


Spatial Shrinkage Via the Product Independent Gaussian Process Prior

Arkaprava Roy, Brian J. Reich, Joseph Guinness, Russell T. Shinohara & Ana-Maria Staicu


To cite this article: Arkaprava Roy, Brian J. Reich, Joseph Guinness, Russell T. Shinohara & Ana-Maria Staicu (2021) Spatial Shrinkage Via the Product Independent Gaussian Process Prior, Journal of Computational and Graphical Statistics, 30:4, 1068-1080, DOI: [10.1080/10618600.2021.1923512](https://doi.org/10.1080/10618600.2021.1923512)

To link to this article: <https://doi.org/10.1080/10618600.2021.1923512>

 View supplementary material 

 Published online: 21 Jun 2021.

 Submit your article to this journal 

 Article views: 387

 View related articles 

 View Crossmark data 

 Citing articles: 1 View citing articles 



Spatial Shrinkage Via the Product Independent Gaussian Process Prior

Arkaprava Roy^a , Brian J. Reich^b, Joseph Guinness^c, Russell T. Shinohara^d, and Ana-Maria Staicu^b

^aDepartment of Biostatistics, University of Florida, Gainesville, FL; ^bDepartment of Statistics, North Carolina State University, Raleigh, NC; ^cDepartment of Statistics and Data Science, Cornell University, Ithaca, NY; ^dDepartment of Biostatistics, University of Pennsylvania, Philadelphia, PA

ABSTRACT

We study the problem of sparse signal detection on a spatial domain. We propose a novel approach to model continuous signals that are sparse and piecewise-smooth as the product of independent Gaussian (PING) processes with a smooth covariance kernel. The smoothness of the PING process is ensured by the smoothness of the covariance kernels of the Gaussian components in the product, and sparsity is controlled by the number of components. The bivariate kurtosis of the PING process implies that more components in the product results in the thicker tail and sharper peak at zero. We develop an efficient computation algorithm based on spectral methods. The simulation results demonstrate superior estimation using the PING prior over Gaussian process prior for different image regressions. We apply our method to a longitudinal magnetic resonance imaging dataset to detect the regions that are affected by multiple sclerosis computation in this domain. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received August 2018
Revised April 2021

KEYWORDS

Bayesian; High dimension;
Image regression; Multiple
sclerosis; Shrinkage; Spatial
data analysis

1. Introduction

In this article, we discuss linear regression models for two- or three-dimensional image responses, image covariates, or both, in which the signal is assumed to be continuous, sparse and piecewise smooth. The methodological development is motivated by a study of multiple sclerosis (MS) using magnetic resonance imaging (MRI) (Sweeney et al. 2016; Pomann et al. 2016; Mejia et al. 2016), where subjects with MS are imaged repeatedly over multiple hospital visits, and the objective is to identify the brain regions that are damaged over time. Although a healthy brain would not change much during the study period, a diseased brain is expected to exhibit changes in a small number of regions of interest that are associated with the disease. This is an example, where using an image-on-scalar regression framework, would require the signal to be continuous, sparse and piecewise smooth.

Modeling a continuous, sparse and piecewise-smooth signal for high-dimensional data poses several challenges such as (i) handling the complex spatial dependence of the data, (ii) accounting for a simultaneously sparse and continuous, where sparsity is defined in terms of the number of nonzero smooth pieces that comprise the signal, and (iii) accommodating a possibly large dimensional signal. Approaches to estimate a sparse signal exist both by using a frequentist perspective (such as Tibshirani 1996; Tibshirani et al. 2005) as well as a Bayesian paradigm. We focus on the latter as it allows to also quantify estimation uncertainty. In a Bayesian framework, parameter sparsity is modeled using the traditional spike and slab prior (Mitchell and Beauchamp 1988), the horseshoe prior

(Carvalho, Polson, and Scott 2010), normal-gamma prior (Griffin and Brown 2010), double-Pareto prior (Armagan, Dunson, and Lee 2013), or Dirichlet-Laplace prior (Bhattacharya et al. 2015). However, none of these priors ensures a smooth spatial structure. In the context of high-dimensional data, this adds computational challenges as well. Reducing computational demand is an important task that we address in this article.

Estimating sparse and spatially smooth signals has been attracting increasing interest in image regression. For image-on-scalar regression, Yan and Liu (2017) and Chen et al. (2016) tackled a similar problem. The first article considers a Laplacian-type penalty and the second article introduces a fused SCAD-type penalty to account simultaneously for spatial smoothness and sparsity. In the context of functional magnetic resonance imaging (fMRI) studies, Zhang et al. (2016) and Musgrove, Hughes, and Eberly (2016) considered similar regression models. Their estimation approach uses a spike-and-slab prior to induce sparsity and considers spatial smoothness for the parameter selection. However, the approach does not guarantee that the estimated signal is smooth. In scalar-on-image regression, there is limited work on sparse and piece-wise smooth signal estimation. Goldsmith, Huang, and Crainiceanu (2014) and Li et al. (2015) proposed priors that account separately for spatial dependence and sparsity. For the same problem, Wang and Zhu (2017) proposed a penalty based on the total variation. Spatial dependence is still not fully incorporated in this approach. In Kang, Reich, and Staicu (2016), the proposed soft-thresholded Gaussian process (GP) prior account for both spatial dependence and sparsity simultaneously. The method is computationally very expensive.

In the context of image response and image predictors, there is a more limited research (Morris et al. 2011; Jog, Carass, and Prince 2013; Sweeney et al. 2013). Noh and Park (2010), and Tang, Wang, and Zhu (2013) considered a varying co-efficient model that accounts for sparsity but not smoothness. To the best of our knowledge, only Boehm-Vock et al. (2015) and Jhuang et al. (2018) considered both smoothness and sparsity for image-on-image regression. Their methodology captures the spatial dependence using copulas, which is computationally expensive for large datasets. Our proposed prior has nice conjugacy structure and leads to computational advantages.

In this article, we propose a novel prior for modeling continuous, sparse and piecewise smooth signals. We construct the prior as the location-wise product of independent GPs with smooth covariance kernel. The proposed prior has both high mass around zero, which creates sparsity in the estimation, and a smooth covariance kernel that ensures large support for the spatially varying function. To handle the heavy computational burden associated with this kind of prior, we propose to use the discrete Fourier transformation (DFT) that decorrelates the stationary part of the response. Specifically, we use the fast Fourier transformation (FFT). The FFT algorithm requires regularly spaced input data. In reality, the datasets are not often on a regular grid. To bypass this issue, we propose a fast imputation technique to transform the data into a regular grid. If the dimension of the dataset is manageable for computation in the spatial domain, one can exploit the conjugacy structure of our prior to get the full conditional distribution of parameters given the error process is Gaussian. We analyze the performance of our prior with respect to commonly used GP prior in different linear image regressions with signals that are sparse, piecewise smooth and continuous.

The remainder of the article is organized as follows. In the next section, we describe the image-on-scalar regression model along with the new sparse prior process. We discuss the usage of our new prior to other image regression models in Section 3. In Section 4, we describe other computational aspects that we use for faster computation. In Section 5, we evaluate the finite sample performance of this new prior in the case of different image regression models. We apply our method to a longitudinal MRI data in Section 6 and end with some concluding remarks in Section 7.

2. The Modeling Framework

Our research is motivated by a longitudinal study of MS via MRI images. We introduce the main ideas for the case when we have images collected at multiple time points for a single subject. Specifically, let $Y_i(\nu)$ be the intensity of the i th MRI image collected at time t_i and for a three-dimensional voxel ν for an MS subject. Consider the following linear image-on-scalar regression model:

$$Y_i(\nu) = \alpha(\nu) + t_i\beta(\nu) + E_i(\nu), \quad (1)$$

where $\alpha(\nu)$ is a spatially varying intercept function and $\beta(\nu)$ is an unknown continuous function that quantifies the effect of time. Additionally, it is assumed that $\beta(\cdot)$ is piecewise smooth and sparse. The error $E_i(\nu)$ is a spatially correlated mean-zero

error process, independent across visits. We hypothesize that $\beta(\nu) = 0$ for most of the voxels except for voxels forming spatially contiguous subsets where $\beta(\nu) \neq 0$. Thus, an alternative formulation of our proposed model could be writing it in terms of a frequentist hypothesis testing problem. However, in our Bayesian framework, we may specify a sparse prior for $\beta(\nu)$ in the above model (1) itself and assess the frequentist properties of our decision rules for identifying the voxels with $\beta(\nu) \neq 0$ in terms of true positive (TP) and false positive (FP) proportions. In this article, our proposed sparse prior for $\beta(\nu)$ is a product of independent GPs. We formally describe its properties and the error process in the remainder of this section.

2.1. PING Process

Let $\beta_1(\nu), \dots, \beta_q(\nu)$ be q independent and identically distributed GPs with mean $E\{\beta_j(\nu)\} = 0$, variance $V\{\beta_j(\nu)\} = 1$, and covariance kernel $\text{cov}\{\beta_j(\nu), \beta_j(\nu')\} = K(\nu, \nu')$ for $j = 1, \dots, q$. The zero-mean product of independent Gaussian (PING) stochastic process is defined as the point-wise product of independent GPs, $\beta(\nu) = \sigma\beta_1(\nu) \cdot \beta_2(\nu) \cdot \dots \cdot \beta_q(\nu)$, where $\sigma > 0$ is a scale parameter. The stochastic process $\beta = \{\beta(\nu) : \nu \in \mathcal{V}\}$ constructed in this way is denoted $\beta \sim \text{PING}(q, \sigma^2, K)$.

2.1.1. Properties of the Marginal Distribution

We first discuss the distribution of the PING process at a single location ν . The theoretical properties of the marginal distribution of $\beta(\nu)$ have been studied by Stojanac, Suess, and Kliesch (2017) and Gaunt (2018). Gaunt (2018) provided detail results on characteristic function and propose estimates for the tail behavior of product normals. We briefly revise some of its properties here for completeness. The marginal density function $f_q(x)$ for the product of q standard normals is given by $f_q(x) = \frac{1}{(2\pi)^{q/2}} G_{0,q}^{q,0}(\frac{x^2}{2^q} | 0)$, where $G(\cdot)$ denotes the Meijer G-function (Stojanac, Suess, and Kliesch 2017). The k th marginal moment is $E\{\beta(\nu)^k\} = [(k-1)!!]^q$ where $n!!$ is the product of all numbers from 1 to n that have the same parity as n . The density is unimodal and symmetric about zero; thus, all the odd-order moments are zero. The variance is $V\{\beta(\nu)\} = \sigma^2$. The marginal kurtosis is equal to $3^q - 3$ which is an increasing function of q . As a result, the marginal density has thicker tail and sharper peak at zero for larger q . This is depicted in the first row of Figure 1. Furthermore, $f_q(x) \sim q \exp(-q(x^2/2^q)^{1/q})$ as x goes to infinity and thus the tail is heavier than Gaussian for $q > 1$ and it gets heavier as q increases; (see Gaunt (2018)).

2.1.2. Properties of the Bivariate Distribution

Next, we study the bivariate properties of the PING process at a pair of locations ν_1 and ν_2 . From the construction of the PING process with q components, this bivariate distribution is in fact the distribution of the product of q bivariate normals. Simple calculations show that its mean is $E\{\beta(\nu_j)\} = 0$ for $j = 1, 2$, and its covariance is $\text{cov}\{\beta(\nu_1), \beta(\nu_2)\} = \sigma^2 K^q(\nu_1, \nu_2)$, implying a correlation coefficient that is smaller than the correlation of each individual Gaussian components and that further decays with the number of components, q . In particular, if $K(\cdot)$ is the powered exponential correlation kernel, $K(\nu_1, \nu_2) = \exp\{-\left(\frac{\|\nu_1 - \nu_2\|_2}{\rho}\right)^\nu\}$, the PING covariance

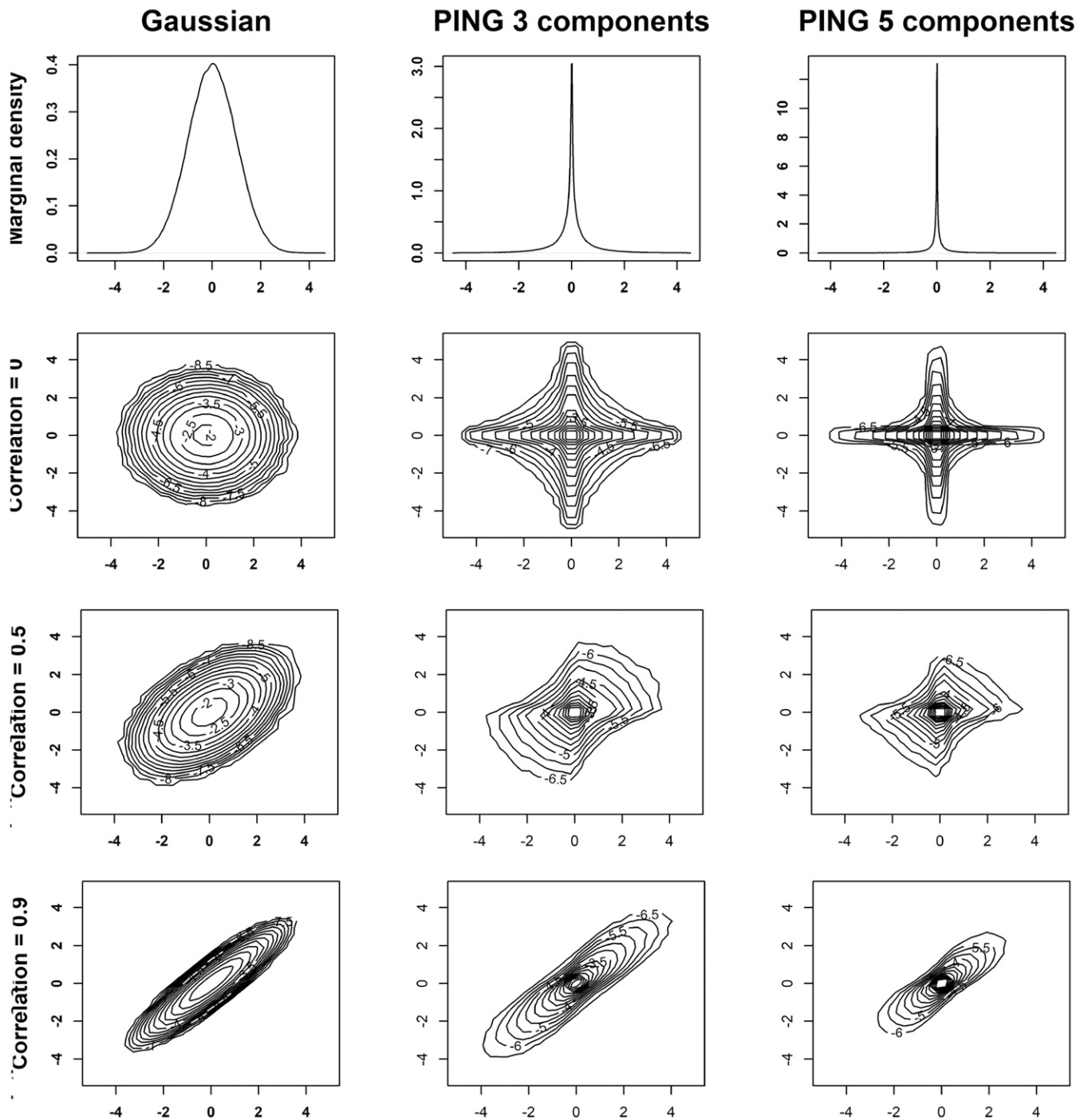


Figure 1. Comparison of Gaussian density with PING-3 and PING-5 densities.

is $\exp\{-\left(\frac{\|v_1 - v_2\|_2}{\rho q^{-1/v}}\right)^v\}$, where the v and ρ are smoothness and range parameters, respectively. Therefore, while the covariance decreases with q for a fixed kernel function, strong spatial correlation can be maintained for large q by simply increasing the parameter ρ with q . The smoothness of the product process is the same as that of its individual components for these power exponential cases. We expect that separating sparsity and spatial dependence hold for other kernel functions as well. To quantify the shrinkage properties, we study the kurtosis of this product distribution. Kurtosis of a multivariate random variable Z of

dimension p with mean μ_Z and covariance matrix Σ_Z is defined as $E[(Z - \mu_Z)^T \Sigma_Z^{-1} (Z - \mu_Z)]^2 - p(p + 2)$ (Mardia 1970). The kurtosis of a general product of bivariate normal random variable is summarized by the following theorem.

Theorem 1. Let Z_1, \dots, Z_q be such that $Z_i \stackrel{\text{ind}}{\sim} \text{BVN}(0, 0, \sigma_{i1}^2, \sigma_{i2}^2, \rho)$ for $i = 1, \dots, q$ and $P_q = Z_1 \odot \dots \odot Z_q$. The mean and the covariance matrix of P_q are $E(P_q) = 0$ and

$$\text{cov}(P_q) = \begin{bmatrix} \prod_{i=1}^q \sigma_{i1}^2 & \rho^q \prod_{i=1}^q \sigma_{i1} \sigma_{i2} \\ \rho^q \prod_{i=1}^q \sigma_{i1} \sigma_{i2} & \prod_{i=1}^q \sigma_{i2}^2 \end{bmatrix}.$$

The kurtosis is $\text{Kurt}(q, m) = \frac{2 \times 3^q}{(1-m^q)^2} [1 + 2(\frac{(1+2m)m}{3})^q + (\frac{1+2m}{3})^q - 4m^q] - 8$, with $m = \rho^2$ and it increases with q .

Here, “BVN” stands for bivariate normal and \odot denotes the element wise product. We allow varying variances for individual components as the kurtosis does not depend on the variances. Supplementary materials for details. The “increasing” property of the kurtosis results from the application of arithmetic and geometric means inequality. The distribution of $\beta(v)$ for two locations v_1 and v_2 under a PING process has the above-mentioned properties with $\sigma_{i1} = \sigma_{i2} = 1$ for $i > 1$. Since it is a unimodal symmetric distribution, higher kurtosis suggests a heavier tail and higher peakedness at zero as q increases; Figure 1 depicts the joint density function of $\beta(v_1)$ and $\beta(v_2)$ for $q = 1, q = 3$ and $q = 5$, and for different correlations. In this plot, we observe that the mass at zero increases with q while they share the same covariance structure with unit variance term. Figure 1 of the supplementary material shows the conditional density of $\beta(v)$ for an arbitrary location v given $\beta(v')$ with $v \neq v'$. The conditional density at one location tends to have a shorter peak as the value at other location moves away from zero. Also, the conditional densities tend to be more positively skewed, as we condition on higher values for the other location.

2.1.3. Multivariate Properties

Let $P_q = Z_1 \odot \dots \odot Z_q$ be random vector of length p where $Z_1 \sim \text{MVN}(0, \Sigma_1)$ and $Z_2, \dots, Z_q \stackrel{\text{ind}}{\sim} \text{MVN}(0, \Sigma_2)$ and Σ_2 has diagonal entries equal to one. Then, we have $E(P_q' P_q) = \text{trace}(\Sigma_1)$ for all q and $\Sigma_2 \odot \dots (q\text{-times}) \odot \Sigma_2 \rightarrow I_p$, as q increases to infinity, where I_p is the $p \times p$ identity matrix. The distribution of $\beta(v)$ for a finite set of locations has the above-mentioned properties as $\beta_2(v), \dots, \beta_q(v)$ have the same covariance kernel with one total variance. In general, it is difficult to explicitly calculate kurtosis in multivariate setup. However, we have following alternative result.

Theorem 2. The multivariate kurtosis of P_q increases with q .

This can be proved using the method of induction and Theorem 1. The proof is in the supplementary materials.

To summarize, for $q = 1$, the PING process is the standard GP and as q increases PING has a mass near zero and its tail probabilities increase. Furthermore, an appropriate rescaling of the spatial correlation parameters, can maintain the smoothness properties of the original GP. Therefore, the PING process is an attractive model for a sparse and smooth signal. However, because of the differentiability property, the posterior mode for PING is never zero, unlike Bayesian LASSO. The number of terms q plays a major role in the application of the PING process, and we recommend to select this tuning parameter using cross-validation. We discuss our cross-validation in detail in Section 5.1.

2.2. Error Distribution and Matern Correlation

Next, we discuss the error process $E_i(v)$. To account for both large- and small-scale spatial deviations of Y_i we consider the

following decomposition, similar to Reich et al. (2018) of $E_i(v)$,

$$E_i(v) = \sum_{j=1}^J Z_j(v) \gamma_{ij} + \epsilon_i(v), \quad (2)$$

where the first term is a linear combination of known basis functions Z_j 's and γ_{ij} are unknown coefficients for i th visit and j th basis and captures The large-scale deviation. The second term ϵ_i is intended to capture small-scale deviations. We assume that ϵ_i is mean-zero GP with stationary and isotropic Matern covariance function as follows:

$$\text{cov}(\epsilon_i(v), \epsilon_i(v')) = C(\theta) = \sigma^2 I(v = v') + \tau^2 \mathcal{M}_\nu \left(\frac{\|v - v'\|}{\phi} \right), \quad (3)$$

where $\mathcal{M}_\nu(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} (3h\sqrt{\nu})^\nu \mathcal{K}_\nu(3h\sqrt{\nu})$ and \mathcal{K} is the modified Bessel function of the second kind. The Matern covariance has four parameters $\theta = (\sigma^2, \tau^2, \phi, \nu)$, that represent the variance of the non-spatial error (nugget), the variance of the spatial process (partial sill), the spatial range and the smoothness of the correlation function respectively.

The large-scale spatial structure is described by J random-effect covariates $\{Z_1, \dots, Z_J\}$. Among many different choices for Z_j 's, we consider outer product of B-spline basis functions. The error in approximating the non-stationary covariance function using B-splines decreases at a rate of $J^{-\alpha/2}$ (Shen and Ghosal 2015), where J is the number of B-spline basis and α is the regularity of the process. Thus, with large J , this can approximate any covariance function. We assume the random effects are normally distributed, that is, $\gamma_i = (\gamma_{i1}, \dots, \gamma_{iJ})^T \sim \text{Normal}(0, \Sigma)$, where Σ is the $J \times J$ covariance matrix. The nonstationary component of the covariance is

$$\text{NS}(v, v') = \sum_{j=1}^J \sum_{l=1}^J Z_j(v) Z_l(v') \Sigma_{jl}. \quad (4)$$

Then, the overall covariance becomes sum of Equations (3) and (4).

3. Extension to Other Image Regression Models

The model in the previous section is designed for image-on-scalar regression. Our sparse prior can be easily adapted for other image regressions, as described next.

3.1. Image-On-image Regression

Consider the case of a linear image-on-image regression model (see, e.g., Gelfand et al. 2003; Morris et al. 2011; Jog et al. 2015, 2017; Hazra et al. 2019)

$$Y_i(v) = \alpha(v) + \sum_{j=1}^p X_{ij}(v) \beta_j(v) + E_i(v), \quad (5)$$

where Y_i is the image response and the X_{ij} 's are the image predictors for subject i . Here, $\alpha(\cdot)$ is an unknown intercept as before and $\beta_j(\cdot)$ are spatially varying piecewise smooth and sparse covariate effects, and $E_i(\cdot)$ is the error process.

We put the PING prior on each of $\beta_j(\cdot)$ for sparse and smooth estimation. The selection of q is done based on a cross validation technique which is discussed in Section 6. This gives local variable selection as the subset of the covariates with beta shrunk toward zero changes with s .

3.2. Scalar-On-Image Regression

Finally, consider the case of a scalar-on-image regression model (see Wang and Zhu 2017; Kang, Reich, and Staicu 2016; Goldsmith, Huang, and Crainiceanu 2014; Li et al. 2015). This model is

$$Y_i = \sum_{j=1}^n X_i(v_j) \beta(v_j) + \epsilon_i, \quad (6)$$

where Y_i is the scalar response and X_i is an image with n spatial locations for subject i . Here, $\beta(\cdot)$ are spatially varying piecewise smooth and sparse covariate effect, and ϵ_i is the error which follows $N(0, \sigma^2)$. We again put a PING prior on $\beta(\cdot)$ for sparse and smooth estimation and its performance is studied Section 5.

4. Computational Details

The prior on intercept (α) as well as the components in the PING prior (β_k 's that comprise the PING prior) are assumed to be mean-zero GP with stationary and isotropic Matern covariance function:

$$\text{cov}(\alpha(v), \alpha(v')) = C(\theta_0) = \sigma_0^2 I(v = v') + \tau_0^2 \mathcal{M}_{\nu_0} \left(\frac{\|v - v'\|}{\phi_0} \right), \quad (7)$$

$$\text{cov}(\beta_k(v), \beta_k(v')) = C(\theta_1) = \tau_1^2 \mathcal{M}_{\nu_1} \left(\frac{\|v - v'\|}{\phi_1} \right). \quad (8)$$

No nugget variance is assumed for the components of PING to ensure smoothness. For small and moderate datasets, standard Markov chain Monte Carlo (MCMC) algorithms apply to the PING model and computation is straightforward. One advantage of the PING prior is the elements of the j th component ($\beta_j(v_1), \dots, \beta_j(v_n)$) have multivariate Gaussian full conditional distribution given the other $(q - 1)$ GPs, and thus Gibbs steps can be used to update the PING process parameters. For large n , however, these updates become slow and we use spectral methods, described in the remainder of this section.

4.1. The Model in Spectral Domain

Similar to Reich et al. (2018), we partially decorrelate the data by using the DFT. Let us denote spectral representation of the processes $Y_i(v), \alpha(v), \beta(v), \beta_k(v), X_i(v), Z_j(v)$ and $\epsilon_i(v)$ as $\tilde{Y}_i(\omega), \tilde{\alpha}(\omega), \tilde{\beta}(\omega), \tilde{\beta}_k(\omega), \tilde{X}_i(\omega), \tilde{Z}_j(\omega)$ and $\tilde{\epsilon}_i(\omega)$ for frequency $\omega \in \mathcal{F} \subset \mathbb{R}^3$. Since DFT preserves linearity, the spatial model in Equation (2) in the spectral domain can be written as

$$\tilde{Y}_i(\omega) = \tilde{\alpha}(\omega) + \tilde{X}_i(\omega) * \tilde{\beta}(\omega) + \sum_{j=1}^J \tilde{Z}_j(\omega) \gamma_{ij} + \tilde{E}_i(\omega), \quad (9)$$

$$\tilde{\beta}(\omega) = \tilde{\beta}_1(\omega) * \tilde{\beta}_2(\omega) * \dots * \tilde{\beta}_q(\omega). \quad (10)$$

The notation $*$ denotes convolution. The GP $\alpha(v), \beta_k(v)$, and $\epsilon_i(v)$ are stationary and defined over a discrete spatial domain. To avoid computationally expensive Bessel function and spectral aliasing calculations, we use the quasi-Matern spectral density (Guinness and Fuentes 2017), which mimics the flexibility of the Matern spectral density for $\tilde{\alpha}(\omega), \tilde{\beta}_k(\omega)$ and $\tilde{\epsilon}_i(\omega)$,

$$\lambda(\omega|\theta) = (\sigma^2, \tau^2, \phi, \nu) = \sigma^2 + \tau^2 \left[\frac{1}{\phi^2} + h(\omega) \right]^{-\nu-d/2}, \quad (11)$$

where d is the dimension, $\omega \in [0, 2\pi]^d$ and $h(\omega) = \sum_{j=1}^d \sin(\omega_j/2)^2$. More specifically,

$$\tilde{E}_i(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta)), \quad (12)$$

$$\tilde{\alpha}(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta_0)),$$

$$\tilde{\beta}_k(\omega) \sim \text{Normal}(0, \tilde{\lambda}(\omega|\theta_1)),$$

where $\tilde{\lambda}(\omega|\theta) = \lambda(\omega|\theta)/2$ if $\omega \in \{0, \pi\}^3$ and $\tilde{\lambda}(\omega|\theta) = \lambda(\omega|\theta)$ otherwise. All the parameters in θ have the same interpretation as in Equation (3). For $\tilde{\beta}_k$, the nugget variance is zero.

4.2. Imputation Method

Each spectral element $Y_i(\omega)$ is a function of $Y_i(v)$ for all $v \in \mathcal{V}$, and thus spectral methods require complete data. However, in practice, data are often not collected on a complete regular grid and thus the response $Y_i(v)$ is missing at many locations $v \in \mathcal{V}$ if we transform it into a regular grid. For example, in brain images, we consider the complete regular grid as the 3D cube in which the skull is inscribed; from this perspective the medical images involve missingness. Missing values are handled naturally in a Bayesian context within a Gibbs sampler that draws the missing values from their conditional distribution given the observed data and the other parameters. Because imputation is applied during each MCMC iteration to account for imputation uncertainty, this step must be computationally efficient.

Denote the conditional mean by

$$\mu_i(v) = \alpha(v) + X_i(v) \beta(v) + \sum_{j=1}^J Z_j(v) \gamma_{ij},$$

and define Y_{i1} to be the vector of observed data for subject i and Y_{i2} to be the vector representing the missing values. Likewise, let μ_{i1} and μ_{i2} be the corresponding vectors of means. The conditional distribution of (Y_{i1}, Y_{i2}) given all of the other parameters is

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} | \text{rest of the parameters} \sim \text{Normal} \left(\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \quad (13)$$

and thus the conditional distribution of Y_{i2} given Y_{i1} and the rest of the parameters is normal with mean $\mu_{i2} + \Sigma_{21} \Sigma_{11}^{-1} (Y_{i1} - \mu_{i1})$ and covariance $\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$.

For large datasets, direct sampling from this distribution is unfeasible. The limiting computational task in computing the conditional mean is solving a linear system with Σ_{11} . Since Σ_{11} is symmetric and positive definite, this can be achieved with a preconditioned conjugate gradient (PCG) algorithm (Golub and Van Loan 2012), an iterative method for solving the linear

system $\Sigma_{11}a = b$. The goal of iterative linear solvers is to generate a sequence a_1, a_2, \dots that converges to $a = \Sigma_{11}^{-1}b$. The algorithms generally require us to compute $\Sigma_{11}a_k$ at each iteration k to check for convergence and to generate the next vector in the sequence, and thus the algorithms are fast when this forward multiplication can be computed quickly. In this case, forward multiplications with Σ_{11} can be computed in $O(n \log n)$ time and $O(n)$ memory with circulant embedding algorithms (Wood and Chan 1994), as can the forward multiplication with Σ_{21} . This is because Σ_{11} and Σ_{21} can be embedded in the larger circulant matrix Σ , that is,

$$\Sigma \begin{bmatrix} a_k \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} a_k \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{11}a_k \\ \Sigma_{21}a_k \end{bmatrix},$$

and fast Fourier transform can be exploited to compute the forward multiplication with the (nested block) circulant matrix Σ , since (nested block) circulant matrices are diagonalizable by the (d -dimensional) DFT. The PCG algorithm uses an approximate inverse of Σ_{11} , called a preconditioner, to encourage the sequence a_k to converge to a is a small number of iterations.

Completing the imputation step requires us to simulate a residual vector with covariance matrix $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. To accomplish this, we first simulate a vector $(\varepsilon_{i1}, \varepsilon_{i2})$ with mean zero and covariance as in Equation (13), which is again efficient with circulant embedding. Then, we form and the residual $\varepsilon_{i2} - \Sigma_{21}\Sigma_{11}^{-1}\varepsilon_{i1}$, which has the desired and can be computed in the same fashion as the conditioned mean. Further computation details for the conditional draws can be found in Stroud, Stein, and Lysen (2017) and Guinness and Fuentes (2017).

4.3. Sampling

Total variances are updated from their posterior inverse gamma distributions. All other Matérn parameters are updated using Metropolis sampling. The DFT of the PING process parameters is the convolution of frequencies as in (10). Conducting a full conditional Gibbs update, even in the spectral domain, is computationally expensive. One can improve the computational effi-

ciency using PCG as described in Algorithm 1 of the supplementary materials. However, for our real-data application, this still imposes a serious computational burden. The existing Metropolis techniques for a joint update of large coefficient vectors, such as the gradient adjusted Metropolis–Hastings (Roberts and Rosenthal 1998) or Hamiltonian Monte Carlo (Duane et al. 1987) mix slowly. Here, we introduce a new sampling technique that uses Metropolis steps for updating each $\beta_k(\cdot)$. The general q -component model can be written as follows:

$$Y_i(v) = \alpha(v) + \beta_k(v)\beta_{-k}(v)t_i + E_i(v), \quad (14)$$

where $\beta_{-k} = \beta_1 \cdots \beta_{k-1} \cdot \beta_{k+1} \cdots \beta_q$. Denote the estimated values at the N -th stage of the MCMC iteration as Y^N (samples using PCG), α^N , β_k^N and β_{-k}^N . We can calculate the error at the N th stage as $E^N = Y^N - \alpha^N - \beta_k^N \beta_{-k}^N t$.

We can rewrite our model in (14) as

$$\frac{Y_i(v)}{\beta_{-k}(v)} = \frac{\alpha(v)}{\beta_{-k}(v)} + \beta_k(v)t_i + E_i(v) \left(\frac{1}{\beta_{-k}(v)} - 1 \right) + E_i(v).$$

Except for $\beta_k(v)$ and the last $E_i(v)$, replacing all other values by the ones from the N th step gives

$$\frac{Y_i^N(v)}{\beta_{-k}^N(v)} = \frac{\alpha^N(v)}{\beta_{-k}^N(v)} + \beta_k(v)t_i + E_i^N(v) \left(\frac{1}{\beta_{-k}^N(v)} - 1 \right) + E_i(v). \quad (15)$$

The notation Y_i^N denotes the a full response dataset including imputed missing values from the N th iteration. To update $\beta_k(\omega)$, we sample $\tilde{\beta}_k^u(\omega)$ according to step (ii) from Algorithm 1, take the inverse DFT to obtain $\beta_k^u(v)$, and then form the Metropolis candidate $\beta_k^c(v) + c \frac{\beta_k^u(v) - \beta_k^N(v)}{\|\beta_k^u(v) - \beta_k^N(v)\|_2}$. Here c acts as a tuning parameter and $\|\cdot\|$ denotes the ℓ_2 norm, defined as $\|\beta\|_2^2 = \int_{v \in \mathcal{V}} \beta^2(v) dv$. We are essentially sampling $\tilde{\beta}_k^u(\omega)$ from an approximated model and then shrinking it back toward $\tilde{\beta}_k^N(\omega)$. Smaller values of c generate higher acceptance rate and vice versa. This step is described in detail in Algorithm 1.

Algorithm 1: Sampling algorithm of β_1 in spectral domain for the model $Y_i(v) = \alpha(v) + \beta_1(v)\beta_{-1}(v)t_i + E_i(v)$

(i) Calculate $Q_i(v) = \frac{Y_i^N(v)}{\beta_{-k}^N(v)} - \frac{\alpha^N(v)}{\beta_{-k}^N(v)} - E_i^N(v) \left(\frac{1}{\beta_{-k}^N(v)} - 1 \right)$. The superscript N denotes the values at N -th step on the MCMC.

Transform $Q_i(v)$ into spectral domain to get $\tilde{Q}_i(\omega)$.

(ii) Generate $\tilde{\beta}_k^u(\omega) \sim \text{Normal}(M(\omega), V(\omega))$, where $V(\omega) = \frac{1}{\sum_i(t_i^2)}(1/\tilde{\lambda}(\omega|\theta_k) + 1/\tilde{\lambda}(\omega|\theta))^{-k}$, where $\tilde{\lambda}(\omega|\theta_k)$ and $\tilde{\lambda}(\omega|\theta)$

are the spectral variances of the prior on $\tilde{\beta}_k(\omega)$ and the error process $\tilde{E}_i(\omega)$ respectively as described in Section 4.1 with θ_1 and θ as corresponding Matérn parameters and $M(\omega) = \sum_i \tilde{Q}_i(\omega)/V(\omega)$.

(iii) Convert $\tilde{\beta}_k^u(\omega)$ into spatial domain $\beta_k^u(v)$ using spectral methods for spatial data as in Section 2 of Reich, Chang, and Foley (2014).

(iv) Adjust the update $\beta_k^c(v) = \beta_k^N(v) + c \frac{\beta_k^u(v) - \beta_k^N(v)}{\|\beta_k^u(v) - \beta_k^N(v)\|_2}$.

(v) Convert $\beta^c = \beta_k^c(v) \cdot \beta_{-k}^N(v)$ back into spectral domain to get $\tilde{\beta}^c$, again by using spectral methods for spatial data as in Section 2 of Reich, Chang, and Foley (2014).

(vi) Calculate the acceptance probability for the MH step,

$$P_{\beta_k^c, \beta_k^N} = \min \left\{ 1, \frac{\exp(-\sum_{i=1}^n \|(\tilde{Y}_i^N - \tilde{\beta}^c t_i)/\tilde{\lambda}(\cdot|\theta)\|_2^2 - \|\tilde{\beta}_k^c/\tilde{\lambda}(\cdot|\theta_k)\|_2^2)}{\exp(-\sum_{i=1}^n \|(\tilde{Y}_i^N - \tilde{\beta}^N t_i)/\tilde{\lambda}(\cdot|\theta)\|_2^2 - \|\tilde{\beta}_k^N/\tilde{\lambda}(\cdot|\theta_k)\|_2^2)} \right\}.$$

In our simulation, we adjust c to maintain an acceptance rate of around 0.6 for this scheme to ensure good mixing. If acceptance is lower than 0.55, we decrease the value of c , and if it is higher than 0.70, we increase c under the restriction that $c \leq \|\beta_1^u - \beta_1^N\|_2$. If we start the algorithm with small $E_i(v)$, the convergence is faster. To ensure that, we propose to get least-square estimates of α and β from the data. Then assign $\beta_i = \beta^{1/q}$ as the starting value if q is odd. We also recommend to take odd q , so that it will be easier to take $1/q$ -th power, although this is not necessary for starting from the least-square estimate. In the supplementary materials, Figure 7 illustrates the output of MCMC chain at the 5000th iteration with a random starting point for our algorithm as well as HMC and shows that our algorithm performs overwhelmingly better than HMC and also with respect to the speed as well.

We use this spectral method for all image-on-scalar regressions in this article. For the simulated image-on-image and scalar-on-image regressions in Section 5 the datasets are small, so we use Gibbs sampling for the PING process parameters; for larger datasets, the Metropolis scheme explained above could also be adapted to image-on-image and scalar-on-image regressions.

5. Simulation Results

In this section, we present simulation results for all three regression setups, namely image-on-scalar regression, image-on-image regression, and scalar-on-image regression. We compare the results in terms of mean squared error (MSE) with respect to the posterior mean, TP, and false negative (FN) for different levels of signal-to-noise ratios (SNR). To compute TP and FN, we need to identify the voxels with nonzero signals. To do that, we construct 95% credible intervals from the 0.025 and 0.975 quantiles of the MCMC samples at each voxel separately and identify the voxels for which the credible intervals do not include zero. Based on this identified set of voxels, we compute the proportions TP and FN, where TP is defined as the proportion of locations such that the credible intervals do not include zero given the true parameter value is non-zero and FN is defined as the proportion of locations such that the credible intervals do not include zero given the true parameter is zero.

While comparing estimation accuracies between our PING prior and other competing methods, we select q based on a leave-one-visit-out cross validation technique. In this approach, we select the value q that produces the smallest out-of-sample prediction MSE for each simulated dataset.

5.1. Image-On-Scalar Regression

Here, we consider the image-on-scalar regression model in Section 2 for images of dimension $20 \times 20 \times 20$ with 20 visits. The model is

$$Y_i(v) = \alpha(v) + t_i\beta(v) + e_i(v), \quad (16)$$

here $v \in \{1, \dots, 20\}^3$ with $i = 1, 2, \dots, 20$ and t_i 's are 20 equidistant points such that $\sum_i t_i = 0$ and $\sum_i t_i^2 = 20$ obtained by standardizing the times $i = 1, 2, \dots, 20$. The true signal is zero

for most of the spatial locations but has subregions that are non-zero. Let, $d_1 = (6, 14, 6)$, $d_2 = (6, 10, 14)$, $d_3 = (14, 6, 14)$, $d_4 = (14, 14, 14)$ and $d_5 = (6, 6, 6)$, $\kappa(v) = 2[\exp(-4\|v - d_1\|_2^2/20) + \exp(-1.5\|v - d_2\|_2^2/20) + \exp(-4\|v - d_3\|_2^2/20) + \exp(-4\|v - d_4\|_2^2/20) + \exp(-4\|v - d_5\|_2^2/20)]$. The true signal is $\beta(v) = \kappa(v)$ if $\kappa(v) \geq 0.1$ and $\beta(v) = 0$ otherwise. The plot of the true slope $\beta(v)$ is in the supplementary material (Figure 2). The error process $e_i(v)$ is assumed to be GP with stationary Matern covariance function. The true reparameterized Matern parameters for intercept process $\alpha(v)$ are (1, 0.95, 10, 1) and the last three parameters for error are (0.90, 10, 1). After generating the data on $20 \times 20 \times 20$ grid, we treat the values missing outside of the inner grid $18 \times 18 \times 18$. Here, we use the imputation technique to impute those missing values in our estimation.

We put GP prior with Matern covariance function on the intercept process $\alpha(v)$ and PING prior (Section 2.1) on the slope. We represent $\theta = (\sigma^2, \tau^2, \phi, v)$, $\theta_0 = (\sigma_0^2, \tau_0^2, \phi_0, v_0)$ and $\theta_1 = (\sigma_1^2, \tau_1^2, \phi_1, v_1)$ as the Matern parameters for the error, intercept and first component in the PING process prior on slope, respectively. For other components of the PING prior, the Matern parameters are (1, τ_1^2 , ϕ_1 , v_1). We reparameterize the Matern parameter $\theta = (\sigma^2, \tau^2, \phi, v)$ to $\theta' = (\vartheta^2, \zeta^2, \phi, v)$ as $\vartheta^2 = (\sigma^2 + \tau^2)$ and $\zeta^2 = \frac{\tau^2}{\sigma^2 + \tau^2}$. Here, ϑ^2 is called the total variance. The total variance of error is set at 0.09, 0.017, and 0.009 to achieve different SNRs which are mentioned in the Table 1. All these results, compiled in Table 1 are based on 50 replications and 10,000 postburn samples after burning in 10000 samples.

We fit the model with priors: $\vartheta^{-2}, \vartheta_0^{-2}, \vartheta_1^{-2} \sim \text{Gamma}(0.1, 0.1)$; $\text{logit}(\zeta), \log \phi, \log v \sim N(0, 1)$, and $\text{logit}(\zeta_0), \log \phi_0, \log v_0 \sim N(0, 1)$. For the PING process: We set $\zeta_1 = 1$ (as nugget variance is zero) and $\log \phi_1, \log v_1 \sim N(0, 1)$. The priors are the same for the next two regressions as well.

Table 1. Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage, and FP for the slope of the image-on-scalar simulation with different SNRs for Gaussian, and PING where q is selected via cross-validation in each dataset.

SNR	Metric	Fitted model	
		Gaussian	PING
1	Total MSE	2.2 (0.25)	2.3 (1.42)
	MSE for $\beta = 0$	2.49 (0.36)	1.25 (0.54)
	TP	0.97	0.94
	FP	0.06	0.00
	Coverage	0.92	0.98
	Total MSE	0.96 (0.12)	0.52 (0.04)
5	MSE for $\beta = 0$	1.49 (0.23)	0.50 (0.06)
	TP	1	1
	FP	2.19×10^{-1}	0.00
	Coverage	0.88	0.98
	Total MSE	0.76 (0.08)	0.28 (0.03)
	MSE for $\beta = 0$	1.19 (0.21)	0.26 (0.04)
10	TP	1	1
	FP	2.72×10^{-1}	0.00
	Coverage	0.85	0.99

NOTE: The reported MSEs and corresponding variances in bracket are all multiplied by 1000 for clarity.

From the values in Table 1, we infer that there is not much difference between PING estimates and GP when the SNR is low. However, as SNR increases, PING estimates become much better than GP. Across datasets, the most commonly selected number of components (35/50 datasets) was $q = 5$. Figure 3 of the supplementary material compares the estimates for one slice of the 3-D slope across different methods along with the true $\beta(v)$. GP prior overestimates the regions where the true value is zero as shown in Figure 3 of the supplementary material. This results in higher FP and higher MSE for locations where the true value is zero. We calculate the MSE for $\beta = 0$ as $\sum_{s: \beta_0(s)=0} (\beta(s))^2$ and total MSE as $\sum_s (\beta(s) - \beta_0(s))^2$, where $\beta_0(\cdot)$ is the truth. Here all methods have high power.

5.2. Image-On-Image Regression

We consider image-on-image regression model as in Section 3

$$Y_i(v) = \alpha(v) + \sum_{j=1}^{10} X_{ij}(v) \eta_j(v) + e_i(v), \quad (17)$$

on data collected over 100 locations, selected at random in $[0, 1]^2$ with $i = 1, \dots, 20$ observations at each location. The ten spatially varying predictors (X_{ij} 's) are generated using the reparameterized Matern parameters, generated randomly. First a random vector of four elements are generated from $N(0, 1)$. We exponentiate the first, third, and fourth element and take inverse logit transformation of the second element to get those reparameterized Matern parameters for each predictor. These predictors are generated only once for the whole simulation. We change the domain of the images for this simulation from previous case and only consider data at 100 locations to construct a dataset of manageable dimension for easier computation.

The error process $e_i(v)$ is assumed to be GP with stationary Matern covariance function, independent over i . And $\eta_j(v) = 0$ for $j = 1, \dots, 5$. Rest of those five η 's have the structures, plotted in Figure 4 of the supplementary material. These are zero at most of the locations with some non-zero subregions. To generate η_6 , we divide the whole $[0, 1]^2$ space into a 50×50 grid. Then we generate a random number h in $\{1, 2, 3\}$. Then, we generate h set of co-ordinates in $[0, 1]^2$. Let these be u_1, \dots, u_h . Let us define $\kappa(v) = \sum_{i=1}^h 2 \exp(-3||v - 50u_i||_2^2/50)$. Then, $\eta_6(v) = \kappa(v)$ if $\kappa(v) \geq 0.1$ and $\eta_6(v) = 0$ otherwise. Other four β_j 's are generated similarly. The true reparameterized Matern parameters for intercept are (1, 0.95, 10, 1) and last three parameters for error are (0.9, 10, 1). The total variance of error is set to 0.57, 0.11 and 0.06 to achieve different SNRs which are mentioned in the table. We report MSE, power, FP and coverage averaged over η . All these results, compiled in Table 2 are based on 50 replications and 5000 postburn samples after burning in 5000 samples.

In Table 2, we see that the PING process prior consistently gives a better estimate in terms of MSE. For most of the replicated datasets (around 60%), the cross-validation-based selection of q again turns out to be 5. The GP prior overestimates the regions where the true value is zero. This results in higher FP for GP prior. Here, all methods have similar power. As the SNR increases, the results using PING are even better than those

Table 2. Total MSE, MSE for the subregion with true $\eta = 0$ along with standard errors in the bracket, power, coverage and FP error for the slope of the image-on-image simulation with different SNRs for Gaussian, and PING as choices of prior.

SNR	Metric	Fitted model	
		Gaussian	PING
1	Total MSE	3.69 (0.44)	0.97 (0.19)
	MSE for $\eta = 0$	2.91 (0.42)	0.49 (0.14)
	TP	0.87	0.84
	FP	0.20×10^{-1}	0.01×10^{-1}
	Coverage	0.97	0.99
	Total MSE	0.97 (0.12)	0.20 (0.05)
5	MSE for $\eta = 0$	0.88 (0.11)	0.12 (0.03)
	TP	0.97	0.96
	FP	0.29×10^{-1}	0.01×10^{-1}
	Coverage	0.97	0.99
	Total MSE	0.54 (0.06)	0.11 (0.02)
	MSE for $\eta = 0$	0.5 (0.06)	0.07 (0.02)
10	TP	0.98	0.97
	FP	0.33×10^{-1}	0.01×10^{-1}
	Coverage	0.96	0.99
	Total MSE	0.96	0.99

NOTE: The number of components in PING are selected via cross-validation. The reported MSEs and corresponding variances in bracket are all multiplied by 1000 for clarity.

using Gaussian. For most of the replicated datasets, the optimal choice for q happens to be 4.

5.3. Scalar-On-Image Regression

Finally, we replicate the simulation from (Kang, Reich, and Staicu 2016) with 100 observations. For each observation, there is a two-dimensional image X_i of dimension 20×20 with an exponential covariance structure having range parameter 3. The model is

$$Y_i \stackrel{\text{ind}}{\sim} \text{Normal}(\sum_{j,k=1}^{20} X_{ijk} \beta_{jk}, \sigma^2), \quad (18)$$

Here, the coefficient $\beta = ((\beta_{jk}))_{1 \leq j,k \leq 20}$ is a matrix of dimension 20×20 . The true β is generated in such a way that it has five peaks. Let, $d_1 = (4, 16)$, $d_2 = (16, 4)$, $d_3 = (4, 4)$, $d_4 = (16, 16)$, and $d_5 = (10, 10)$ and $\kappa(v) = \sum_{i=1}^5 2 \exp(-20||v - d_i||_2^2/50)$. The true beta is $\beta(v) = \kappa(v)$ if $\kappa(v) \geq 0.1$ and $\beta(v) = 0$ otherwise. Only in this set up, we have number of observations much less than number of parameters to be estimated. We consider three choices of σ^2 in generating the data, 0.1, 1 and 1.5.

The prior for σ^{-2} is Gamma(0.1, 0.1). Rest of parameters have the same prior as in the previous subsections. In this subsection, we also compare our method with fused lasso (Tibshirani et al. 2005) and functional principal component analysis (fPCA) (Jones and Rice 1992). Fused lasso estimates are computed using the `genlasso` package in R (Arnold and Tibshirani 2020). After smoothing the images using `fbps` function of `refund` package (Goldsmith et al. 2020), eigendecomposition of the sample covariance is computed. After that lasso regularized principal components regression is performed. The leading eigenvectors that explain 95% of the variation in the

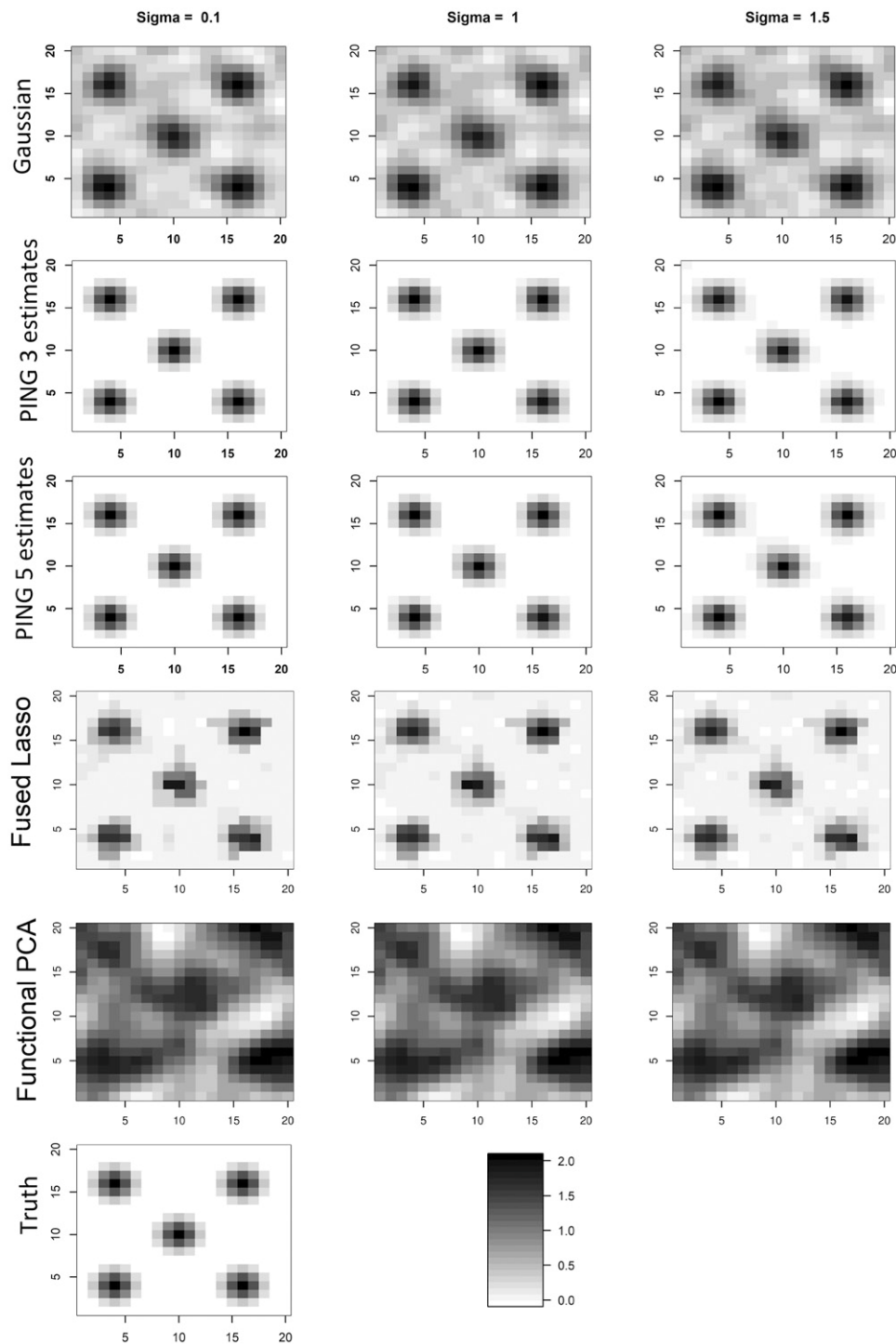


Figure 2. Comparison plot of the estimates for the slope of the scalar-on-image simulation with different true variances for Gaussian, PING 3, and PING 5 as choices of prior along with fused lasso and functional PCA estimates.

sample images are used to get the final estimate. The MSEs due to functional PCA estimates turn out to be overwhelming larger than the other competitive. Thus, they are not included in Table 3. However, Figure 2 includes pictorial illustrations for all the estimates.

We report MSE, TP, FP rate, and coverage in the estimation of the slope β matrix. All these results, compiled in Table 3, are based on 50 replications and 5000 postburn samples after burning in 5000 samples. Here, we see that the estimates from

the PING process prior are superior to those of the GP, fused lasso and functional PCA (fPCA) in all metrics. In particular, fPCA results are very noisy. In Figure 2 the estimated parameters from PING are less noisy than all other estimates.

We also compare PING with STGP method (Kang, Reich, and Staicu 2016). While comparing with STGP, we consider the low-rank approximation of each component of PING. The low-rank approximation is incorporated following the works in the kernel convolution of Higdon, Swall, and Kern (1999) and

Table 3. Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power, coverage and false-positive error for the slope of the scalar-on-image regression model with different true variances for Gaussian, and PING as choices of prior along with fused lasso and functional PCA estimates.

σ	Metric	Fitted model		
		Gaussian	PING	Fused lasso
1.5	Total MSE	30.1 (1.8)	6.2 (2.1)	28.2 (76.9)
	MSE for $\beta = 0$	20.1 (1.32)	2.1 (0.66)	7 (29.6)
	TP	0.79	0.75	0.56
	False positive	8.14×10^{-3}	0.34×10^{-3}	0.00
	Coverage	0.97	0.96	–
	Total MSE	25.6 (1.71)	3.3 (0.77)	20.9 (56.32)
	MSE for $\beta = 0$	1.66 (1.12)	0.11 (0.32)	4.36 (18.64)
1	TP	0.86	0.91	0.69
	False positive	14.24×10^{-3}	0.68×10^{-3}	3.39×10^{-3}
	Coverage	0.96	0.96	–
	Total MSE	20.90 (0.98)	0.31 (0.02)	14.73 (39.92)
	MSE for $\beta = 0$	12.75 (0.44)	0.18 (0.01)	2.28 (10.99)
0.1	TP	0.94	1.00	0.90
	False positive	1.76×10^{-2}	0.37×10^{-2}	0.15
	Coverage	0.94	0.96	–

NOTE: The reported MSEs and corresponding variances in bracket are all multiplied by 1000 for clarity.

Table 4. Total MSE, MSE for the subregion with true $\beta = 0$ along with standard errors in the bracket, power and Type I error for the slope of the scalar-on-image simulation with different true variances for soft-thresholded Gaussian process (STGP) and PING as choices of prior.

σ	Metric	Fitted model	
		STGP	PING
1.5	Total MSE	4.22 (1.57)	1.30 (0.59)
	MSE for $\beta = 0$	0.11 (0.15)	0.48 (0.18)
	TP	0.85	0.95
	False positive	0.10	0.06
	Total MSE	3.8 (1.54)	0.68 (0.28)
	MSE for $\beta = 0$	0.08 (0.13)	0.32 (0.09)
	TP	0.86	0.96
1	False positive	0.10	0.06
	Total MSE	2.86 (1.26)	0.026 (0.08)
	MSE for $\beta = 0$	0.03 (0.09)	0.12 (0.02)
	TP	0.86	0.96
	False positive	0.11	0.07

NOTE: The reported MSEs and corresponding variances in bracket are all multiplied by 1000 for clarity.

Nychka et al. (2015) as in Kang, Reich, and Staicu (2016). Due to this modification, all the results change from the previous table and the comparison results with STGP are provided separately in Table 4. We show results using PING with q , selected via our one-leave-out cross-validation. For most of the replicated datasets (around 75%), the cross-validation-based selection of q turns out to be 3. In the supplementary materials, a plot of MSEs for different choices of q is given.

We can see that the estimates from the PING process prior are superior STGP in terms of overall MSE, the TP rate, and

coverage, but not in terms of the MSE at the subregion where the truth is zero. In fact STGP and PING-based estimates are nearly indistinguishable graphically; results not included here. Due to thresholding, STGP estimates are more conservative. In this case, the cross-validation-based selection of q turns out to be 8 for around 70% of the replicated datasets.

In all the simulations, PING estimates work well as a shrinkage estimate, producing much better results than other alternatives. All the PING estimates are presented for the q , selected by our proposed cross validation. Thus the proposed cross validation technique works well to make good selection for q .

6. Application to Longitudinal MRI Data

Next, we turn to the study of MS using MRI images. In a natural history cohort followed at the National Institute for Neurological Disorders and Stroke, each subject was scanned approximately once per month over several hospital visits. In the subset of the study published in Sweeney et al. (2016), several individuals were scanned over 3 years. We focus on the set of images from a single subject. Using a 1.5T GE scanner with clinically optimized scanning parameters, whole-brain magnetization transfer fluid attenuation inversion recovery (FLAIR) volumes were acquired. All the modalities were interpolated to a voxel size of 1 mm^3 yielding images of dimension $182 \times 218 \times 182$. We use normalized FLAIR images in our study by z-scoring using normal-appearing white matter (Shinohara et al. 2011, 2014). We also use subtraction-based logistic inference for modeling and estimation (SuBLIME) mask. The SuBLIME mask is a 4D image with three dimensions for space and one for time. For each time point, the 3D image is a map of where there were new/enlarging lesions between the corresponding pair of time points. All images were registered longitudinally and across the modalities and rigidly aligned to the Montreal Neurological Institute standard space (Fonov et al. 2009). Sweeney et al. (2016) have a complete description of the study along with the acquisition parameters. We are interested to study how the image intensity varies over time. Let $Y_i(\nu)$ denote the image intensity at a 3-dimensional voxel ν of i -th image at time t_i , which denotes the number of days passed between i th and the first visit of a single subject. In general, ν is used to denote voxel. We normalize the time covariate t_i and set the image of the first visit as the baseline. We consider the following model from Section 2, $Y_i(\nu) = \alpha(\nu) + \beta(\nu)t_i + E_i(\nu)$ where $\alpha(\nu)$ is the spatially varying intensity image at baseline visit and $\beta(\nu)$ quantifies the brain regions that are deteriorated over time due to MS. We consider a linear model in time after performing some exploratory analysis on model selection among the higher order polynomials in t at each voxel; the linear model corresponds to the smallest AIC and BIC. Preliminary analysis that confirms linearity in the change overtime is included in the supplementary material. It is expected that the healthy brain tissue does not change much, while changes occur in the MS-affected brain regions and the number of such regions is small. Thus, the effect $\beta(\nu)$ is expected to be sparse, and in addition, it is also desired to be piecewise smooth and continuous, due to the complex spatial dependence in the brain. However, we have

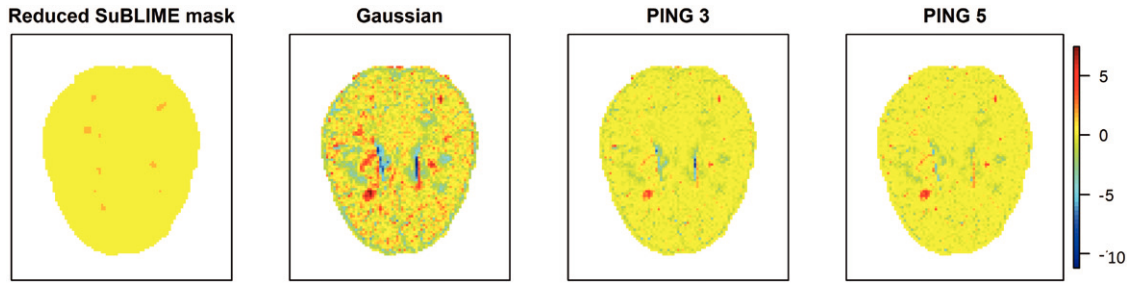


Figure 3. Estimated slope $\beta(v)$ of the middle slice using different priors along with the color scale along the reduced SuBLIME mask.

also fitted following nonlinear model:

$$Y_i(v) = \sum_{j=1}^K \beta_j(v) B_j(t_i) + E_i(v), \quad (19)$$

here $B_j(\cdot)$ stands for the B-Spline basis functions and $\beta_j(v)$ is the spatially varying coefficients. Since $\beta_1(v)$ corresponds to the effect at the 0th time point, we put PING(1) which is standard Gaussian prior on $\beta_1(v)$ and PING(q) on $\beta_j(v)$ for $j \geq 2$. We consider linear B-spline bases with four equidistant knots after performing prediction MSE based cross validation on the degree and number of bases.

For the error process, we consider the nonstationary covariance model as discussed in Section 2.2. We use GP prior with Matern covariance function for α . Further details of the model specifications are $\mathbf{y}_i \sim N(0, \Sigma)$ where $\Sigma \sim IW(J + 0.1, \frac{c}{J+0.1} I_J)$ and $c^{-1} \sim \text{Gamma}(0.1, 0.1)$, where “IW” stands for inverse Wishart, $J = 6^3 = 216$ B-spline basis functions. We reduce the dimensionality of the images to $91 \times 109 \times 91$ using `resize` function of `imager` package of R due to computational and storage issues; the reduced images preserve the overall structure of the original images. The time of the visits is roughly every month. We normalize the time vector such that the sum of squares of the times is one. We present the analysis for one MS subject in the study. The real data plot of an axial view for the subject’s first 12 visits is in Figure 6 of the supplementary material.

We use both the proposed method with the signal modeled using the PING process, and with the signal modeled via a GP. We select optimal q based on a leave-one-out cross-validation method as described in Section 5.1. The estimates are based on 5000 postburn MCMC samples after 5000 burn-in. We sample the values in the image outside of the brain using the techniques of Section 4.2 after every 30 iterations. We performed the cross validation for q between 3 and 9 and found that $q = 5$ gave the smallest MSE. To show the effect of q , we present the results for $q = 3$ and $q = 5$ in Figures 3 and 4. For the nonlinear model in Equation (19), the prediction MSE is the best for PING(3) prior on the B-spline coefficients $\beta_j(v)$ for $j \leq 2$. For PING-3, it is around 0.09 which is slightly better than the prediction MSE values from the linear model. For, PING-5 and Gaussian, prediction MSEs are 0.12 and 0.10, respectively. The overall prediction MSE is marginally improved for the nonlinear model over the linear model which had 0.13 as the best prediction MSE. Figure 3 compares the estimates from Gaussian, PING-3, and PING-5 for the middle cross section along with the reduced SuBLIME mask. The reduced SuBLIME is a 3-D mask which aggregates

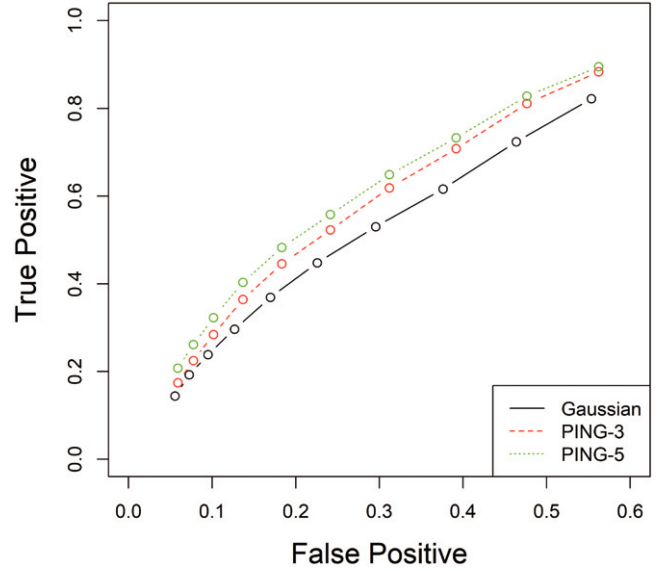


Figure 4. ROC curve constructed from the Gaussian, PING-3 and PING-5 estimates in detecting lesions flagged by reduced SubLIME mask.

the original SuBLIME masks over time. Thus, this reduced SuBLIME mask identifies all the brain regions that turn into lesions between the subject’s first visit to the 11th visit. The estimated effect with a Gaussian prior is very noisy, and it is difficult to be used by practitioners who wish to identify regions affected by the MS. In contrast, the PING-based estimates clearly highlight the regions of interest that are affected by the disease. The results agree with the ones obtained with the SuBLIME. To further investigate this, we plot the receiver operating characteristic (ROC) curve in detecting the lesions flagged by SuBLIME mask in Figure 4 based on the estimates of β from the model in 1. We see that PING-3 estimate is better than Gaussian and PING-5 is slightly better than PING-3. Due to storage issues of the posterior samples, the ROC curves are constructed based on the same level of cutoffs on the posterior estimates in detecting new lesions, flagged by the SuBLIME mask. PING estimates perform much better than the corresponding Gaussian estimate based on the ROC curve. For the nonlinear model, it is difficult to prepare such plots combining the B-spline coefficients.

7. Conclusion and Discussion

We propose a new class of prior, called the PING prior, for estimating spatially sparse and smooth signals. We analyze the performance of our prior in different kinds of image regressions,

namely image-on-scalar, image-on-image, and scalar-on-image. We develop techniques to tackle huge dimensional datasets by transforming into the spectral domain. Our simulations show that this new prior outperforms the Gaussian prior for all the image regressions, we considered. An R package to fit PING for different image regression is available at <https://github.com/royarkaprava/PING>.

Our simulation results suggest that PING priors give better estimates than Gaussian at the locations where true value is zero. This results in lower FP for PING. All of the methods have high TP for both image-on-scalar and image-on-image regression models. For the scalar-on-image model, they even have better TP than Gaussian along with lower FP and MSEs. The versatility in application of this prior is well studied in the simulation section of this article. In the MRI data acquired longitudinally in a patient with MS, although there is little improvement in prediction MSE from Gaussian to PING, the disease affected areas are more easily distinguishable in PING estimates due to the shrinkage. For the nonlinear model, wavelet bases can also be considered instead of B-splines (Torabi et al. 2007; Hackmack et al. 2012; Zhang et al. 2016; Wang et al. 2016). In this case, selecting the PING prior for each basis coefficient allows for abrupt changes in space and time.

An area of future work is the selection strategy for q , the number of components in PING. In this article, our selection criterion is based on leave-one-out cross-validation performance. The selected q has also shown the best performance in Figure 4 in detecting the affected brain regions. However, one possible future direction is to put a prior on q and develop computations using reversible jump MCMC (RJMCMC). RJMCMC algorithms are in general computationally expensive. To circumvent the computational complexities in RJMCMC, another possible alternative is to write the PING prior as,

$$\beta(v) = \sigma \prod_{i=1}^Q \beta_i^{Z_i}(v), \quad \beta_i(v) \sim \text{GP}(0, K),$$

$$Z_i \sim \text{Ber}(p_i), \quad p_i \sim \text{Beta}(\alpha, \beta),$$

for some fixed large enough Q . Due to the huge dimensionality of MRI data, the convergence is rather slow using this prior. However, for smaller dimensional datasets, this is a possible alternative to selecting the number of components in the PING prior.

The applicability of our proposed PING prior is not restricted to the models, described in this article. This prior can be used to estimate any sparse and piece-wise smooth function. As long as the dimension of the data is manageable, one can develop efficient Gibbs sampler to estimate each component of the PING prior. One can also consider to use PCG to update each component of PING from its full conditional using the methods developed in Section 4.2 for moderately large images exploiting conditional distribution of multivariate normal. To do that, we need to stack all the observations and one of the PING components in one large vector and get a conditional distribution of that large vector given other components of PING. Based on that expression, one can sample from the full conditional distribution of that PING component given the stack of observations using PCG. Details of this computation technique are in the supplementary materials.

Acknowledgments

We thank to the editor, associate editor, and the anonymous referees for their constructive comments which helped to improve the overall presentation of the article.

Funding

This research was supported partially by the National Institutes of Health grant R01 ES027892 (Guinness), R01 NS085211 (Staicu), R01 MH086633 (Staicu), R21NS093349 (Shinohara), R01NS085211 (Shinohara) and R01MH112847 (Shinohara), and the National Science Foundation grant DMS 1613219 (Guinness) and DMS 1454942 (Staicu). Shinohara was also partially supported by RG-1707-28586 from the National Multiple Sclerosis Society. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Supplementary Material

Supplementary materials of Sparse and smooth signal. Proof of the theorem, tables, and figures, referenced in Sections 2, 5, and 6 are available in this pdf. (pdf). R-code: In submissioncode.zip file all the necessary R-codes are placed.

ORCID

Arkaprava Roy  <http://orcid.org/0000-0001-6976-9967>

References

- Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [1068]
- Arnold, T. B., and Tibshirani, R. J. (2020), *genlasso: Path Algorithm for Generalized Lasso Problems*, R package version 1.5. [1075]
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), "Dirichlet-Laplace Priors for Optimal Shrinkage," *Journal of the American Statistical Association* 110, 1479–1490. [1068]
- Boehm-Vock, L., Reich, B., Fuentes, M., and Dominici, F. (2015), "Spatial Variable Selection Methods for Investigating Acute Health Effects of Fine Particulate Matter Components," *Biometrics*, 71, 167–177. [1069]
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1068]
- Chen, Y., Wang, X., Kong, L., and Zhu, H. (2016), "Local Region Sparse Learning for Image-on-Scalar Regression," arXiv:1605.08501 [1068]
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216–222. [1073]
- Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009), "Unbiased Nonlinear Average Ageappropriate Brain Templates From Birth to Adulthood," *NeuroImage*, 47, S102. [1077]
- Gaunt, R. E. (2018), "Products of Normal, Beta and Gamma Random Variables: Stein Operators and Distributional Theory," *Brazilian Journal of Probability and Statistics*, 32, 437–466. [1069]
- Gelfand, A., Kim, H., Sirmans, C. F., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396. [1071]
- Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014), "Smooth scalar-On-Image Regression Via Spatial Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*, 23, 46–64. [1068,1072]
- Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2020), *Refund: Regression With Functional Data*, R package version 0.1-23. [1075]
- Golub, G. H., and Van Loan, C. (2012). *Matrix Computations*, Chapter 10. Baltimore, MD: JHU Press. [1072]

- Griffin, J. E., and Brown, P. J. (2010), "Inference With Normal-Gamma Prior Distributions in Regression Problems," *Bayesian Analysis*, 5, 171–188. [1068]
- Guinness, J., and Fuentes, M. (2017), "Circulant Embedding of Approximate Covariances for Inference From Gaussian Data on Large Lattices," *Journal of Computational and Graphical Statistics*, 26, 88–97. [1072,1073]
- Hackmack, K., Paul, F., Weygandt, M., Allefeld, C., Haynes, J.-D., Initiative, A. D. N., et al. (2012), "Multi-scale Classification of Disease Using Structural MRI and Wavelet Transform," *Neuroimage*, 62, 48–58. [1079]
- Hazra, A., Reich, B. J., Reich, D., Shinohara, R., and Staicu, A. (2019), "A Spatio-Temporal Model for Longitudinal Image-On-Image Regression," *Statistics in Biosciences*, 11, 22–46. [1071]
- Higdon, D., Swall, J., and Kern, J. (1999), "Non-Stationary Spatial Modeling," *Bayesian Statistics 6—Proceedings of the Sixth Valencia Meeting*, 761–768. [1076]
- Jhuang, A., Fuentes, M., Jones, J., Esteves, G., Fancher, C., Furman, M., and Reich, B. (2018), "Spatial Signal Detection Using Continuous Shrinkage Priors," in *reversion, Technometrics* [1069]
- Jog, A., Carass, A., and Prince, J. (2013), "Magnetic Resonance Image Example-based Contrast Synthesis," *IEEE Transaction on Medical Imaging*, 32, 2348–2363. [1069]
- Jog, A., Carass, A., Roy, S., Pham, D., and Prince, J. (2015), "Mr Image Synthesis by Contrast Learning on Neighborhood Ensembles," *IEEE Transaction on Medical Imaging*, 34, 63–76. [1071]
- Jog, A., Carass, A., Roy, S., Pham, D., and Prince, J. (2017), "Random Forest Regression for Magnetic Resonance Image Synthesis," *IEEE Transaction on Medical Imaging*, 35, 475–488. [1071]
- Jones, M., and Rice, J. A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46, 140–145. [1075]
- Kang, J., Reich, B. J., and Staicu, A. M. (2016), "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process," arXiv: 1604.03192. [1068,1072,1075,1076,1077]
- Li, F., Zhang, T., Wang, Q., Gonzalez, M., Maresh, E., and Coan, J. (2015), "Spatial Bayesian Variable Selection and Grouping in High-Dimensional Scalar-On-Image Regressions," *Annals of Applied Statistics*, 9, 687–713. [1068,1072]
- Mardia, K. (1970), "Measures of Multivariate Skewness and Kurtosis With Applications," *Biometrika*, 57, 519–530. [1070]
- Mejia, A., Sweeney, E., Dewey, B., Nair, G., Sati, P., Shea, C., Reich, D., and Shinohara, R. (2016), "Statistical Estimation of T_1 Relaxation Times Using Conventional Magnetic Resonance Imaging," *Neuroimage*, 133, 176–188. [1068]
- Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036. [1068]
- Morris, J., Baladandayuthapani, V., Herrick, R., Sanna, P., and Gutstein, H. (2011), "Automated Analysis of Quantitative Image Data Using Isomorphic Functional Mixed Models, With Application to Proteomics Data," *The Annals of Applied Statistics*, 5, 894–923. [1069,1071]
- Musgrove, D. R., Hughes, H., and Eberly, L. E. (2016), "Fast, Fully Bayesian Spatiotemporal Inference for fMRI Data," *Biostatistics*, 17, 291–303. [1068]
- Noh, H., and Park, B. (2010), "Sparse Varying Coefficient Models for Longitudinal Data," *Statistica Sinica*, 20, 1183–1202. [1069]
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599. [1077]
- Pomann, G., Staicu, A., Lobaton, E., Mejia, A., Dewey, B., Reich, D., Sweeney, E., and Shinohara, R. (2016), "A Lag Functional Linear Model for Prediction of Magnetization Transfer Ratio in Multiple Sclerosis Lesions," *The Annals of Applied Statistics*, 10, 2325–2348. [1068]
- Reich, B., Guinness, J., Vandekar, S., Shinohara, R., and Staicu, A. (2018), "Fully-Bayesian Spectral Methods for Imaging Data," *Biometrics*, 74, 645–652. [1071,1072]
- Reich, B. J., Chang, H. H., and Foley, K. M. (2014), "A Spectral Method for Spatial Downscaling," *Biometrics*, 70, 932–942. [1073]
- Roberts, G. O., and Rosenthal, J. S. (1998), "Optimal Scaling of Discrete Approximations to Langevin Diffusions," *Journal of the Royal Statistical Society, Series B*, 60, 255–268. [1073]
- Shen, W., and Ghosal, S. (2015), "Adaptive Bayesian Procedures Using Random Series Priors," *Scandinavian Journal of Statistics*, 42, 1194–1213. [1071]
- Shinohara, R., Crainiceanu, C., Caffo, B., Gaitán, M., and Reich, D. (2011), "Population-Wide Principal Component-Based Quantification of Blood-Brain-Barrier Dynamics in Multiple Sclerosis," *NeuroImage*, 57, 1430–1446. [1077]
- Shinohara, R., Sweeney, E., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., for the Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, the Alzheimer's Disease Neuroimaging Initiative (2014), "Statistical Normalization Techniques for Magnetic Resonance Imaging," *NeuroImage*, 6, 9–19. [1077]
- Stojanac, Z., Suess, D., and Kliesch, M. (2017), "On the Distribution of a Product of N Gaussian Random Variables," *Wavelets and Sparsity XVII* 10394, 1039419. [1069]
- Stroud, J. R., Stein, M., and Lysen, S. (2017), "Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice," *Journal of Computational and Graphical Statistics*, 26, 108–120. [1073]
- Sweeney, E., Shinohara, R., Shea, C., Reich, D., and Crainiceanu, C. (2013), "Automatic Lesion Incidence Estimation and Detection in Multiple Sclerosis Using Multisequence Longitudinal MRI," *American Journal of Neuroradiology*, 34, 68–73. [1069]
- Sweeney, E. M., Shinohara, R. T., Dewey, B. E., Schindler, M. K., Muschelli, J., Reich, D. S., Crainiceanu, C. M., and Eloyan, A. (2016), "Relating Multi-sequence Longitudinal Intensity Profiles and Clinical Covariates in Incident Multiple Sclerosis Lesions," *NeuroImage*, 10, 1–17. [1068,1077]
- Tang, Y., Wang, H., and Zhu, Z. (2013), "Variable Selection in Quantile Varying Coefficient Models With Longitudinal Data," *Computational Statistics & Data Analysis*, 57, 435–449. [1069]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1068]
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness Via the Fused Lasso," *Journal of the Royal Statistical Society, Series B*, 67, 91–108. [1068,1075]
- Torabi, M., Moradzadeh, H., Vaziri, R., Ardekani, R. D., and Fatemizadeh, E. (2007), "Multiple Sclerosis Diagnosis Based on Analysis of Subbands of 2-D Wavelet Transform Applied on MR-Images," *2007 IEEE/ACS International Conference on Computer Systems and Applications*, 717–721. [1079]
- Wang, S.-H., Zhan, T.-M., Chen, Y., Zhang, Y., Yang, M., Lu, H.-M., Wang, H.-N., Liu, B., and Phillips, P. (2016), "Multiple Sclerosis Detection Based on Biorthogonal Wavelet Transform, RBF Kernel Principal Component Analysis, and Logistic Regression," *IEEE Access*, 4, 7567–7576. [1079]
- Wang, X., and Zhu, H. (2017), "Generalized Scalar-on-Image Regression Models Via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [1068,1072]
- Wood, A., and Chan, G. (1994), "Simulation of Stationary Gaussian Process in $[0, 1]^d$," *Journal of Computational and Graphical Statistics*, 3, 409–432. [1073]
- Yan, B., and Liu, Y. (2017), "Smooth Image-on-Scalar Regression for Brain Mapping," arXiv:1703.05264. [1068]
- Zhang, L., Guindani, M., Versace, F., Engelmann, Jeffrey M., and Vannucci, M. (2016), "A Spatiotemporal Nonparametric Bayesian Model of Multi-Subject fMRI Data," *The Annals of Applied Statistics*, 10, 638–666. [1068]
- Zhang, Y., Lu, S., Zhou, X., Yang, M., Wu, L., Liu, B., Phillips, P., and Wang, S. (2016), "Comparison of Machine Learning Methods for Stationary Wavelet Entropy-Based Multiple Sclerosis Detection: Decision Tree, k-Nearest Neighbors, and Support Vector Machine," *Simulation*, 92, 861–871. [1079]