcal{S}$ and the function encoding $p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$ are seen as model parameters. Subspace projection is based on the notion that the conditional distribution of y can be linked to $\mathbf{x}$ through linear combinations of $\mathbf{x}$. This approach refrains from complete zeroing out of a coordinate of $\mathbf{x}$, instead attaches different weights to the coordinates to suitably calibrate their relative influence on y. Such a summary of $\mathbf{x}$ is quite appealing in studies where tens of predictor variables are hand picked, most possessing a moderate ability to predict y. In such cases, subspace projection is also likely to offer better out of sample prediction than sparse variable selection which may zero out many of the variables with mild effects. We demonstrate this phenomenon with the Boston house price data (Harrison and Rubinfeld 1978) in Section 5.

The formulation (1) where $p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$ is non-parametric, has been well researched in the non-Bayesian literature. However, the focus of this research has been estimation of only the minimal $\mathcal{S}$ for which (1) holds. This minimal subspace, commonly referred to as the *central subspace*, exists and is unique under mild conditions (see Section 3) . Estimation of the central subspace is often referred to as *sufficient dimension reduction* for which a number of non-Bayesian techniques now exist in the literature; see Li (1991); Cook and Weisberg (1991) for early developments and Zhu and Zeng (2006); Xia (2007) for recent accounts. In contrast, our Bayesian model provides a joint inference on $\mathcal{S}$ and $p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$, facilitating dimension reduction and prediction simultaneously. This simultaneous approach correctly reflects the uncertainty about both these parameters when producing a summary of either object or when predicting y at a new value of $\mathbf{x}$. Moreover, it leads to a consistent estimation of the central subspace (see Theorem 3.2) under far milder conditions than those needed by the non-Bayesian approaches and offers better empirical performance as illustrated in Section 5. Our approach also provides a model based choice of the dimension of $\mathcal{S}$ and we demonstrate in Section 5 that this choice relates well to out of sample prediction.

We demonstrate in the following sections that the LGP combines well with subspace projection (SP) to produce a theoretically sound and computationally tractable statistical procedure for estimating conditional densities, prediction of response and inference on predictor importance. Hereafter, we shall call this procedure spLGP. Section 2 gives details of the model and Section 3 establishes its theoretical properties in terms of predictive accuracy and consistent estimation of $\mathcal{S}$. In Section 4 we discuss in detail model fitting via Markov chain Monte Carlo, prediction and estimation of $\mathcal{S}$ and its dimensionality. Section 5 presents three simulation studies highlighting empirical performance of spLGP in estimating $\mathcal{S}$. These are followed by two real data examples. The first (Tecator data) shows a superior out of sample prediction by spLGP than previously recorded studies of this data set. In the second (Boston house price) we illustrate how subspace projection may offer improved prediction over variable selection when dealing with correlated predictor variables many of which have a mild effect on the response. In this example we also discuss calibration of predictor importance that correctly reflects posterior uncertainty.

2 Model

Below we describe the spLGP model specific to a given dimension $d = \dim(\mathcal{S})$. A value of d can be fixed by considering the number of linear summaries of $\mathbf{x}$ one is willing to entertain in prediction of y. A small value of $d = 2$ or 3 is often appealing as a parsimonious, nonparametric extension of single summary models such as the linear or generalized linear models. In latter sections we shall discuss how to select a dimension from a range of candidate values by comparing the spLGP models corresponding to these values. A more ambitious approach is to include d as a model parameter and fit an overarching model that spans over a given range of its values. However, this is not pursued in the current paper due to the additional computing challenges involved in fitting such models.

For $\dim(\mathcal{S}) = d$, we model $\mathcal{S}$ as a uniform random variate in $\mathcal{G}_{p,d}$ - the set of all d-dimensional linear subspaces of $\mathbb{R}^p$. Note that any $\mathcal{S} \in \mathcal{G}_{p,d}$ can be expressed as the linear span $\mathcal{C}(\mathbf{B})$ of the columns of some $p \times d$ matrix $\mathbf{B} \in \mathcal{B}_d = \{\mathbf{B} = [\mathbf{b}_1 : \cdots : \mathbf{b}_d] : \mathbf{b}_j \in \mathbb{R}^p, \|\mathbf{b}_j\| = 1, j \geq 1, \mathbf{Ba} = 0 \implies \mathbf{a} = 0\}$. It turns out that if $\mathbf{B}$ is a uniform random variate in $\mathcal{B}_d$ then $\mathcal{S} = \mathcal{C}(\mathbf{B})$ defines a uniform random variate in the Grassmannian $\mathcal{G}_{p,d}$; see Lemma 6.1 in Appendix (see also Mattila 1995, Ch 3 for more on Grassmannian manifolds). This allows us to work directly with the representation $\mathbf{B}$. However, since $\mathbf{B}$ does not define a unique basis of $\mathcal{S} = \mathcal{C}(\mathbf{B})$, all inference about $\mathcal{S}$ will be done with respect to $\mathcal{C}(\mathbf{B})$ or $\mathbf{P_B} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}' = \mathbf{P}_\mathcal{S}$ – the unique orthogonal projection matrix onto this space.

With $\mathbf{B} \in \mathcal{B}_d$ representing a basis of $\mathcal{S}$, the spLGP model can be written as

$$p(y \mid \mathbf{x}) = f(\mathbf{B}'\mathbf{x}, y) \tag{2}$$

for some f in $\mathcal{F}_d = \{f : \mathbb{R}^d \times \mathbb{R} \to (0, \infty) : \int f(\mathbf{z}, y)dy = 1 \ \forall \mathbf{z} \in \mathbb{R}^d\}$. We specify a prior distribution Π_d on $(\mathbf{B}, f)$ as $\Pi_d = \Pi_d^{\mathbf{B}} \times \Pi_d^f$ where $\Pi_d^{\mathbf{B}}$ is the uniform distribution over $\mathcal{B}_d$ and Π_d^f is a logistic Gaussian process prior over $\mathcal{F}_d$ defined below. The definition of Π_d^f closely follows the one in Tokdar (2007, Section 6) where a Gaussian process defined over a compact subset of $\mathbb{R}^d$ is suitably transformed to produce a random density function with support $\mathbb{R}^d$.

Let $W(\mathbf{s}, t)$ be an almost surely continuous, zero-mean, Gaussian process indexed by $\mathbf{s} = (s_1, \cdots, s_d) \in [-1, 1]^d$, $t \in [0, 1]$ with the square exponential covariance kernel

$$\mathbb{C}\mathrm{ov}(W(\mathbf{s}_1, t_1), W(\mathbf{s}_2, t_2)) = \tau^2 \exp\left(-\sum_{j=1}^{d} \beta_j^2 (s_{1j} - s_{2j})^2 - \beta_y^2 (t_1 - t_2)^2 \right)$$

parametrized by $\tau > 0$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d) \in (0, \infty)^d$ and $\beta_y > 0$. Fix a density g_0 on $\mathbb{R}$ and let G_0 denote its cumulative distribution function. Also fix continuous, monotonically increasing functions F_j, $1 \leq j \leq d$, from $\mathbb{R}$ onto $(-1, 1)$ and let $\mathbf{F}(\mathbf{z})$ denote $(F_1(z_1), \cdots, F_d(z_d))'$. We define Π_d^f to be the distribution governing the stochastic process

$$f(\mathbf{z}, y) = g_0(y) \frac{e^{W(\mathbf{F}(\mathbf{z}), G_0(y))}}{\int e^{W(\mathbf{F}(\mathbf{z}), v)} dv} \tag{3}$$

with a suitable prior distribution $H(\tau, \boldsymbol{\beta}, \beta_y)$ on the hyper-parameters $(\tau, \boldsymbol{\beta}, \beta_y)$. Note that $p(y \mid \mathbf{z}) = f$ is equivalent to saying $p(G_0(y) \mid \mathbf{F}(\mathbf{z})) = f_W$ where $f_W(\mathbf{s}, t) = e^{W(s,t)} / \int_0^1 e^{W(\mathbf{s}, u)} du$.

The compactification provided by $\mathbf{F}$ and G_0 is useful in many ways. It allows us to work with a Gaussian process W defined on a compact set. The known continuity, maxima and support properties of these processes make it easy to investigate the asymptotic properties of spLGP. It also ensures that the oscillations in the sample paths of f tapers off to zero at infinity. The identity $\mathbb{E} \log f(\mathbf{z}, y) = \log g_0(y)$ represents g_0 as a central sample path under Π_d^f. As discussed in Tokdar (2007), it is possible to extend g_0 to incorporate a parametric model of dependence of y on $\mathbf{z}$. For example, by defining $g_0(\mathbf{z}, y) = \text{Normal}(y \mid \mu_0 + \mu'\mathbf{z}, \sigma^2)$ one can embed the multiple linear regression at the center of Π_d^f. The simpler choice of a fixed g_0 avoids complication in posterior computation.

We make data driven choices of g_0 and $\mathbf{F}$. Theorem 3.1 requires g_0 to be heavy tailed. We set it equal to $t_3(\bar{y}, s_y/1.134)$ where $\bar{y}$ and s_y are the sample mean and standard deviation of y_i's. Note that the third quartile of a standard t_3 distribution equals the third quartile of a $\text{Normal}(0, 1.134^2)$ distribution. To fix F_j's, we first assume that each coordinate variable of $\mathbf{x}$ is scaled to have mean zero and variance 1. Each F_j is then set to be $F_j(z_j) = z_j/(A+\delta)$ for $|z_j| < A$ and is left unspecified outside $[-A, A]$, where $A = \max_{1\le i\le n} \|\mathbf{x}_i\|$. Outside $[-A, A]$, each F_j is hypothesized to be extended properly to satisfy the constraints discussed before.

The linearity of each F_j in $[-A, A]$ offers an important simplification in computing the data likelihood. It turns out that each column $\mathbf{b}_j$ of $\mathbf{B}$ appears only in the form of $\beta_j^2(\mathbf{x}_i'\mathbf{b}_j - a)^2$. This allows us to reparametrize the pair $(\beta_j, \mathbf{b}_j)$ as a single vector $\boldsymbol{\beta}_j^{\mathbf{x}} = \beta_j \mathbf{b}_j$. Note that the unit-norm constraint on $\mathbf{b}_j$ ensures that the original parametrization can be easily retrieved as $\beta_j = \|\boldsymbol{\beta}_j^{\mathbf{x}}\|$ and $\mathbf{b}_j = \boldsymbol{\beta}_j^{\mathbf{x}}/\|\boldsymbol{\beta}_j^{\mathbf{x}}\|$. This reparametrization gets rid of the unit-norm condition and allows us model $\boldsymbol{\beta}_j^{\mathbf{x}}$'s as independent multivariate normal variates with mean zero and covariance σ^2 times the identity matrix. Under this specification $\mathbf{b}_j$'s are indeed independently distributed according to the uniform distribution on the p-dimensional sphere. Moreover, β_j^2's are independently distributed according to the gamma density with shape $p/2$ and rate $\sigma^2/2$ and are independent of $\mathbf{b}_j$'s.

It remains to specify prior distributions on τ^2, σ^2 and β_y. We model these parameters independently with inverse chi-square prior distributions on τ^2 and σ^2, with parameters (ν_τ, τ_0^2) and (ν_σ, σ_0^2) respectively; see Gelman et al. (2004) for notations. An *extreme gamma* distribution is specified on β_y under which $\exp(\beta_y) - 1$ follows a gamma distribution with shape ν_β and scale μ_β. This ensures a thin tail on β_y, $\Pr(\beta_y > b) = O(\exp(-\exp(b)))$, providing a safeguard against large values of β_y which can lead to undesirable overfitting. See also Remark 1 of Ghosal and Roy (2006) where such thin tail distributions are recommended to meet certain technical requirements. For our choice of covariance kernel these requirements are also met with a simple gamma distribution, as has been specified on β_j's.

3 Posterior Consistency

We investigate asymptotic frequentist properties of spLGP for a given $\dim(\mathcal{S}) = d$ under the assumption that $p(y \mid \mathbf{x}) = p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$ for some $\mathcal{S}_0 \in \mathcal{G}_{p,d}$. We first show that our model on $(\mathcal{S}, p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$ assigns positive prior probabilities around $(\mathcal{S}_0, p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$, and consequently the posterior on $p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$ concentrates around small neighborhoods of $p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$ as more data accumulate from $(\mathbf{x}, y) \sim p(\mathbf{x})p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$, where $p(\mathbf{x})$ is the density of $\mathbf{x}$. We next explore consistency in estimating $\mathcal{S}$. For this we need to assume that the given dimension d is the minimum dimension for which (1) holds. Such an assumption is necessary because for any higher dimension, $\mathcal{S}_0$ is not unique, and thus its estimation is not meaningful. When the density $p(\mathbf{x})$ is positive on all of $\mathbb{R}^p$, a minimum d exists and the corresponding $\mathcal{S}_0$ is unique (Cook 1994). This $\mathcal{S}_0$ is precisely the central subspace, i.e., $\mathcal{S}_0 = \cap\mathcal{S}$ where the intersection is taken over all linear subspaces $\mathcal{S} \subset \mathbb{R}^p$ for which (1) holds. Derivation of the asymptotic properties of the posterior on $\mathcal{S}$ when the specified d does not match the minimum dimension remains an open problem. We note that the existing theory for non-Bayesian approaches to estimating the central subspace also require the knowledge of the dimension (Xia 2007).

Under our assumption, there exists $(\mathbf{B}_0, f_0) \in \mathcal{B}_d \times \mathcal{F}_d$ satisfying the model assumption in (2). We will show that the posterior distribution

$$\Pi_d^{(n)}(d\mathbf{B}, df) \propto \prod_{i=1}^{n} f(\mathbf{B}'\mathbf{x}_i, y_i)\Pi_d(d\mathbf{B}, df)$$

concentrates around small neighborhoods of $(\mathbf{B}_0, f_0)$ as $n \to \infty$. The neighborhoods, however, are to be carefully defined since $(\mathbf{B}, f)$ is not uniquely determined under (2). We first consider neighborhoods derived from a topology on the space of joint densities of $(\mathbf{x}, y)$:

$$L_\epsilon = \{(\mathbf{B}, f) \in \mathcal{B}_d \times \mathcal{F}_d : \int \|f(\mathbf{B}'\mathbf{x}, \cdot) - f_0(\mathbf{B}_0'\mathbf{x}, \cdot)\|_1 p(\mathbf{x}) d\mathbf{x} < \epsilon\}$$

where $\|\cdot\|_1$ denotes the L_1 norm and $\epsilon > 0$ is arbitrary. In the following $\mathrm{KL}(g, f)$ denotes the Kullback-Leibler divergence $\int g \log(g/f)$ of a density f from another density g defined with respect to the same dominating measure.

Theorem 3.1. *Assume that*

1. *f_0 is continuous in both arguments.*
2. *There exists $a > 0$ such that $\mathrm{KL}(f_0(\mathbf{z}_1, \cdot), f_0(\mathbf{z}_2, \cdot)) < a\|\mathbf{z}_1 - \mathbf{z}_2\|^2$.*
3. *$\int \|\mathbf{x}\|^2 p(\mathbf{x}) d\mathbf{x} < \infty$.*
4. *For all $\mathbf{B} \in \mathcal{B}_d$, $\int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \left|\log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{g_0(y)}\right| dy d\mathbf{x} < \infty$.*
5. *For every $\mathbf{z} \in \mathbb{R}^d$,*

$$\lim_{y \to -\infty} \frac{f_0(\mathbf{z}, y)}{g_0(y)} = 0, \lim_{y \to \infty} \frac{f_0(\mathbf{z}, y)}{g_0(y)} = 0$$

and these convergences are uniform over every compact subset of $\mathbb{R}^d$.

Then $\Pi_d^{(n)}(L_\epsilon) \to 1$ almost surely as $n \to \infty$.

This result holds because the induced prior distribution on the joint density $h(\mathbf{x}, y) = p(\mathbf{x}) f(\mathbf{B}'\mathbf{x}, y)$ satisfies the Kullback-Leibler support condition and the L_1 entropy condition of Ghosal et al. (1999); a sketch of a proof is given in Appendix. The continuity assumption (a) and the heavy tail assumption (e) together ensure that the map $(\mathbf{s}, t) \mapsto \log f_0(\mathbf{F}^{-1}(\mathbf{s}), G_0^{-1}(t)) - \log g_0(\mathbf{F}^{-1}(\mathbf{s}), G_0^{-1}(t))$ is well approximated by a continuous function on $[-1, 1]^d \times [-1, 1]$. This approximating function belongs to the supremum-norm support of the Gaussian process W – a basic result required to verify the Kullback-Leibler condition. The other assumptions make sure that the error in approximation is negligible in the tails.

To measure accuracy in estimating $\mathcal{S}$, we need to focus on neighborhoods derived from a topology on the space of d dimensional linear subspaces of $\mathbb{R}^p$. As any such subspace is uniquely represented by the rank d (orthogonal) projection matrix associated with it, we can simply work with a distance metric on the latter objects. Two distance metrics would be considered here:

$$\begin{aligned} \rho_{\text{trace}}(\mathbf{P}_1, \mathbf{P}_2) &= [1 - \text{trace}(\mathbf{P}_1\mathbf{P}_2)/d]^{1/2} && (4) \\ \rho_{\text{op}}(\mathbf{P}_1, \mathbf{P}_2) &= \|\mathbf{P}_1 - \mathbf{P}_2\|_2 && (5) \end{aligned}$$

where $\|\cdot\|_2$ denotes the operator norm for linear transforms from $\mathbb{R}^p$ to $\mathbb{R}^p$. The first metric measures an average distance between the projections of $\mathbf{x}$ under the two matrices. In particular when $\mathbf{x}$ has mean zero and covariance identity,

$$\mathbb{E}\|\mathbf{P}_1\mathbf{x} - \mathbf{P}_2\mathbf{x}\|^2 = 2d\rho_{\text{trace}}(\mathbf{P}_1, \mathbf{P}_2)^2 \tag{6}$$

The second metric measures the maximum possible distance between the projections of a normalized $\mathbf{x}$ under the two matrices. Given a distance metric ρ on the rank d projection matrices of $\mathbb{R}^p$, a neighborhood of $\mathbf{B}_0$ is defined as follows

$$R_{\rho,\delta} = \{\mathbf{B} \in \mathcal{B}_d : \rho(\mathbf{P}_\mathbf{B}, \mathbf{P}_0) < \delta\}$$

where $\mathbf{P}_\mathbf{B} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ denotes the projection matrix onto the subspace spanned by the columns of $\mathbf{B}$ and $\mathbf{P}_0 = \mathbf{P}_{\mathbf{B}_0}$.

Theorem 3.2. *If $0 < p(\mathbf{x}) < \bar{p}_\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^p$ for some $\bar{p}_\mathbf{x} < \infty$, then under the assumptions of Theorem 3.1, $\Pi^{(n)}(R_{\rho,\delta}) \to 1$ almost surely as $n \to \infty$ where $\rho = \rho_{\text{trace}}$ or $\rho = \rho_{\text{op}}$. Moreover, if $\hat{\mathbf{B}}_n$ is a minimizer of $\text{width}(\mathbf{B}) = \int \rho_{\text{trace}}(\mathbf{P}_{\mathbf{B}_1}, \mathbf{P}_\mathbf{B})^2 \Pi_d^{(n)}(d\mathbf{B}_1)$, then $\rho_{\text{trace}}(\mathbf{P}_{\hat{\mathbf{B}}_n}, \mathbf{P}_0) \to 0$ almost surely as $n \to \infty$.*

This result follows from the fact that $(\mathbf{B}, f)$ cannot be close to $(\mathbf{B}_0, f_0)$ in the L_1 topology defined earlier, unless $\rho(\mathbf{P}_\mathbf{B}, \mathbf{P}_0)$ is small; see the appendix for a proof. The positivity of $p(\mathbf{x})$ and continuity of f_0 together ensure that $\mathbf{P}_{\mathbf{B}_1} \neq \mathbf{P}_{\mathbf{B}_2}$ implies $\|f_0(\mathbf{B}_1'\mathbf{x}, \cdot) - f_0(\mathbf{B}_2'\mathbf{x}, \cdot)\|_1 > 0$ with a positive probability under p.

Remark Note that the first assertion of Theorem 3.2 stays valid if we take ρ to be the metric induced by the Frobenius norm, which essentially is the Euclidean distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ seen as elements in $\mathbb{R}^{p \times p}$. Both ρ_{trace} and ρ_{op} are dominated by this metric. The second assertion is valid for both ρ_{op} and the Frobenius distance. However, defining $\hat{\mathbf{B}}_n$ in terms of ρ_{trace} has a certain computational advantage as would be discussed later.

4 Model Fitting and Estimation

4.1 Sampling from the Posterior

As indicated in Tokdar (2007), working directly with the Gaussian process W poses serious problems to model fitting even for numerical methods like the Markov Chain Monte Carlo; see, however Adams, Murray, and MacKay (2009) for some recent developments on this issue. Instead, by replacing W in (3) with the conditional expectation process

$$Z_{\mathcal{N}}(\mathbf{s},t)=\mathbb{E}[W(\mathbf{s},t)\mid W(\mathbf{s}_1,t_1),\cdots,W(\mathbf{s}_k,t_k)], \tag{7}$$

one obtains a useful approximation to the prior distribution Π_d^f that tracks randomness of W only at the node set $\mathcal{N}=\{(\mathbf{s}_1,t_1),\cdots,(\mathbf{s}_k,t_k)\}\subset[-1,1]^d\times[0,1]$. In fact $Z_{\mathcal{N}}$ can be simply written as

$$Z_{\mathcal{N}}(\mathbf{s},t)=\sum_{m=1}^{k}\lambda_m K((\mathbf{s},t),(\mathbf{s}_m,t_m)\mid\boldsymbol{\beta},\beta_y)$$

where $K((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2)\mid\boldsymbol{\beta},\beta_y)=\exp(-\sum_{j=1}^{d}\beta_j^2(s_{1j}-s_{2j})^2-\beta_y^2(t_1-t_2)^2)$ and $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_k)'$, given $\boldsymbol{\beta}$, β_y and τ, has a multivariate normal distribution with mean zero and covariance $\tau^2\mathbf{K}(\mathcal{N}\mid\boldsymbol{\beta},\beta_y)^{-1}$ with the (m,l)-th element of $\mathbf{K}(\mathcal{N}\mid\boldsymbol{\beta},\beta_y)$ given by $K((\mathbf{s}_m,t_m),(\mathbf{s}_l,t_l)\mid\boldsymbol{\beta},\beta_y)$, $1\le m,l\le k$. A similar approximation is used by Tokdar (2007) for LGP models of a single density and by Banerjee, Gelfand, Finley, and Sang (2008) for Gaussian process spatial models.

With the node set $\mathcal{N}$ fixed, model fitting with the approximate prior distribution is done by running a Markov Chain to update the parameters $(\mathbf{B},\boldsymbol{\beta},\beta_y,\boldsymbol{\lambda},\tau^2,\sigma^2)$. Note that $\mathbf{B}$ enters the likelihood computation only through evaluation of $Z_{\mathcal{N}}$ at $(\mathbf{F}(\mathbf{B}'\mathbf{x}_i),t)$ for $1\le i\le n$ for various $t\in[0,1]$. Since each $|\mathbf{B}'\mathbf{x}_i|\le A$, linearity of each F_j in $[-A,A]$ ensures that these evaluations can be made by computing $\sum_{j=1}^{d}\beta_j^2(\mathbf{b}_j'x_i-\mathbf{s}_m)^2$ for $1\le m\le k$. This simplification was used in Section 2 to collapse $\boldsymbol{\beta}$ and $\mathbf{B}$ into a single block of parameters $\mathbf{B}^{\mathbf{x}}=[\boldsymbol{\beta}_1^{\mathbf{x}}:\cdots:\boldsymbol{\beta}_d^{\mathbf{x}}]$. Note that in updating $(\mathbf{B}^{\mathbf{x}},\beta_y,\boldsymbol{\lambda},\tau^2,\sigma^2)$, one achieves further simplification by integrating out τ^2 and σ^2 from the model. The prior distribution on $(\mathbf{B}^{\mathbf{x}},\beta_y,\boldsymbol{\lambda})$ can be written as

$$t_{\nu_\sigma}(\mathbf{B}^{\mathbf{x}}\mid 0,\sigma_0^2\mathbf{I}_{dp})\mathrm{ExGam}(\beta_y\mid\nu_y,\mu_y)t_{\nu_\tau}(\boldsymbol{\lambda}\mid 0,\tau_0^2\mathbf{K}(\mathcal{N}\mid(\|\boldsymbol{\beta}_1^{\mathbf{x}}\|,\cdots,\|\boldsymbol{\beta}_d^{\mathbf{x}}\|),\beta_y)^{-1})$$

where $\mathbf{B}^{\mathbf{x}}$ is seen as a vector in $\mathbb{R}^{dp}$, $\mathbf{I}_{dp}$ is the identity matrix of dimension dp, $t_\nu(\cdot\mid\mu,\Sigma)$ denotes the multivariate t distribution with degrees of freedom ν, location μ and scale Σ, and $\mathrm{ExGam}(\cdot\mid\nu,\mu)$ denotes the extreme gamma distribution with shape ν and scale μ.

In our implementation, we make a data driven choice of $\mathcal{N}$, by augmenting the parameter set to $(\mathbf{N}_x,\mathbf{n}_y,\mathbf{B}^{\mathbf{x}},\beta_y,\boldsymbol{\lambda})$, where $\mathbf{N}_x=[\mathbf{s}_1:\cdots:\mathbf{s}_k]$ and $\mathbf{n}_y=(t_1,\cdots,t_k)$. The columns of $\mathbf{N}_x$ are modeled as independent, uniform random variates in $[-1,1]^d$, $\mathbf{n}_y$ is modeled as a uniform random variate in $[0,1]^k$ and these are taken to be independent of each other and also of the remaining parameters. We run a block Metropolis-Hastings sampler where each of $\mathbf{N}_x$, $\mathbf{n}_y$, $\mathbf{B}^{\mathbf{x}}$, β_y and $\boldsymbol{\lambda}$ is updated as a block holding the rest fixed. A current realization $\mathbf{B}^{\mathbf{x},\text{curr}} = ((b_{ij}^{\mathbf{x},\text{curr}}))$ is updated by proposing a new realization $\mathbf{B}^{\mathbf{x},\text{prop}} = ((b_{ij}^{\mathbf{x},\text{prop}}))$ as $b_{ij}^{\mathbf{x},\text{prop}} = b_{ij}^{\mathbf{x},\text{curr}} + \sigma_B \epsilon_{ij}$ where ϵ_{ij} are independent Normal$(0,1)$ variates. Here σ_B is a scalar that controls the magnitude of these perturbations. Similar multivariate normal perturbations are used for $\boldsymbol{\lambda}$, $\mathbf{N}_x$ and $\mathbf{n}_y$, with σ_λ, σ_x and σ_y controlling the respective perturbation magnitude. For the latter two blocks, however, perturbation is applied to the constraint-free arctan and logit transforms of their elements. A univariate normal perturbation is used on the logarithm of β_y: $\log \beta_y^{\text{prop}} = \log \beta_y^{\text{curr}} + \sigma_\beta \epsilon$ where $\epsilon \sim$ Normal$(0,1)$. Tuning of σ_B, σ_λ, σ_x, σ_y and σ_β to attain desired levels of acceptance of the corresponding updates is discussed in detail in Example 1. The cardinality k of $\mathcal{N}$ is taken to be fixed, usually between 5 to 10. Note that the computing cost of evaluating the likelihood function at a given parameter configuration is roughly $O(nk^3 g)$ where g is the number of grid points on $[0,1]$ used to approximate the integral in (3).

4.2 Posterior Summaries

Once a sample $(\mathbf{B}_l, f_l)$, $1 \le l \le L$ is obtained from the (approximate) posterior distribution, it is easy to estimate the predictive conditional density of y at a given covariate value $\mathbf{x}^*$ by the Monte Carlo average

$$\hat{p}(y \mid \mathbf{x} = \mathbf{x}^*) = \frac{1}{L} \sum_{l=1}^{L} f_l(\mathbf{B}_l' \mathbf{x}^*, y).$$

In our implementation, this density is evaluated only on a finite number of grid points, from which its quantiles can be easily approximated. For making a prediction at $\mathbf{x}^*$, we use the median of this predictive conditional density, and its 50% (and 95%) equal tailed intervals are used to indicate possible spread.

Approximating $\hat{\mathbf{B}}_n$ – the Bayes estimate under the ρ^2_{trace} loss – however, requires more attention. Let $\mathbf{z}_{li}$ denote the projection $\mathbf{P}_{\mathbf{B}_l}\mathbf{x}_i$. For any $\mathbf{B} \in \mathcal{B}_d$, the identity (6) implies that a Monte Carlo approximation to width$(\mathbf{B}) = \int \rho^2_{\text{trace}}(\mathbf{B}, \tilde{\mathbf{B}}) \Pi_d^{(n)}(d\tilde{\mathbf{B}})$ is given by

$$\widehat{\text{width}}(\mathbf{B}) = \frac{1}{2dnL} \sum_{i=1}^{n} \sum_{l=1}^{L} \|\mathbf{z}_{li} - \mathbf{P}_{\mathbf{B}}\mathbf{x}_i\|^2.$$

A simple bias-variance argument shows that minimizing $\widehat{\text{width}}(\mathbf{B})$ is equivalent to minimizing $\sum_{i=1}^{n} \|\bar{\mathbf{z}}_i - \mathbf{P}_{\mathbf{B}}\mathbf{x}_i\|^2$ with $\bar{\mathbf{z}}_i = (1/L)\sum_{l=1}^{L} \mathbf{z}_{li}$. We find $\hat{\mathbf{B}}_n$ by carrying out this latter minimization restricted to the set of the sampled $\mathbf{B}_l$ values. Note that the computing cost of this approach of estimating $\hat{\mathbf{B}}_n$ from the MCMC sample is linear in the sample size L. This would not have been the case if ρ_{trace} were to be replaced with ρ_{op} in the definition of $\hat{\mathbf{B}}_n$; that would have resulted in a computing cost quadratic in L. Once $\hat{\mathbf{B}}_n$ is computed, a measure of *standard error* in estimating $\mathcal{S}_{y|\mathbf{x}}$ is found in $\widehat{\text{width}}(\hat{\mathbf{B}}_n)$. While presenting $\hat{\mathbf{B}}_n$, we shall rearrange its columns in decreasing order of β_j – stored from the same MCMC draw that produced $\hat{\mathbf{B}}_n$. The idea here is to bring forward the column which corresponds to most rapid changes in the conditional densities $f(\mathbf{z}, \cdot)$, much in the spirit of automatic relevance detection (Neal 1996). This idea is further explored in Example 5 of Section 5 in the context of assessing relative importance of the predictor variables.

4.3 Selecting $\dim(\mathcal{S})$

It is natural to ask what minimum dimension of $\mathcal{S}$ is needed to well approximate $p(y \mid \mathbf{x})$ by $p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$. One can pose this question as a model selection problem and compare the marginal data likelihood

$$m(d) = \int \prod_{i=1}^{n} f(\mathbf{B}'\mathbf{x}_i, y_i)\Pi_d(d\mathbf{B}, df) \tag{8}$$

of the spLGP models for a set of candidate values $1 \le d \le d_{\max}$ for some prespecified $d_{\max} \le p$, and then select the one with the maximum $m(d)$ value. This procedure amounts to modeling $\dim(\mathcal{S})$ with a discrete uniform prior distribution over $\{1, \cdots, d_{\max}\}$; suitable adaptations can be used if other prior distributions are to be considered. To tackle the difficult numerical problem of evaluating $m(d)$, we follow Chib and Jeliazkov (2001), who proposed a Monte Carlo approximation to marginal likelihoods based on sequential reduced runs of a block Metropolis-Hastings sampler.

One disadvantage of using the above formal approach is the increased computing cost. The Chib and Jeliazkov method requires five additional runs of the MCMC sampler, one for each of $\mathbf{N}_x$, $\mathbf{n}_y$, $\mathbf{B}^{\mathbf{x}}$, β_y and $\boldsymbol{\lambda}$. Moreover, likelihood evaluations are needed for two perturbed versions of the saved samples from each of these runs. This motivates us to develop a quicker alternative to judge plausibility of different candidate values of $\dim(\mathcal{S})$. This works directly with the MCMC output from the model fitting stage, no additional runs are required. On the flip side, it requires model fits across a set of contiguous candidate values even while comparing between just two candidates. We recommend this approach more as an ad-hoc, quick visualization tool to be used as a first step toward estimating a minimum $\dim(\mathcal{S})$. Details of this procedure are given below.

For a candidate value d,

$$\mathrm{rad}(d, \alpha) = \inf\{r : \Pi_d^{(n)}(\rho_{\mathrm{trace}}(\mathbf{P}_{\mathbf{B}}, \mathbf{P}_{\hat{\mathbf{B}}_n}) > r) \le \alpha\}$$

gives a measure of the concentration of $\Pi_d^{(n)}$ around the point estimate of $\mathcal{S}$ it produces. Theorem 3.2 says that $\mathrm{rad}(d_0, \alpha)$ would be asymptotically quite small for any $\alpha \in (0, 1)$ if there is a unique $\mathcal{S}_0$ of dimension d_0 satisfying $p(y \mid \mathbf{x}) = p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$, i.e., $\mathcal{S}_0$ is the central subspace and d_0 is the minimum dimension for which (1) holds. Let $\hat{\mathbf{B}}_n^*$ denote the Bayes estimate $\hat{\mathbf{B}}_n$ corresponding to $d = d_0$. Now consider the case where a $d > d_0$ is specified. We will work under the assumption that this larger model will put posterior mass around subspaces $\mathcal{S}$ which contain $\mathcal{S}_0$ as a further subspace. A basis $\mathbf{B}$ of such an $\mathcal{S}$ must be of the form $\mathbf{B} = [\hat{\mathbf{B}}_n^* : \mathbf{B}^\dagger]\mathbf{M}$, where $\mathbf{B}^\dagger \in \mathcal{B}^{d-d_0}$ is linearly independent of $\hat{\mathbf{B}}_n^*$ and $\mathbf{M}$ is a $d \times d$ non-singular matrix. The maximum ρ_{trace} separation attainable for two such $\mathbf{B}$ matrices is $\overline{\text{rad}}(d_0, d) = \sqrt{1 - d^+/d}$ where $d^+ = d_0 + \max(0, 2d - p - d_0)$. Since it is likely that $\text{rad}(d, \alpha)$, for small α, would be close to this maximum ρ_{trace} separation, $\overline{\text{rad}}(d_0, d)$ provides a ballpark value for $\text{rad}(d, \alpha)$ for $d > d_0$. Note that $\overline{\text{rad}}(d_0, d)$ monotonically increases in $d \geq d_0$ until d exceeds $(p - d_0)/2$ and monotonically decreases to zero after that.

The above heuristic argument leads to the following procedure for estimating $\dim(\mathcal{S})$, where for every candidate d, we compare $\text{rad}(\tilde{d}, \alpha)$, $\tilde{d} > d$ with the ballpark value $\overline{\text{rad}}(d, \tilde{d})$ and choose that d for which these two quantities are close to each other. Formally, fix a small $\alpha > 0$ such as $\alpha = 0.05$. For each candidate value $d \in \{1, \cdots, d_{\max}\}$, approximate $\text{rad}(d, \alpha)$ by the $(1 - \alpha)\%$ quantile of $\rho_{\text{trace}}(\mathbf{B}_l, \hat{\mathbf{B}}_n)$ where $\mathbf{B}_l$ are the saved MCMC draws of $\mathbf{B}$. For each candidate d compute the average deviance between $\{\text{rad}(\tilde{d}, \alpha)\}_{d \leq \tilde{d} \leq d_{\max}}$ and $\{\overline{\text{rad}}(d, \tilde{d})\}_{d \leq \tilde{d} \leq d_{\max}}$ by

$$\text{dev}(d, \alpha) = \sum_{d_1 = d}^{d_{\max}} |\text{rad}(d_1, \alpha) - \overline{\text{rad}}(d, d_1)| / (d_{\max} - d + 1) \tag{9}$$

and select $\dim(\mathcal{S}) = \arg\min_{1 \leq d \leq d_{\max}} \text{dev}(d, \alpha)$. In Section 5 we demonstrate satisfactory empirical performance of this heuristic selection rule, although a rigorous theoretical justification is yet to be established.

5 Numerical Illustration

Example 1. Figure 1 shows a snapshot of the output of spLGP applied to a simulated dataset with $n = 100$ and $p = 10$ where

$$p(y \mid \mathbf{x}) = \text{Normal}(\sin(\pi \mathbf{b}_0' \mathbf{x}/2), 0.1^2)$$

with $\mathbf{b}_0 = (0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0)'$ and $\mathbf{x} \sim \text{Normal}(0, \mathbf{I}_{10})$. The top row corresponds to a spLGP model fit with $d = 1$. Each of the hyper parameter ν_τ, τ_0^2, ν_σ, σ_0^2, ν_β and μ_β was fixed at 1. The cardinality of the node set $\mathcal{N}$ was fixed at $k = 10$. The MCMC sampler discussed in Section 4 was run for 100,000 iterations of which first 60,000 were thrown out as burn-in. The sampler was tuned to attain approximately 45% acceptance rate for β_y and 22% for the other blocks. The above values of the hyper parameters were used for all other examples presented in this section.

The trace plots of each coordinate of $\mathbf{B} = \mathbf{b}_1$ are shown in the top-left panel. Note that $\mathbf{B}$ appears to fluctuate around a neighborhood of $\mathbf{b}_0$ in the sense of either ρ_{trace} or ρ_{op}, but not in the sense of the Euclidean distance on $\mathbb{R}^{10}$. The top-right plot shows the contours (light grey) of $Z_{\mathcal{N}}$ and the cumulative distribution functions of the resulting f from one iteration of the sampler. The grey filled circles show the scatter plot of $G_0(y_i)$ versus $\mathbf{F}(\mathbf{B}'\mathbf{x}_i)$ where $\mathbf{B}$ is taken from the same iteration. The checked circles show the points in the corresponding node set $\mathcal{N}$.

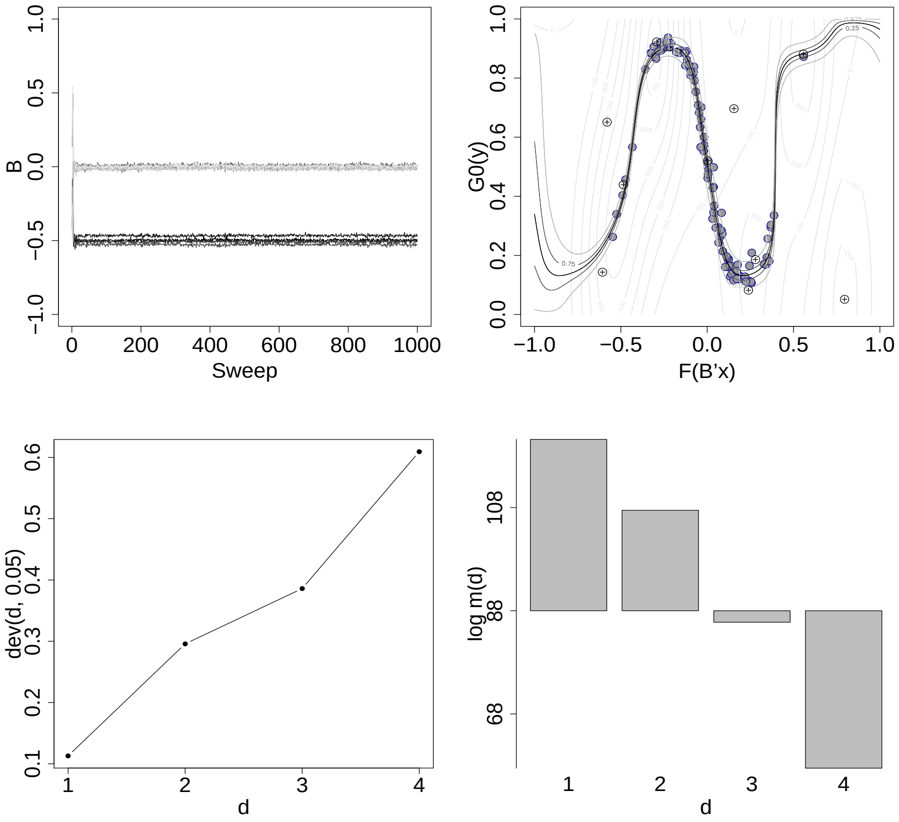

Figure 1: Example 1: spLGP in action. Top row: (left) trace plots of each coordinate of $\mathbf{B}$ from a model fit with $d = 1$ and (right) contours of $\mathcal{Z}_{\mathcal{N}}$ (light grey) and quantiles of $f_{\mathcal{Z}_{\mathcal{N}}}$ overlaid on the transformed data $(\mathbf{F}(\mathbf{B}'\mathbf{x}_i), G_0(y))$ (filled grey circles). Bottom row: plots of (left) $\text{dev}(d, 0.05)$ and (right) $\log m(d)$ against d. Both the heuristic and the formal method correctly estimated $d_0 = 1$.

a	n	spLGP	dMAVE	SIR	SAVE
2	100	0.40 (0.14)	0.46 (0.16)	0.96 (0.09)	0.90 (0.06)
2	200	0.23 (0.05)	0.28 (0.06)	0.95 (0.07)	0.87 (0.11)
2	400	0.15 (0.04)	0.19 (0.04)	0.95 (0.09)	0.85 (0.12)

Table 1: Example 2: Comparison of spLGP to other SDR methods for estimating $\mathcal{S}$.

The bottom row of Figure 1 shows the diagnostics for estimating $\dim(\mathcal{S})$. The left plot shows the deviance $\mathrm{dev}(d, 0.05)$ across candidate values 1 through 4. The right plot shows the estimates of $\log m(d)$ obtained by the Chib-Jeliazkov method against the same candidate values. Both the formal and our heuristic approach produced a correct estimation of $d_0 = 1$. The estimate $\hat{B}_n$ equaled $\hat{B}_n' =$(-0.50, -0.47, -0.51, -0.52, 0.01, -0.01, 0.00, 0.00, -0.01, 0.01).

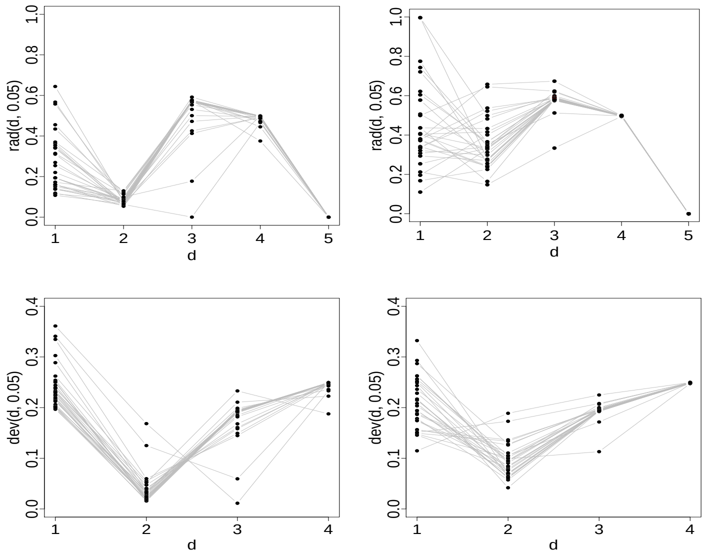

Figure 2: Example 3: Estimation of d_0 from data simulated with $d_0 = 2$. Plots of $\mathrm{rad}(d, 0.05)$ (top) and $\mathrm{dev}(d, 0.05)$ (bottom) are shown against d for 30 data sets simulated under low noise (left) and high noise (right) settings; see text. In the low noise case, detection of d_0 is sharp as $\mathrm{dev}(d, 0.05)$ shows a marked dip at $d = 2$ for most data sets, except for two cases where d_0 is overestimated to be 3. In the high noise simulations, detection is less sharp with flatter $\mathrm{dev}(d, 0.05)$ curves, and in two cases d_0 is underestimated at 1.

Example 2. To compare finite sample frequentist properties of spLGP with other methods, we tested it on the simulation setting of Example 4.2 of Xia (2007). Here, the minimal $\dim(\mathcal{S})$ is assumed to be known and the focus is only on the error in estimating $\mathcal{S}$ of the following regression problem:

$$p(y \mid \mathbf{x}) = \text{Normal}(2(\mathbf{b}_1'\mathbf{x})^a, (2\exp(\mathbf{b}_2'\mathbf{x}))^2)$$

with $\mathbf{b}_1 = (1,2,0,0,0,0,0,0,0,2)'/3$, $\mathbf{b}_2 = (0,0,3,4,0,0,0,0,0,0)'/5$, $a = 1,2$ and $\mathbf{x} \sim \text{Uniform}([0,1]^{10})$. The "spLGP" column in Table 1 shows mean (and standard deviation) of $\rho_{\text{op}}(\mathbf{B}_n, \mathbf{B}_0)$, where $\mathbf{B}_0 = (\mathbf{b}_1, \mathbf{b}_2)$, obtained from 200 data sets of size n generated from this model. We only report the case of $a = 2$, as the other case showed very similar numbers. The latter columns of Table 1 are taken from Table 2 of Xia (2007). In this simulation spLGP clearly outperformed the other methods, in particular dMAVE, which is perhaps the best non-Bayesian procedure for central subspace estimation (see the simulation studies presented in Xia 2007).

Example 3. In this example we investigated the behavior of our heuristic approach for estimating d_0. Following Zeng (2008), we generated 50 observations from the model

$$p(y \mid \mathbf{x}) = \text{Normal}(\mathbf{x}'\mathbf{b}_1 + (\mathbf{x}'\mathbf{b}_2)^2, \sigma^2)$$

where $\mathbf{x} \sim \text{Normal}(0, \mathbf{I}_5)$, $\mathbf{b}_1 = (1,1,0,0,0)'$, $\mathbf{b}_2 = (0,0,0,1,1)'$. The top left plot of Figure 2 shows $\text{rad}(d, 0.05)$ against $d = 1, \cdots, 5$ and the bottom left plot shows $\text{dev}(d, 0.05)$ against $d = 1, \cdots, 4$, for 30 data sets generated from this model with $\sigma = 0.2$ (low noise). The right plots show the same for data sets generated with a larger $\sigma = 1.0$ (noisy). In either set, for 28 out of the 30 cases, d_0 was correctly estimated to be 2. Among the incorrectly estimated cases, the low noise ones had the estimate equal to 3 – more liberal than the true value of 2, while the high noise ones had it equal to 1 – more conservative than the true value. Among the correctly estimated cases, $\text{rad}(2, 0.05)$ values are substantially larger for the examples with $\sigma = 1$, indicating a more imprecise estimation of $\mathcal{S}$ compared to their low noise counterparts. Similarly, although the *success rate* in estimating d_0 remained the same across these two examples, the flatness of the $\text{dev}(d, \alpha)$ curves in the second set indicates possibility of nearly inconclusive situations.

Example 4. The Tecator data, available at the StatLib[1] website, were collected with the aim of predicting the fat content in a meat sample from its near infrared absorbance spectrum. Data were recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. The spectra were represented via 22 principal components, of which the first 13 were recommended for use in model fitting. The training dataset contained 172 meat samples, and another 45 samples were reserved for testing prediction power.

On applying spLGP to the 172 observations in the training dataset we found $\log m(d)$ to peak at $d = 2$ and $\text{dev}(d, 0.05)$ to dip at $d = 2$ (see Figure 3, bottom left and middle plots). Thus we concluded $d_0 = 2$. The top row of Figure 3 shows the scatter plots of y against each coordinate of $\mathbf{z} = \hat{\mathbf{B}}_n'\mathbf{x}$. The bottom right plot of this figure shows the median and the 50% two-tail interval of the conditional predictive density for each of the

[1] http://lib.stat.cmu.edu/datasets/tecator

45 test cases against the actual recorded y values. Note the elongation of this interval for the two cases at the top-right corner – rightly indicating a less precise prediction for the corresponding $\mathbf{x}^*$ values. The root mean square prediction error turned out to be 0.31, which marginally bettered the previously recorded best root mean square performance (0.35) mentioned in the StatLib website.

The mean absolute deviation prediction error with the posterior medians of spLGP with $d = 1, \cdots, 4$ turned out to be 1.01, 0.25, 0.27, 0.28, thus lending external support to the estimate $d_0 = 2$.

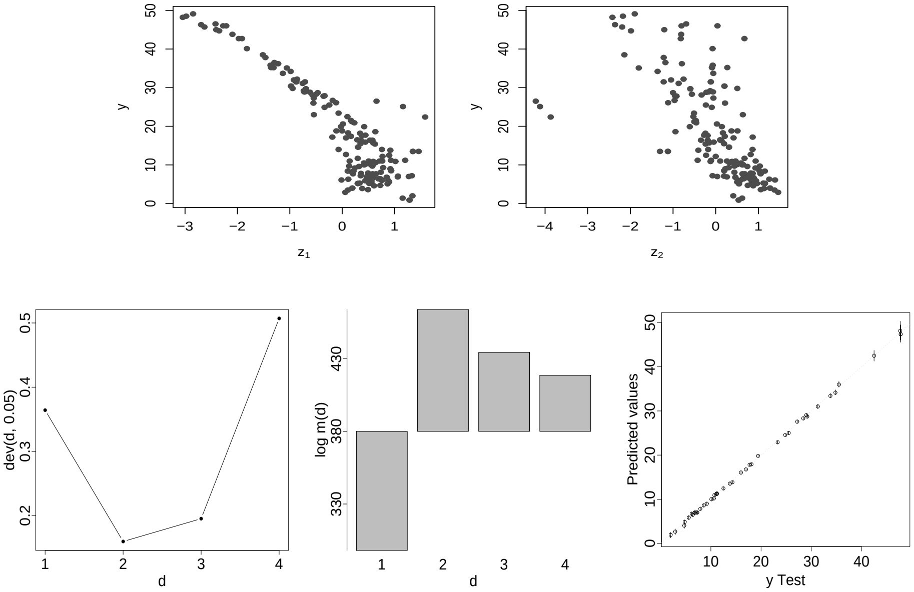

Figure 3: Tecator fat content data: Top: plots of the response versus $\mathbf{z} = \hat{\mathbf{B}}_n' \mathbf{x}$ for $d = 2$. Bottom: plots of (left) $\text{dev}(d, 0.05)$ and (middle) $\log m(d)$ – both point to $d_0 = 2$. Bottom right : Predicted values (medians) and spreads (50% two-tail intervals) versus the recorded values from the test set.

Example 5. The Boston house price data set (Harrison and Rubinfeld 1978) records the median price (in \$1000) of owner occupied homes in each of the 506 census tracts in Boston Standard Metropolitan Statistical Areas. Thirteen other variables (see Table 2) were measured for each tract capturing its socio-economic, geographic and demographic features that may have influenced this price. Chen and Li (1998) give a detailed account of analyzing this data with the classic central subspace estimation technique *sliced inverse regression* (SIR, Li 1991). Following their treatment, we removed a group of tracts with exceptionally high crime rate (`crim`) and identical records for five other variables. This left us with 374 observations which were then randomly divided into a training set

Variable	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	Variable	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
crim	0.23	0.37	-0.05	dis	0.41	0.12	0.07
zn	-0.07	0.09	-0.42	rad	-0.11	0.00	-0.24
indus	0.04	0.07	-0.16	tax	0.14	-0.06	-0.28
chas	-0.05	0.03	-0.14	ptratio	0.35	0.19	-0.31
nox	0.10	0.14	0.18	black	-0.08	-0.06	-0.16
rm	-0.59	0.85	-0.49	lstat	0.31	-0.20	0.45
age	0.39	-0.09	0.20				

Table 2: Example 5: Boston house price, columns of $\hat{\mathbf{B}}_n$ for $d = 3$.

with 249 cases and a test set with the remaining 125 cases.

On applying spLGP on the training dataset, we found $\log m(d)$ to peak at $d = 3$. The heuristic criterion was undecided between $d_0 = 2$ and $d_0 = 3$; see Figure 4 bottom left and middle plots. A fair amount of non-linear confounding among the variables (see Chen and Li 1998) makes it difficult to make a sharp estimation of a minimal $\mathcal{S}$ for this data set. For example, all the spLGP models corresponding to $d = 1, \cdots, 5$ made similar prediction on the test set; see Figure 5. A closer inspection of the mean absolute predictive errors (1.85, 1.81, 1.75, 1.68 and 1.73), however, revealed a steady improvement until $d = 4$, but the margin of improvement remained small. The choice of $d_0 = 3$ appeared to strike a reasonable balance between predictive ability and model complexity.

The top row of Figure 4 shows the scatter plot of y versus each coordinate of $\mathbf{z} = \hat{\mathbf{B}}_n\mathbf{x}$ corresponding to $d = 3$. Table 2 shows the columns of $\hat{\mathbf{B}}_n$. The first column has moderate contribution from a number of variables, e.g., the average number of rooms per unit (`rm`), the weighted average distance to five Boston employment centers (`dis`) and the proportion of units built prior to 1940 (`age`). The second column is mostly made up of `rm`, with moderate contribution from crime rate (`crim`). The third column is again an ensemble of many variables, prominent among them are `rm`, the percentage of poor (`lstat`) and the proportion of residential land zoned for large lots (`zn`). From this the average number of rooms appears as the most important factor in determining median house price, a similar conclusion was drawn by Chen and Li (1998). Note that all the variables were standardized before performing the spLGP analysis, and hence the entries in the columns of $\hat{\mathbf{B}}_n$ are scale-corrected.

Judging relative importance of predictors from $\hat{\mathbf{B}}_n$ seems rather inappropriate when there is substantial posterior uncertainty about $\mathbf{B}$ itself, as was the case with the Boston house price data. It is also crucial to take into account that all columns of $\mathbf{B}$ do not affect the conditional density equally. We address these two points as follows. For two $\mathbf{x}$-configurations $\mathbf{x}_1$ and $\mathbf{x}_2$ that differ only in the i-th coordinate by a unit amount, the covariance between $f(\mathbf{B}'\mathbf{x}_1, \cdot)$ and $f(\mathbf{B}'\mathbf{x}_2, \cdot)$ is determined by $\text{imp}_i = [\sum_{j=1}^{d} \beta_j^2 b_{ji}^2]^{1/2}$. Since covariance encodes how rapidly the conditional density is changing between these two points, imp_i serves as a measure of influence the i-th predictor variable exerts on the conditional behavior of y; see also the literature on automatic relevance detection (in

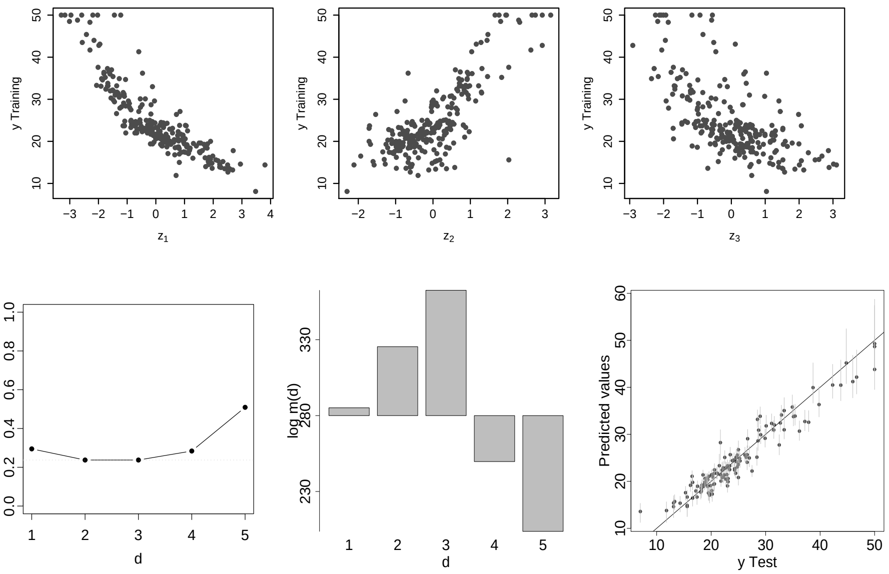

Figure 4: Boston house price: Top row: plots of the response versus $\mathbf{z} = \hat{\mathbf{B}}_n'\mathbf{x}$ corresponding to $d = 3$. Bottom row: plots of (left) $\text{dev}(d, 0.05)$ and (middle) $\log m(d)$ – both point to $d_0 = 3$. Bottom right : Predicted values (medians) and spreads (50% two-tail intervals) versus the recorded values from the test set.

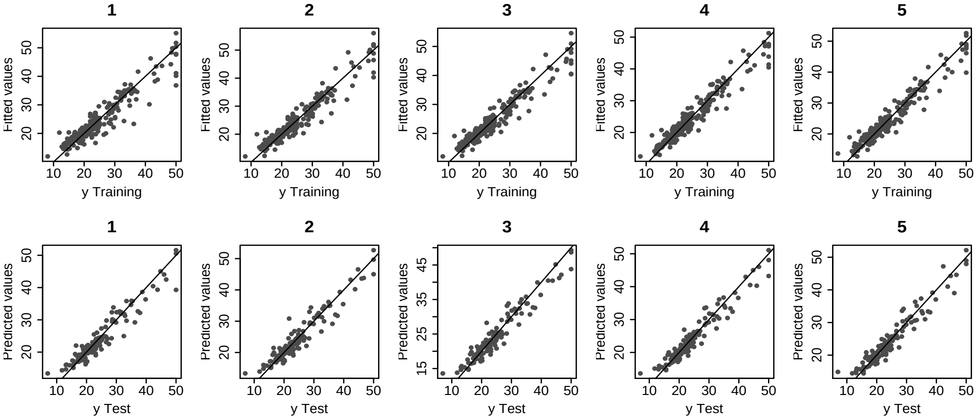

Figure 5: Boston house price: Fitted (top) and predicted (bottom) values for $d = 1, \cdots, 5$, against observed median house price.

particular Neal 1996) – a variable selection criterion based on a similar idea. Relative importance of the variables then can be judged by comparing the posterior distribution of imp_i, $i = 1, \cdots, p$. Variables for which imp_i remains consistently high across the MCMC sample can be thought of as the most important ones. Figure 6 shows boxplots of the MCMC draws of imp_i across the thirteen predictor variables for Boston house price data. Based on this plot the most influential variables appears to be `rm` and `lstat`, with a number of other variables competing with each other at a moderate level. The two most interesting linear projections noted by Chen and Li (1998) were `rm` and `crim + 30 lstat` – very close to our finding. It is, however, important to remember that in presence of linear or non-linear dependence among predictors, any calibration of their relative importance must be interpreted with caution.

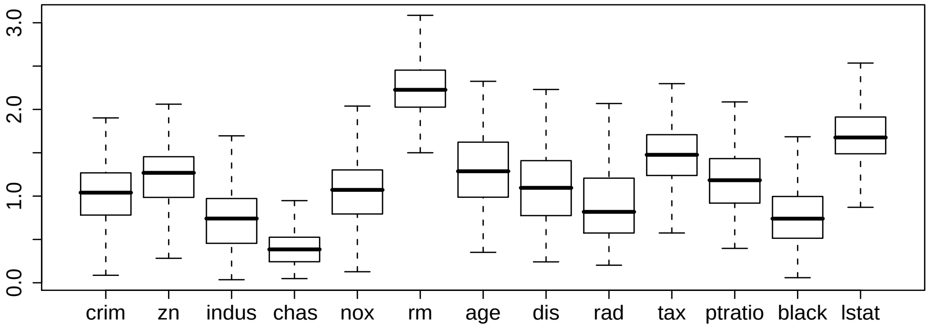

Figure 6: Boston house price: Relative importance of predictors. The boxplots present the posterior draw of imp_i for the thirteen predictors. The average room per unit (`rm`) and the percentage of poor (`lstat`) appeared the most influential predictors, but a number of others had a mild influence as well.

We conclude this example with a comparison of spLGP with the probit stick breaking process (PSBP) based density regression approach of Chung and Dunson (2009). PSBP, among a rapidly growing class of density regression models based on stick breaking processes, explicitly encodes variable selection through a binary inclusion parameter $\boldsymbol{\gamma}$ with $\gamma_j = 1$ indicating x_j is included in the model, and it is excluded otherwise. We applied PSBP to the training subset of our Boston house price data. The posterior inclusion probabilities of the thirteen variables are listed in Table 3. It is striking that these posterior inclusion probabilities took quite extreme values, most variables were zeroed out while a few were almost certainly in the model. It is interesting that `lstat` received a very low posterior probability of being in the model, although it was found to be one of the influential predictors by spLGP. It is likely that this difference arose because the variables in this example are highly dependent on each other, and in its quest of selecting only a few variables PSBP ignored many that are related to one that was already included. The PSBP out of sample prediction error turned out to be 2.35,

	Inclusion		Inclusion
Variable	probability	Variable	probability
`crim`	.03	`dis`	.17
`zn`	.04	`rad`	.04
`indus`	.04	`tax`	.06
`chas`	.03	`ptratio`	1
`nox`	.04	`black`	.33
`rm`	1	`lstat`	.07
`age`	1		

Table 3: Posterior inclusion probabilities of predictor variables with a PSBP density regression of Boston house price data.

about 50% more than that of spLGP. This relatively poor performance can be linked to PSBP's exclusion of many variables from the model, each of which probably had some influence on the response. In fact, when we ran spLGP ($d = 3$) with only those variables that received more than 10% posterior inclusion probability from PSBP (namely, `rm`, `age`, `dis`, `ptratio` and `black`), the out of sample prediction error jumped up to 1.90.

6 Discussion

The examples described in the previous section illustrate that spLGP offers good out of sample prediction, accurate estimation of the central subspace and a reasonably satisfactory estimation of dimensionality that relates well with predictive performance of the competing models. In addition we have proposed a novel technique to judge relative importance of variables within the subspace projection setting. We have also demonstrated that subspace projection may offer better prediction than sparse variable selection when many predictor variables have mild influence on the response and are possibly dependent on each other.

We note that many aspects of our implementation of spLGP may prove to be too simpleminded despite its good performance in the examples we reported above. Our parametrization of $\mathcal{S}$ through a basis $\mathbf{B}$, done in a quest to obtain an Euclidean parameter suitable for random walk type Metropolis exploration, introduces identifiability issues that may lead to poor mixing of a MCMC sampler. In our examples, however, samplers started from different parameter configurations generated from the prior distribution appeared to mix well in the $\mathcal{S}$ space. But in general, a direct parametrization based on $\mathbf{P}_{\mathcal{S}}$ seems more desirable. In addition to removing identifiability issues, such an approach may also make it feasible to explore the posterior over $(d, \mathcal{S}, p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$ via reversible jump MCMC. However, further research is needed to identify efficient proposals for $\mathbf{P}_{\mathcal{S}}$, keeping in mind its non-Euclinear structure.

Estimation of dimensionality is another area that requires further research. The formal approach of computing $\log m(d)$ is computationally prohibitive. The heuristic selection based on $\text{dev}(d, \alpha)$ needs further theoretical backing. The basic intuition behind the heuristic approach is that in a nested model setting, a model that is larger than ideal will lead to a more diffuse posterior than an ideal model. However, there seems to be room to improve the formalization discussed in Section 4 toward a more rigorous one with a stronger Bayesian interpretation.

Nonlinear confounding among the covariates can make the spLGP posterior spread over distant subspaces, making it challenging to explore via Markov chain samplers. However, we note that our naive block Metropolis-Hastings sampler did a reasonably good job of exploring such a posterior distribution in the Boston house price example. The MCMC samples of imp_i showed substantial negative correlation for many pairs of variables, most notably the two influential predictors `rm` and `lstat`. This is an indication that subspace projections that align well with either of these two predictors received considerable posterior mass, and that our sampler was able to move between such subspaces.

The Boston house price example highlights a possible advantage of subspace projection over variable selection when predictor variables are well chosen and are likely to contribute toward predicting the response, even if individual influences are small. On the other hand, this advantage is likely to be lost in studies where a huge number of predictor variables are assembled, only a few of which are likely to influence the response. Moreover, in cases where variable selection is a primary objective, subspace projection may appear inadequate. However, the importance calibration described in the earlier section provides a partial answer. In summary, both these approaches to dimension reduction have their respective merits in terms of performance as well as embedding specific modeling objectives.

Acknowledgement

We would like to thank the Associate Editor and the referee who reviewed an earlier manuscript of this article and gave many constructive suggestions. We would also like to thank Professor David Dunson and Dr. Yeonseung Chung for sharing their PSBP code with us.

Appendix: Technical Details

Lemma 6.1. *Let $\mathcal{B}$ be a uniform random variate in $\mathcal{B}_d$. Define a probability measure μ on $\mathcal{G}_{p,d}$ (equipped with the Borel σ-field generated by a suitable metric ρ) as $\mu(A) = \Pr(\mathcal{C}(\mathbf{B}) \in A)$ for $A \subset \mathcal{G}_{p,d}$. Then μ defines a uniform measure over $\mathcal{G}_{p,d}$ in the sense that it is invariant under rotation: $\mu(\mathbf{R}A) = \mu(A)$ for any $p \times p$ unitary matrix $\mathbf{R}$ and any $A \subset \mathcal{G}_{p,d}$, where $\mathbf{R}A = \{\mathbf{R}\mathcal{S} : \mathcal{S} \in A\}$ with $\mathbf{R}\mathcal{S} = \{\mathbf{R}\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$.*

Proof. Let $\mathcal{S}_{\text{can}}$ denote the linear subspace $\{\mathbf{x} = (x_1, \cdots, x_p) \in \mathbb{R}^p : x_{d+1} = \cdots = x_p = 0\}$. Let $\mathbf{V}$ be a $p \times p$ random matrix whose elements are independent standard normal variates. Then μ can be expressed as $\mu(A) = \Pr(\mathbf{V}\mathcal{S}_{\text{can}} \in A)$, $A \subset \mathcal{G}_{p,d}$. Now, for any unitary $\mathbf{R}$, the random matrix $\mathbf{U} = \mathbf{R}'\mathbf{V}$ has the same distribution as $\mathbf{V}$, and hence for any $A \subset \mathcal{G}_{p,d}$, $\mu(\mathbf{R}A) = \Pr(\mathbf{R}'\mathbf{V}\mathcal{S}_{\text{can}} \in A) = \Pr(\mathbf{U}\mathcal{S}_{\text{can}} \in A) = \mu(A)$.

□

Proof of Theorem 3.1. Recall the representation $f(\mathbf{z}, y) = g_0(y) f_W(\mathbf{F}(\mathbf{z}), G_0(y))$ where W is the underlying Gaussian process and for any continuous function $(w(\mathbf{s}, t)$ on $[-1,1]^d \times [0,1]$, $f_w(\mathbf{s}, t) = e^{w(\mathbf{s},t)} / \int_0^1 e^{w(\mathbf{s},u)} du$. Also recall that $h(\mathbf{x}, y) = p(\mathbf{x}) f(\mathbf{B}'\mathbf{x}, y)$ is the joint density on $(\mathbf{x}, y)$ induced by $(\mathbf{B}, f)$, and similarly $h_0(\mathbf{x}, y) = p(\mathbf{x}) f_0(\mathbf{B}_0'\mathbf{x}, y)$ denotes the true joint density. Here we discuss how to verify the Kullback-Leilber condition

$$\forall \epsilon > 0, \Pr(\text{KL}(h_0, h) < \epsilon) > 0. \tag{10}$$

Verification of the entropy conditions of Ghosal *et al.* (1999) are easily derived by using infinite differentiability of the square-exponential covariance kernel; see Section 5 of Tokdar and Ghosh (2007) for an accessible construction.

Note that

$$\begin{aligned}\text{KL}(h_0, h) &= \int p(x) \text{KL}(f_0(\mathbf{B}_0'\mathbf{x}, \cdot), f_0(\mathbf{B}'\mathbf{x}, \cdot)) d\mathbf{x} \\ &\quad + \int p(x) \int f_0(\mathbf{B}_0'\mathbf{x}) \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{f(\mathbf{B}'\mathbf{x}, y)} dy d\mathbf{x} \\ &\leq a\|\mathbf{B}_0 - \mathbf{B}\|^2 \int \|\mathbf{x}\|^2 p(\mathbf{x}) d\mathbf{x} \\ &\quad + \int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{f(\mathbf{B}'\mathbf{x}, y)} dy d\mathbf{x}.\end{aligned}$$

Fix an $\epsilon > 0$. Let $\mathcal{B}_\epsilon$ be the set of $\mathbf{B} \in \mathcal{B}_d$ for which the first term in the above expression is smaller than $\epsilon/2$. Since $\Pr(\mathcal{B}_\epsilon) > 0$, it suffices to show that for every $\mathbf{B} \in \mathcal{B}_\epsilon$

$$\Pr\left(\int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{f(\mathbf{B}'\mathbf{x}, y)} dy d\mathbf{x} < \epsilon/2 \mid \mathbf{B} \right) > 0. \tag{11}$$

Fix a $\mathbf{B} \in \mathcal{B}_\epsilon$. From Assumption (d), there exists an $r_x > 0$ such that

$$\int_{\|\mathbf{x}\| > r_x} p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \left| \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{g_0(y)} \right| dy d\mathbf{x} < \frac{\epsilon}{16}.$$

Define

$$f_1(\mathbf{s}, t) = \frac{f_0(\mathbf{F}^{-1}(\mathbf{s}), G_0^{-1}(t))}{g_0(G_0^{-1}(t))}$$

on $\mathbf{s} \in (-1,1)^d$ and $t \in (0,1)$. This function is well defined and satisfies

$$f_0(\mathbf{z}, y) = g_0(y) f_1(\mathbf{F}(\mathbf{z}), G_0(y)), \ \mathbf{z} \in \mathbb{R}^d, y \in \mathbb{R}.$$

Define $r_z = \sup\{\|\mathbf{B}'\mathbf{x}\| : \|\mathbf{x}\| \leq r_x\}$ and let $\mathcal{S} \subset [-1,1]^d$ denote the compact set $\{\mathbf{F}(\mathbf{z}) : \|\mathbf{z}\| \leq r_z\}$. From assumption (e), $f_1(\mathbf{s}, t) \to 0$ as $t \to 0$ or $t \to 1$, uniformly on

$\mathcal{S}$. Define f_2 on $[-1,1]^d \times [0,1]$ as

$$f_2(\mathbf{s},t) = \begin{cases} f_1(\mathbf{s},t) & \mathbf{s} \in \mathcal{S}, t \in (0,1) \\ 0 & \mathbf{s} \in \mathcal{S}, t = 0, 1 \\ 1 & \text{otherwise} \end{cases} \tag{12}$$

Therefore for $\|\mathbf{z}\| \le r_z$, $f_0(\mathbf{z},y) = g_0(y) f_2(\mathbf{F}(\mathbf{z}), G_0(y))$ and for $\|\mathbf{z}\| > r_z$, $f_0(\mathbf{z},y) = f_0(\mathbf{z},y) f_2(\mathbf{F}(\mathbf{z}), G_0(y))$. Thus, splitting $\mathbf{B}'\mathbf{x}$ at norm r_z we get

$$\begin{aligned} &\int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{f(\mathbf{B}'\mathbf{x}, y)} dy d\mathbf{x} \\ &\qquad\qquad \le \int p(x) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_2(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))}{f_W(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))} dy d\mathbf{x} \\ &\qquad\qquad + \int_{\|\mathbf{x}\| > r_x} p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \left| \log \frac{f_0(\mathbf{B}'\mathbf{x}, y)}{g_0(y)} \right| dy d\mathbf{x}. \end{aligned}$$

The second term in the last expression above is bounded by $\epsilon/16$. Therefore it suffices to show that the first term can be bounded by $\epsilon/16$ with positive conditional probability given $\mathbf{B}$. Find a $\delta > 0$ small enough such that $\log(1+\delta) < \epsilon/16$. Define

$$f_3(\mathbf{s},t) = \begin{cases} \frac{f_2(\mathbf{s},t)+\delta}{1+\delta} & \mathbf{s} \in \mathcal{S}, t \in [0,1] \\ 1 & \text{otherwise} \end{cases} . \tag{13}$$

Clearly,

$$\begin{aligned} &\int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_2(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))}{f_W(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))} dy d\mathbf{x} \\ &\quad \le \log(1+\delta) + \int p(\mathbf{x}) \int f_0(\mathbf{B}_0'\mathbf{x}, y) \log \frac{f_3(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))}{f_W(\mathbf{F}(\mathbf{B}'\mathbf{x}), G_0(y))} dy d\mathbf{x} \\ &\quad \le \frac{\epsilon}{16} + \int p_B(\mathbf{s}) \sup_t \left| \log \frac{f_3(\mathbf{s},t)}{f_W(\mathbf{s},t)} \right| d\mathbf{s} \end{aligned}$$

where $p_B(\mathbf{s})$ is the density of $\mathbf{F}(\mathbf{B}'\mathbf{x})$ under $\mathbf{x} \sim p(\mathbf{x})$. Let $A = \sup_{\mathbf{s},t} |\log f_3(\mathbf{s},t)|$. Find an open set $\mathcal{S}' \subset [-1,1]^d$ containing $\mathcal{S}$ such that $\int_{\mathcal{S}' \setminus \mathcal{S}} p_B(\mathbf{s}) d\mathbf{s} < \epsilon/(16A)$. Find a continuous function $\lambda : [-1,1]^d \to [0,1]$ with $\lambda(\mathbf{s}) = 1$, $\mathbf{s} \in \mathcal{S}$ and $\lambda(v) = 0$, $\mathbf{s} \in (\mathcal{S}')^c$. Extend f_3 to the following continuous function:

$$f_4(\mathbf{s},t) = \lambda(\mathbf{s}) f_3(\mathbf{P}\mathbf{s},t) + (1 - \lambda(\mathbf{s}))$$

where $\mathbf{P}\mathbf{s}$ is the projection of $\mathbf{s}$ onto $\mathcal{S}$. Since f_4 is everywhere positive on $[-1,1]^d \times [0,1]$, we can define

$$w_4(\mathbf{s},t) = \log f_4(\mathbf{s},t), \mathbf{s} \in [-1,1]^d, t \in [0,1].$$

As w_4 is continuous, we have $\Pr(\|W - w_4\|_\infty < \epsilon/16) > 0$. Therefore,

$$\begin{aligned}\int p_B(\mathbf{s}) \sup_t \left|\log \frac{f_3(\mathbf{s},t)}{f_W(\mathbf{s},t)}\right| d\mathbf{s} &\leq \int p_B(\mathbf{s}) \sup_t \left|\log \frac{f_3(\mathbf{s},t)}{f_4(\mathbf{s},t)}\right| d\mathbf{s} \\ &\quad + \int p_B(\mathbf{s}) \sup_t \left|\log \frac{f_4(\mathbf{s},t)}{f_W(\mathbf{s},t)}\right| d\mathbf{s} \\ &\leq 2A \int_{\mathcal{S}'\setminus\mathcal{S}} p_B(\mathbf{s}) d\mathbf{s} + 2\|W - w_4\|_\infty\end{aligned}$$

From this the result follows.

□

Proof of Theorem 3.2. Let $L(\mathbf{B}, f)$ denote $\int \|f(\mathbf{B}'\mathbf{x}, \cdot) - f_0(\mathbf{B}_0'\mathbf{x}, \cdot\|_1 p(\mathbf{x}) d\mathbf{x}$. Theorem 3.1 leads to a direct proof of $\Pi^{(n)}(R_{\rho,\delta}) \to 1$ if we can show

$$\forall \delta > 0, \exists \epsilon > 0 \text{ such that } \rho(\mathbf{P_B}, \mathbf{P}_0) > \delta \implies \inf_{f \in \mathcal{F}_d} L(\mathbf{B}, f) > \epsilon \tag{14}$$

Suppose (14) is false. Then, there exists a $\delta > 0$ and sequences $\mathbf{B}_k \in \mathcal{B}_d$ and $f_k \in \mathcal{F}_d$ such that $\rho(\mathbf{P}_{\mathbf{B}_k}, \mathbf{P}_0) > \delta$ but $L(\mathbf{B}_k, f_k) \to 0$. Without loss of generality, assume $\mathbf{B}_k$ to have orthogonal columns. Then $\mathbf{B}_k$ can be extended to a $p \times p$ orthonormal matrix $\mathbf{M}_k = [\mathbf{B}_k : \mathbf{C}_k]$. Lemma 6.2 asserts that there exist bounded open intervals $\mathcal{J} \subset \mathbb{R}^d$, $\mathcal{I}_1, \mathcal{I}_2 \subset \mathbb{R}^{p-d}$ and constants $\epsilon > 0$, $a < b$ such that for infinitely many k

$$f_0(\mathbf{B}_0'\mathbf{B}_k\mathbf{u} + \mathbf{B}_0'\mathbf{C}_k\mathbf{v}_1, y) > f_0(\mathbf{B}_0'\mathbf{B}_k\mathbf{u} + \mathbf{B}_0'\mathbf{C}_k\mathbf{v}_2, y) + \epsilon$$

for every $\mathbf{u} \in \mathcal{J}$, $\mathbf{v}_1 \in \mathcal{I}_1$, $\mathbf{v}_2 \in \mathcal{I}_2$ and $y \in (a, b)$. For any such k, by applying the change of variable $(\mathbf{u}, \mathbf{v}) = (\mathbf{B}_k'\mathbf{x}, \mathbf{C}_k'\mathbf{x})$, we get

$$\begin{aligned}L(\mathbf{B}_k, f_k) &= \int |f_k(\mathbf{u}, y) - f_0(\mathbf{B}_0'\mathbf{B}_k\mathbf{u} + \mathbf{B}_0'\mathbf{C}_k\mathbf{v}, y)|\, p(\mathbf{M}_k(\mathbf{u}', \mathbf{v}')') dy d\mathbf{v} d\mathbf{u} \\ &\geq \frac{(b-a)\epsilon}{2} \int_{\mathbf{u} \in \mathcal{J}} \left\{ \min_{j=1,2} \int_{\mathbf{v} \in \mathcal{I}_j} p(\mathbf{M}_k(\mathbf{u}', \mathbf{v}')') d\mathbf{v} \right\} d\mathbf{u}\end{aligned}$$

which is bounded away from 0 by Lemma 6.2 – a contradiction to $L(\mathbf{B}_k, f_k) \to 0$. So (14) must be true. The second assertion is easy to prove once we note compactness of $\mathcal{B}_d$, continuity of the map $\mathbf{B} \mapsto \mathbf{P_B}$ and boundedness of ρ_{trace}; see also Proposition 4.2.1 of Ghosh and Ramamoorthi (2003).

□

Lemma 6.2. *Let there be a unique d-dimensional $\mathcal{S}_0$ such that $p(y \mid \mathbf{x}) = p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x})$ and let $\mathbf{B}_0$ be a basis of $\mathcal{S}_0$. Fix an $f_0 \in \mathcal{F}_d$ such that $p(y \mid \mathbf{P}_{\mathcal{S}_0}\mathbf{x}) = f_0(\mathbf{B}_0\mathbf{x}, y)$. Assume f_0 to be continuous. Let $\mathbf{M}_k = (\mathbf{B}_k : \mathbf{C}_k)$ be a sequence of $p \times p$ orthonormal matrices where $\mathbf{B}_k \in \mathcal{B}_d$ satisfies $\rho(\mathbf{B}_k, \mathbf{B}_0) > \delta$ for all $k \geq 1$. Then there exist bounded open intervals $\mathcal{J} \subset \mathbb{R}^d$, $\mathcal{I}_1, \mathcal{I}_2 \subset \mathbb{R}^{p-d}$ and constants $\epsilon > 0$, $a < b$ such that,*

1. *for infinitely many* k

$$f_0(\mathbf{B}_0'\mathbf{B}_k\mathbf{u} + \mathbf{B}_0'\mathbf{C}_k\mathbf{v}_1, y) > f_0(\mathbf{B}_0'\mathbf{B}_k\mathbf{u} + \mathbf{B}_0'\mathbf{C}_k\mathbf{v}_2, y) + \epsilon \tag{15}$$

   for every $\mathbf{u} \in \mathcal{J}$, $\mathbf{v}_1 \in \mathcal{I}_1$, $\mathbf{v}_2 \in \mathcal{I}_2$ *and* $y \in (a, b)$

2. *and*

$$\liminf_k \int_{\mathbf{u}\in\mathcal{J}} \left\{ \min_{j=1,2} \int_{\mathbf{v}\in\mathcal{I}_j} p(\mathbf{M}_k(\mathbf{u}', \mathbf{v}')')d\mathbf{v} \right\} d\mathbf{u} > 0. \tag{16}$$

Proof. Since $\mathcal{S}_0$ is unique, it cannot have a lower dimensional subspace $\mathcal{S}$ satisfying $p(y \mid \mathbf{x}) = p(y \mid \mathbf{P}_{\mathcal{S}}\mathbf{x})$. Therefore, for every non-trivial subspace $\mathcal{S}$ of $\mathbb{R}^d$ there must exist $\mathbf{z}_1^*, \mathbf{z}_2^* \in \mathbb{R}^d$ with $\mathbf{z}_1^* - \mathbf{z}_2^* \in \mathcal{S}$ such that $f_0(\mathbf{z}_1^*, \cdot) \neq f_0(\mathbf{z}_2^*, \cdot)$. Continuity of f_0 then implies existence of $\epsilon > 0$, $\eta > 0$ and $a < b$ such that

$$f_0(\mathbf{z}_1, y) > f_0(\mathbf{z}_2, y) \text{ for all } \|z_1 - \mathbf{z}_1^*\| < \eta, \|\mathbf{z}_2 - \mathbf{z}_2^*\| < \eta, a < y < b. \tag{17}$$

By compactness, $\mathbf{M}_k = (\mathbf{B}_k : \mathbf{C}_k)$ converges to an orthonormal matrix $\mathbf{M} = (\mathbf{B} : \mathbf{C})$ along a subsequence, which we again index by k. Take $\mathbf{P}_k = \mathbf{P}_{\mathbf{B}_k}$ and $\mathbf{P} = \mathbf{P}_{\mathbf{B}}$. Then $\mathcal{S} = \{\mathbf{z} = \mathbf{B}_0'(\mathbf{I} - \mathbf{P})\mathbf{x} : \mathbf{x} \in \mathbb{R}^p\}$ is a subspace of $\mathbb{R}^d$ and $\mathcal{S} \neq \{0\}$, because, otherwise $0 = \rho(\mathbf{P}, \mathbf{P}_0) = \lim_n \rho(\mathbf{P}_n, \mathbf{P}_0) \geq \delta$ - a contradiction! Find $\mathbf{z}_1^*, \mathbf{z}_2^* \in \mathcal{S}$, $\epsilon > 0$ and $\eta > 0$ satisfying (17). Then $\Delta z^* = z_1^* - z_2^* = \mathbf{B}_0'\Delta x^*$ for some Δx^* satisfying $\mathbf{B}'\Delta x^* = 0$. Set $x_2^* = \mathbf{B}_0(\mathbf{B}_0'\mathbf{B}_0)^{-1}\mathbf{z}_2^*$ and $\mathbf{x}_1^* = \mathbf{x}_2^* + \Delta x^*$.

Define $\mathbf{u}^* = \mathbf{B}'\mathbf{x}_1^* (= \mathbf{B}'\mathbf{x}_2^*)$ and $\mathbf{v}_i^* = \mathbf{C}'\mathbf{x}_i^*$, $i = 1, 2$. Then $\mathbf{B}_0'(\mathbf{B}\mathbf{u}^* + \mathbf{C}\mathbf{v}_i^*) = \mathbf{z}_i^*$. Therefore there exist open neighborhoods $\mathcal{J} \subset \mathbb{R}^d$ of $\mathbf{u}^*$, $\mathcal{I}_i \subset \mathbb{R}^{p-d}$ of $\mathbf{v}_i^*$, $i = 1, 2$, such that for $\mathbf{u} \in \mathcal{J}$ and $\mathbf{v}_i \in \mathcal{I}_i$, $i = 1, 2$, $\|\mathbf{B}_0'(\mathbf{B}\mathbf{u}^* + \mathbf{C}\mathbf{v}_i^*) - \mathbf{z}_i^*\| < \eta/2$. This implies (15) since

$$\|\mathbf{B}_0'\mathbf{B}\mathbf{u} + \mathbf{B}_0'\mathbf{C}\mathbf{v} - \mathbf{B}_0\mathbf{B}_k\mathbf{u} - \mathbf{B}_0'\mathbf{C}_k\mathbf{v}\| \leq (\|\mathbf{u}\| + \|\mathbf{v}\|)\|\mathbf{B}_0\|\|\mathbf{M}_k - \mathbf{M}\|$$

which can be made arbitrarily small for all large k uniformly over $\mathbf{u} \in \mathcal{J}$ and $\mathbf{v} \in \mathcal{I}_i$, $i = 1, 2$.

Since q is bounded, an application of the dominated convergence theorem implies,

$$\liminf_k \int_{\mathbf{u}\in\mathcal{J}} \left[\min_{j=1,2} \int_{\mathbf{v}\in\mathcal{I}_j} p(\mathbf{M}_k(\mathbf{u}', \mathbf{v}')')d\mathbf{v} \right] d\mathbf{u} = \int_{\mathbf{u}\in\mathcal{J}} \left[\min_{j=1,2} \int_{\mathbf{v}\in\mathcal{I}_j} p(\mathbf{M}(\mathbf{u}', \mathbf{v}')')d\mathbf{v} \right] d\mathbf{u}.$$

The last quantity is strictly positive since $\mathcal{J}$, $\mathcal{I}_1$ and $\mathcal{I}_2$ are open intervals and $p(\mathbf{M}(\mathbf{u}', \mathbf{v}')') > 0$ for all $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^d \times \mathbb{R}^{p-d}$. This proves (16).

□

References

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009). "The Gaussian process density sampler." In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, 9–16. 326

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society - Series B*, 70: 825–848. 326

Chen, C. H. and Li, K. C. (1998). "Can SIR be as Popular as Multiple Linear Regression?" *Statistica Sinica*, 8: 289–316. 333, 334, 336

Chib, S. and Jeliazkov, I. (2001). "Marginal Likelihood from the Metropolis-Hastings Output." *Journal of the American Statistical Association*, 96: 270–281. 328

Chung, Y. and Dunson, D. B. (2009). "Nonparametric Bayes conditional distribution modeling with variable selection." *Journal of the American Statistical Association*, 104(488): 1646–1660. 319, 320, 336

Cook, R. D. (1994). "Using dimension-reduction subspaces to identify important inputs in models of physical systems." In *Proceedings of the Section on Physical Engineering Sciences*, 18–25. Washington: American Statistical Association. 324

Cook, R. D. and Weisberg, S. (1991). "Discussion of Sliced Inverse Regression for Dimension Reduction by K. C. Li." *Journal of the American Statistical Association*, 86: 328–332. 321

Dunson, D. B. and Park, J. H. (2008). "Kernel Stick-breaking Processes." *Biometrika*, 95: 307–323. 319

Dunson, D. B., Pillai, N., and Park, J. H. (2007). "Bayesian Density Regression." *Journal of the Royal Statistical Society - Series B*, 69: 163–183. 319

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, 2 edition. 323

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27: 143–158. 324

Ghosal, S. and Roy, A. (2006). "Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression." *The Annals of Statistics*, 34: 2413–2429. 323

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag. 341

Griffin, J. E. and Steel, M. F. J. (2006). "Order-Based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 101: 179–194. 319

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management*, 5: 81–102. 320, 333

Lenk, P. J. (1988). "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities." *Journal of the American Statistical Association*, 83: 509–516. 320

— (1991). "Towards a Practicable Bayesian Nonparametric Density Estimator." *Biometrika*, 78: 531–543. 320

— (2003). "Bayesian Semiparametric Density Estimation and Model Verification Using a Logistic Gaussian Process." *Journal of Computational and Graphical Statistics*, 12: 548–565. 320

Li, K. C. (1991). "Sliced Inverse Regression for Dimension Reduction (with discussions)." *Journal of the American Statistical Association*, 86: 316–342. 321, 333

MacEachern, S. M. (1999). "Dependent Nonparametric Processes." In *Proceedings of the Section on Bayesian Statistical Science*, 50–55. Alexandria, VA: American Statistical Association. 320

— (2000). "Dependent Dirichlet Processes." Ohio State University, Dept. of Statistics Technical Report. 320

Mattila, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces*. New York: Cambridge University Press. 322

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer Verlag. 328, 336

Tokdar, S. T. (2007). "Towards a Faster Implementation of Density Estimation with Logistic Gaussian Process Priors." *Journal of Computational and Graphical Statistics*, 16: 633–655. 320, 322, 326

Tokdar, S. T. and Ghosh, J. K. (2007). "Posterior consistency of logistic Gaussian process priors in density estimation." *Journal of Statistical Planning and Inference*, 137: 34–42. 320, 339

van der Vaart, A. W. and van Zanten, H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *The Annals of Statistics*, 36: 1435–1463. 320

Xia, Y. (2007). "A constructive approach to the estimation of dimension reduction directions." *The Annals of Statistics*, 35: 2654–2690. 321, 324, 331, 332

Zeng, P. (2008). "Determining the dimension of the central subspace and central mean subspace." *Biometrika*, 95: 469–479. 332

Zhu, Y. and Zeng, P. (2006). "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression." *Journal of the American Statistical Association*, 101: 1638–1651. 321