Bayesian Data Sketching for Varying Coefficient Regression Models

Abstract

Varying coefficient models are popular tools in estimating nonlinear regression functions in functional data models. Their Bayesian variants have received limited attention in large data applications, primarily due to the prohibitively slow posterior computations using Markov chain Monte Carlo (MCMC) algorithms. We introduce Bayesian data sketching for varying coefficient models to obviate computational challenges presented by large sample sizes. To address the challenges of analyzing large data, we compress functional response vector and predictor matrix by a random linear transformation to achieve dimension reduction and conduct inference on the compressed data. Our approach distinguishes itself from several existing methods for analyzing large functional data in that it requires neither the development of new models or algorithms nor any specialized computational hardware while delivering fully model-based Bayesian inference. Well-established methods and algorithms for varying coefficient regression models can be applied to the compressed data. We establish posterior contraction rates for estimating the varying coefficients and predicting the outcome at new locations under the randomly compressed data model. We use simulation experiments and conduct a spatially varying coefficient analysis of remote sensed vegetation data to empirically illustrate the inferential and computational efficiency of our approach.

Keywords: B-splines; Predictive Process; Posterior contraction; Random compression matrix; Varying coefficient models.

1 Introduction

We develop an inferential framework for functional data analysis using Bayesian data sketching to achieve scalable inference for massive functional data sets. "Data sketching" (Vempala, 2005; Halko et al., 2011; Mahoney, 2011; Woodruff, 2014; Guhaniyogi and Dunson, 2015, 2016) is a method of compression that is being increasingly employed for analyzing massive amounts of data. The entire data set is compressed before being analyzed for computational efficiency. Data sketching proceeds by transforming the original data through a random linear transformation to produce a much smaller number of data samples and we conduct the analysis on the compressed data thereby achieving dimension reduction. Furthermore, the original data is neither accessed nor exactly recoverable from the compressed data, which preserves data confidentiality.

While such developments have primarily focused on ordinary linear regression and penalized linear regression (Zhang et al., 2013; Chen et al., 2015; Dobriban and Liu, 2018; Drineas et al., 2011; Ahfock et al., 2017; Huang, 2018), our innovation lies in developing such methods for functional regression models. The primary challenge distinguishing the current manuscript from existing data sketching methods is our pursuit of inference for the underlying effects of functional coefficients in the context of varying regression models. While bearing some similarities, our current contribution differs from compressed sensing (Donoho, 2006; Ji et al., 2008; Candes and Tao, 2006; Eldar and Kutyniok, 2012; Yuan et al., 2014) in the inferential objectives. Specifically, compressed sensing solves an inverse problem by "nearly" recovering a sparse vector of responses from a smaller set of random linear transformations. In contrast, our functionally indexed response vector is not necessarily sparse. Also, we do not seek to (approximately) recover the original values in the response vector so our method is applicable to situations where preserving confidentiality of the response (and predictors) is important.

We consider a varying-coefficient model (VCM) where all functional variables (response and predictors) are defined on a d-dimensional indexed space $\mathcal{D} \subseteq \mathbb{R}^d$. For temporal data $d = 1$ and for spatial data applications $d = 2$, while for spatial-temporal applications the domain is $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R}^+$ and the index is a space-time tuple ($\boldsymbol{u} = (\boldsymbol{s}, t)$). For each index $\boldsymbol{u} \in \mathcal{D}$, the functional response $y(\boldsymbol{u}) \in \mathcal{Y} \subseteq \mathbb{R}$ and P functional predictors $x_1(\boldsymbol{u}), ..., x_P(\boldsymbol{u}) \in \mathcal{X} \subseteq \mathbb{R}$, are related according to a posited varying coefficients regression model

$$y(\boldsymbol{u}) = \sum_{j=1}^{P} x_j(\boldsymbol{u})\beta_j + \sum_{j=1}^{\tilde{P}} \tilde{x}_j(\boldsymbol{u})w_j(\boldsymbol{u}) + \epsilon(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{\beta} + \tilde{\boldsymbol{x}}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{w}(\boldsymbol{u}) + \epsilon(\boldsymbol{u}) \ , \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^{\mathrm{T}}$ is a $P \times 1$ vector of functionally static coefficients, $\tilde{\boldsymbol{x}}(\boldsymbol{u}) = (\tilde{x}_1(\boldsymbol{u}), \tilde{x}_2(\boldsymbol{u}), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ is a $\tilde{P} \times 1$ vector comprising a subset of predictors from $\boldsymbol{x}(\boldsymbol{u})$ (so $\tilde{P} \leq P$) whose impact on the response is expected to vary over the functional inputs, $\boldsymbol{w}(\boldsymbol{u}) = (w_1(\boldsymbol{u}), w_2(\boldsymbol{u}), \ldots, w_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ is a $\tilde{P} \times 1$ vector of functionally varying regression slopes, and $\epsilon(\boldsymbol{u}) \overset{iid}{\sim} N(0, \sigma^2)$ captures measurement error variation at location $\boldsymbol{u}$. Such functionally-varying regression coefficient models are effective tools for estimating the functionally varying impact of predictors on the response in time series (see, e.g., Chen and Tsay, 1993; Cai et al., 2000, and references therein), in spatial applications (see, e.g., Gelfand et al., 2003; Wheeler and Calder, 2007; Finley et al., 2011; Guhaniyogi et al., 2013; Kim and Wang, 2021, and references therein) and in spatial-temporal data analysis (see, e.g., Lee et al., 2021, and references therein). When $d = 2$, customary geostatistical regression models with only a spatially-varying intercept emerge if the first column of $x(\boldsymbol{u})$ is the intercept and $\tilde{P} = 1$ with $\tilde{x}_1(\boldsymbol{u}) = 1$. Spatially-varying coefficient models, a class of varying coefficient models for $d = 2$, also offer a process-based alternative to widely used geographically weighted regression (see, e.g., Brunsdon et al., 1996) for modeling nonstationary behavior in the mean. Finley (2011) offers a comparative analysis and highlights the richness of (1) in ecological applications.

Bayesian inference for (1) is computationally expensive for large data sets, as are commonplace today, due to the high-dimensional covariance matrix introduced by $w(\boldsymbol{u})$ in (1). Modeling high-dimensional dependent functional data has been attracting significant interest and the burgeoning literature on diverse aspects of scalable methods—which has largely adapted and built from scalable spatial models (see, e.g., Banerjee, 2017; Heaton et al., 2019, for reviews in spatial statistics)—is too vast to be comprehensively reviewed here. Briefly, model-based dimension reduction in functional data models have proceeded from low-rank or fixed rank representations (e.g., Cressie and Johannesson, 2008; Banerjee et al., 2008; Wikle, 2010; Snelson and Ghahramani, 2005; Burt et al., 2020), multi-resolution approaches (e.g., Nychka et al., 2015; Katzfuss, 2017; Guhaniyogi and Sansó, 2018), sparsity-inducing processes (e.g., Vecchia, 1988; Datta et al., 2016; Katzfuss and Guinness, 2021; Peruzzi et al., 2020) and divide-and-conquer approaches such as meta-kriging (Guhaniyogi and Banerjee, 2018; Guhaniyogi et al., 2020b,a).

While most of the aforementioned methods entail new classes of models and approximations, or very specialized high-performance computing architectures, Bayesian data sketching has the advantage that customary exploratory data analysis tools, well-established methods and well-tested available algorithms for implementing (1) can be applied to the sketched data set without recourse to new algorithmic or software development. We pursue fully model-based Bayesian data sketching, where inference proceeds from a hierarchical model (Cressie and Wikle, 2015; Banerjee et al., 2014). The hierarchical approach to functional data analysis is widely employed for inferring on model parameters that may be weakly identified from the likelihood alone and, more relevantly for substantive inference, for estimating the functional relationship between response and predictors over the domain of interest. For analytic tractability we model the varying coefficients using basis expansions (Wikle, 2010; Wang et al., 2008; Wang and Xia, 2009; Bai et al., 2019) rather than Gaussian processes.

We exploit and adapt some recent developments in the theory of random matrices to relate the inference from the compressed data with the full scale functional data model. We establish consistency of the posterior distributions of the varying coefficients and analyze the predictive efficiency of our models based upon the compressed data. Posterior contraction of varying coefficient (VC) models have been investigated by a few recent articles. For example, Guhaniyogi et al. (2020a) derive minimax-optimal posterior contraction rates for Bayesian VC models under GP priors when the number of predictors P is fixed. Deshpande et al. (2020) also derived near-optimal posterior contraction rates under BART priors, and Bai et al. (2019) showed asymptotically optimal rate of estimation for varying coefficients with a variable selection prior on varying coefficients. We address these questions in the context of data compression, which has largely remained unexplored.

The balance of this article proceeds as follows. Section 2 develops our data sketching approach and discusses Bayesian implementation of VC models with sketched data. Section 3 establishes posterior contraction rates for varying coefficients under data sketching. Section 4 demonstrates performance of the proposed approach with simulation examples and a forestry data analysis. Finally, Section 6 concludes the paper with an eye toward future extensions. All proofs of the theoretical results are placed in the Supplement.

2 Bayesian Compressed Varying Coefficient Models

We model each varying coefficient $w_j(\boldsymbol{u})$ in (1) as

$$w_j(\boldsymbol{u}) = \sum_{h=1}^{H} B_{jh}(\boldsymbol{u})\gamma_{jh}\ , \quad j = 1, ..., \tilde{P}\ , \tag{2}$$

where each $B_{jh}(\boldsymbol{u})$ is a basis function evaluated at an index $\boldsymbol{u}$ for $h = 1, ..., H$, and γ_{jh}'s are the corresponding basis coefficients. The distribution of these γ_{jh}'s yields a multivariate process with $\text{cov}(w_i(\boldsymbol{u}), w_j(\boldsymbol{u}')) = \boldsymbol{B}_i(\boldsymbol{u})^{\mathrm{T}}\text{cov}(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j)\boldsymbol{B}_j(\boldsymbol{u})$, where $\boldsymbol{B}_i(\boldsymbol{u})$ and $\boldsymbol{\gamma}_i$ are $H \times 1$ with elements $B_{ih}(\boldsymbol{u})$ and γ_{ih}, respectively, for $h = 1, \ldots, H$.

Appropriate choices for basis functions can produce appropriate classes of multivariate functional processes. A number of choices are available. For example, Biller and Fahrmeir (2001) and Huang et al. (2015) use splines to model the $B_{jh}(\boldsymbol{u})$'s and place Gaussian priors on the basis coefficients γ_{jh}. Li et al. (2015) propose a scale-mixture of multivariate normal distributions to shrink groups of basis coefficients towards zero. More recently, Bai et al. (2019) proposed using B-spline basis functions and multivariate spike-and-slab discrete mixture prior distributions on basis coefficients to aid functional variable selection. Other popular choices for basis functions include the wavelet basis (Vidakovic, 2009; Cressie and Wikle, 2015), radial basis (Bliznyuk et al., 2008) and locally bi-square (Cressie and Johannesson, 2008) or elliptical basis functions (Lemos and Sansó, 2009). Alternatively, a basis representation of $w_j(\boldsymbol{u})$ can be constructed by envisioning $w_j(\boldsymbol{u})$ as the projection of a Gaussian process $w_j(\boldsymbol{u})$ onto a set of reference points, or "knots", which yields predictive processes or sparse Gaussian processes and other variants (Snelson and Ghahramani, 2005; Banerjee et al., 2008; Guhaniyogi et al., 2013). More generally, each $w_j(\boldsymbol{u})$ can also be modelled using multi-resolution analogues to the aforesaid models to carefully capture global variations at the lower resolution and local variations at the higher resolutions (Katzfuss, 2017; Guhaniyogi and Sansó, 2018).

Let $\{y(\boldsymbol{u}_i), \boldsymbol{x}(\boldsymbol{u}_i)\}$ be observations at N index-points $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N\}$. Using (2) in (1) yields the Gaussian linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}\boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)\,. \tag{3}$$

where $\boldsymbol{y} = (y(\boldsymbol{u}_1), y(\boldsymbol{u}_2), \ldots, y(\boldsymbol{u}_N))^{\mathrm{T}}$ and $\boldsymbol{\epsilon} = (\epsilon(\boldsymbol{u}_1), \epsilon(\boldsymbol{u}_2), \ldots, \epsilon(\boldsymbol{u}_N))^{\mathrm{T}}$ are $N \times 1$ vectors of responses and errors, respectively, $\boldsymbol{X}$ is $N \times P$ with n-th row $\boldsymbol{x}(\boldsymbol{u}_n)^{\mathrm{T}}$, $\tilde{\boldsymbol{X}}$ is the $N \times N\tilde{P}$ block-diagonal matrix with (n,n)-th block $\tilde{\boldsymbol{x}}(\boldsymbol{u}_n)^{\mathrm{T}}$, $\boldsymbol{B} = (\boldsymbol{B}(\boldsymbol{u}_1)^{\mathrm{T}}, \ldots, \boldsymbol{B}(\boldsymbol{u}_N)^{\mathrm{T}})^{\mathrm{T}}$ is $N\tilde{P} \times H\tilde{P}$ with $\boldsymbol{B}(\boldsymbol{u}_n)$ a block-diagonal $\tilde{P} \times H\tilde{P}$ matrix whose j-th diagonal block is $(B_{j1}(\boldsymbol{u}_n), \ldots, B_{jH}(\boldsymbol{u}_n))$. The coefficient $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, ..., \boldsymbol{\gamma}_{\tilde{P}}^{\mathrm{T}})^{\mathrm{T}}$ is $H\tilde{P} \times 1$ with each $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jH})^{\mathrm{T}}$ being $H \times 1$. Bayesian methods for estimating (3) typically employ a multivariate normal prior (Biller and Fahrmeir, 2001; Huang et al., 2015) or its scale-mixture (discrete as well as continuous) variants (Li et al., 2015; Bai et al., 2019) on $\boldsymbol{\gamma}$.

While the basis functions project the coefficients into a low-dimensional space, working with (3) will be still be expensive for large N and will be impracticable for delivering full inference (with robust probabilistic uncertainty quantification) for data sets with $N \sim 10^5+$ on modest computing environments. Furthermore, as is well understood in linear regression, specifying a small number of basis functions in (3) can lead to substantial over-smoothing and, consequently, biased residual variance estimates in functional varying coefficient models(see, e.g., the discussion in Section 2.1 of Banerjee, 2017, including Figures 1 and 2 in the paper). Instead, we consider data compression or sketching using a random linear mapping to reduce the size of the data from N to M observations. For this, we use M one-dimensional linear mappings of the data encoded by an $M \times N$ compression matrix $\boldsymbol{\Phi}$ with $M << N$. This compression matrix is applied to $\boldsymbol{y}$, $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ to construct the $M \times 1$ compressed response vector $\boldsymbol{y}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\boldsymbol{y}$ and the matrices $\boldsymbol{X}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\tilde{\boldsymbol{X}}$. We will return to the specification of $\boldsymbol{\Phi}$, which, of course, will be crucial for relating the inference from the compressed data with the full model. For now assuming that we have fixed $\boldsymbol{\Phi}$, we construct a Bayesian hierarchical model for the compressed data

$$p(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{y_\Phi}, \boldsymbol{\Phi}) \propto p(\boldsymbol{\psi}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}) \times N(\boldsymbol{y_\Phi} \,|\, \boldsymbol{X_\Phi \beta} + \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}} \boldsymbol{B\gamma}, \sigma^2 I_M) \;, \tag{4}$$

where $\boldsymbol{\psi}$ denotes additional parameters specifying the prior distributions on either $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$. For example, a customary specification is

$$p(\boldsymbol{\psi}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{\tilde{P}} IG(\tau_i^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{\gamma} \,|\, \boldsymbol{0}, \boldsymbol{\Delta}) \;, \tag{5}$$

where $\boldsymbol{\psi} = \{\tau_1^2, ..., \tau_{\tilde{P}}^2\}$ and $\boldsymbol{\Delta}$ is $H\tilde{P} \times H\tilde{P}$ block-diagonal with j-th block given by $\tau_j^2 \boldsymbol{I}_H$, for $j = 1, ..., \tilde{P}$. While (5) is a convenient choice for empirical investigations due to conjugate full conditional distributions, our method applies broadly to any basis function and any discrete or continuous mixture of Gaussian priors on the basis coefficients. In applications where the associations among the latent regression slopes is of importance, one could, for instance, adopt $p(\boldsymbol{\psi}, \boldsymbol{\gamma}) = IW(\boldsymbol{\psi} \,|\, r, \boldsymbol{\Omega}) \times N(\boldsymbol{\gamma} \,|\, 0, \boldsymbol{\Delta}_{\boldsymbol{\psi}})$ with $\boldsymbol{\psi}$ as the $H\tilde{P} \times H\tilde{P}$ covariance matrix for $\boldsymbol{\gamma}$. Our current focus is not, however, on such multivariate models, so we do not discuss them further except to note that (4) accommodates such extensions.

The likelihood in (4) is different from that by applying $\boldsymbol{\Phi}$ to (3) because the error distribution in (4) is retained as the usual noise distribution without any effect of $\boldsymbol{\Phi}$. Hence, the model in (4) is a model analogous to (3) but applied to the *new* compressed data set $\{\boldsymbol{y_\Phi}, \boldsymbol{X_\Phi}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\}$. Working with a $\boldsymbol{\Phi}$-transformed model (3), where the distribution of the noise will be transformed according $\boldsymbol{\Phi\epsilon}$, will not deliver the computational benefits, and is somewhat detrimental to the cause of data confidentiality (as in that case, the analyst need to know $\boldsymbol{\Phi}$) that are provided by (4).

For specifying $\boldsymbol{\Phi}$ we pursue the idea of data oblivious Gaussian sketching (Sarlos, 2006), where we draw the elements of $\boldsymbol{\Phi} = (\Phi_{ij})$ independently from $\mathrm{N}(0, 1/N)$ and fix them.

The dominant computational operations for obtaining the sketched data using Gaussian sketches is $O(MN^2\tilde{P})$. While alternative computationally efficient data oblivious options such as the Hadamard sketch (Ailon and Chazelle, 2009) and the Clarkson - Woodruff sketch (Clarkson and Woodruff, 2017) are available for $\boldsymbol{\Phi}$, it is less pertinent in Bayesian settings since computation time of (4) far exceeds that for the sketching matrix. The compressed data serves as a surrogate for the Bayesian regression analysis with varying coefficients. Since the number of compressed records is much smaller than the number of records in the uncompressed data matrix, model fitting becomes computationally efficient and economical in terms of storage as well as the number of floating point operations (flops). Importantly, original data are not recoverable from the compressed data, and the compressed data effectively reveal no more information than would be revealed by a completely new sample (Zhou et al., 2008). In fact, the original uncompressed data does not need to be stored or accessed at any stage in the course of the analysis.

2.1 Posterior Computations & Predictive Inference

In what follows, we discuss efficient computation offered by the data sketching framework. With prior distributions on parameters specified as in (5), posterior computation requires drawing Markov chain Monte Carlo (MCMC) samples sequentially from the full conditional posterior distributions of $\boldsymbol{\gamma}|-$, $\boldsymbol{\beta}|-$, $\sigma^2|-$ and $\tau_j^2|-$, $j = 1, \ldots, \tilde{P}$. To this end, $\sigma^2|- \sim IG(a_\sigma + M/2, b_\sigma + ||\boldsymbol{y}_{\boldsymbol{\Phi}} - \boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\gamma}||^2/2)$, $\tau_j^2|- \sim IG(a_\tau + H/2, b_\tau + ||\boldsymbol{\gamma}_j||^2/2)$ and $\boldsymbol{\beta}|- \sim N\left((\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{\Phi}}/\sigma^2 + \boldsymbol{I})^{-1}\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}(\boldsymbol{y}_{\boldsymbol{\Phi}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\gamma})/\sigma^2, (\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{\Phi}}/\sigma^2 + \boldsymbol{I})^{-1}\right)$ do not present any computational obstacles. The main computational bottleneck lies with $\boldsymbol{\gamma}|-$,

$$N\left(\left(\frac{\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}}{\sigma^2} + \boldsymbol{\Delta}^{-1}\right)^{-1}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\frac{(\boldsymbol{y}_{\boldsymbol{\Phi}} - \boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta})}{\sigma^2}, (\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}/\sigma^2 + \boldsymbol{\Delta}^{-1})^{-1}\right). \quad (6)$$

Efficient sampling of $\boldsymbol{\gamma}$ relies upon the Cholesky decomposition of $\left(\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}/\sigma^2+\boldsymbol{\Delta}^{-1}\right)$ and solves triangular linear systems to draw a sample from (6). While numerically robust for small to moderately large H, computing and storing the Cholesky factor of this matrix involves $O((H\tilde{P})^3)$ and $O((H\tilde{P})^2)$ floating point operations, respectively (Golub and Van Loan, 2012). This produces computational and memory bottlenecks for a large number of basis functions, which is required to estimate the unknown functional coefficients with sufficient local variation.

To achieve computational efficiency, we adapt a recent algorithm proposed in Bhattacharya et al. (2016) (in the context of ordinary linear regression with uncompressed data and small sample size) to our setting: (i) draw $\tilde{\gamma}_1\sim N(\mathbf{0},\boldsymbol{\Delta})$ and $\tilde{\gamma}_2\sim N(\mathbf{0},\boldsymbol{I}_M)$; (ii) set $\tilde{\gamma}_3=\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\tilde{\gamma}_1/\sigma+\tilde{\gamma}_2$; (iii) solve $(\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}/\sigma^2+\boldsymbol{I}_M)\tilde{\gamma}_4=((\boldsymbol{y}_{\boldsymbol{\Phi}}-\boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta})/\sigma-\tilde{\gamma}_3)$; and (iv) set $\tilde{\gamma}_5=\tilde{\gamma}_1+\boldsymbol{\Delta}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\gamma}_4/\sigma$. The resulting $\tilde{\gamma}_5$ is a draw from the full conditional posterior distribution of $\boldsymbol{\gamma}$. The computation is dominated by step (iii), which comprises $O(M^3+M^2H\tilde{P})$. Finally, note that when basis functions involve parameters, they are updated using Metropolis-Hastings steps since no closed form full conditionals are generally available for them.

Predictive inference on $y(\boldsymbol{u}_0)$ will proceed from the posterior predictive distribution

$$\mathbb{E}[p(y(\boldsymbol{u}_0)\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2)]=\int p(y(\boldsymbol{u}_0)\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2)p(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\Phi})d\boldsymbol{\beta}d\boldsymbol{\gamma}d\sigma^2\ , \qquad (7)$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to the posterior distribution in (4). This is easily achieved by drawing $y(\boldsymbol{u}_0)^{(l)}\sim N(\sum_{p=1}^{P}x_p(\boldsymbol{u}_0)\beta_p^{(l)}+\sum_{j=1}^{\tilde{P}}\tilde{x}_j(\boldsymbol{u}_0)w_j(\boldsymbol{u}_0)^{(l)},\sigma^{2(l)})$ for each posterior sample $\{\boldsymbol{\beta}^{(l)},\boldsymbol{\gamma}^{(l)},\sigma^{2(l)}\}$ drawn from (4), where $w_j(\boldsymbol{u}_0)^{(l)}$ is obtained from $\boldsymbol{\gamma}^{(l)}$ using (2) and $l=1,2,\ldots,L$ indexes the L (post-convergence) posterior samples. The next section offers theoretical results related to the large sample consistency of the posterior distribution from the compressed varying coefficients model (4) and the posterior predictive distribution in (7) with respect to the probability law for the uncompressed oracle model in (1).

3 Posterior contraction from data sketching

3.1 Definitions and Notations

This section proves the posterior contraction properties of varying coefficients under the proposed framework. In what follows, we add a subscript N to the compressed response vector $\boldsymbol{y}_{\boldsymbol{\Phi},N}$, compressed predictor matrix $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}$, dimension of the compression matrix M_N and the number of basis functions H_N to indicate that all of them increase with the sample size N. Naturally, the dimension of the basis coefficient vector $\boldsymbol{\gamma}$ and the compression matrix $\boldsymbol{\Phi}$ are also functions of N, though we keep this dependence implicit. Since we do not assume a functional variable selection framework, we keep P fixed throughout, and not a function of N. We assume that $\boldsymbol{u}_1, ..., \boldsymbol{u}_N$ follow i.i.d. distribution G on $\mathcal{D}$ with G having a Lebesgue density g, which is bounded away from zero and infinity uniformly over $\mathcal{D}$. The true regression function is also given by (1), with the true varying coefficients $w_1^*(\boldsymbol{u}), ..., w_{\tilde{P}}^*(\boldsymbol{u})$ belonging to the class of functions

$$\mathcal{F}_\xi(\mathcal{D}) = \{f : f \in L_2(\mathcal{D}) \cap \mathcal{C}^\xi(\mathcal{D}), E_{\mathcal{U}}[|f|] < \infty\}, \tag{8}$$

where $L_2(\mathcal{D})$ is the set of all square integrable functions on $\mathcal{D}$, $\mathcal{C}^\xi(\mathcal{D})$ is the class of at least ξ-times continuously differentiable functions in $\mathcal{D}$ and $E_{\mathcal{U}}$ denotes the expectation under the density of g. The probability and expectation under the true data generating model are denoted by P^* and E^*, respectively. For algebraic simplicity, we make a few simplifying assumptions in the model. To be more specific, we assume that $\boldsymbol{\beta} = \boldsymbol{0}$ and $\sigma^2 = \sigma^{*2}$ is known and fixed at 1. The first assumption is mild since P does not vary with

N and we do not consider variable selection. The second assumption is also customary in asymptotic studies (Vaart and Zanten, 2011). Furthermore, the theoretical results obtained by assuming σ^2 as a fixed value is equivalent to those obtained by assigning a prior with a bounded support on σ^2 (Van der Vaart et al., 2009).

For a vector $\boldsymbol{v} = (v_1, ..., v_N)^{\mathrm{T}}$, we let $|| \cdot ||_1, || \cdot ||_2$ and $|| \cdot ||_\infty$ denote the L_1, L_2 and L_∞ norms, respectively, defined as $||\boldsymbol{v}||_2 = (\sum_{n=1}^N v_n^2)^{1/2}$, $||\boldsymbol{v}||_1 = \sum_{n=1}^N |v_n|$ and $||\boldsymbol{v}||_\infty = \max_{n=1,..,N} |v_n|$, respectively. The number of nonzero elements in a vector is given by $|| \cdot ||_0$. In the case of a square integrable function $f(\boldsymbol{u})$ on $\mathcal{D}$, we denote the integrated L_2-norm of f by $||f||_2 = \left(\int_{\mathcal{D}} f(\boldsymbol{u})^2 g(\boldsymbol{u}) d\boldsymbol{u}\right)^{1/2}$ and the sup-norm of f by $||f||_\infty = \sup_{\boldsymbol{u} \in \mathcal{D}} |f(\boldsymbol{u})|$. Thus $||\cdot||_\infty$ and $||\cdot||_2$ are used both for vectors and functions, and they should be interpreted based on the context. Finally, $e_{\min}(\boldsymbol{A})$ and $e_{\max}(\boldsymbol{A})$, respectively, represent the minimum and maximum eigenvalues of the square matrix $\boldsymbol{A}$. The Frobenius norm of the matrix $\boldsymbol{A}$ is given by $||\boldsymbol{A}||_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$. For two nonnegative sequences $\{a_N\}$ and $\{b_N\}$, we write $a_N \asymp b_N$ to denote $0 < \liminf_{N \to \infty} a_N/b_N \leq \limsup_{N \to \infty} a_N/b_N < \infty$. If $\lim_{N \to \infty} a_N/b_N = 0$, we write $a_N = o(b_N)$ or $a_N \prec b_N$. We use $a_N \lesssim b_N$ or $a_N = O(b_N)$ to denote that for sufficiently large N, there exists a constant $C > 0$ independent of N such that $a_N \leq C b_N$.

3.2 Assumption, Framework and Main Results

For simplicity, we assume $\boldsymbol{\Delta} = \boldsymbol{I}$ and that the random covariates $x_p(\boldsymbol{u})$, $p = 1, ..., P$ follow distributions which are independent of the distribution of the idiosyncratic error ϵ. We now state the following assumptions on the basis functions, H_N, M_N, covariates and the sketching or compression matrix.

(A) For any $w_j^*(\boldsymbol{u}) \in \mathcal{F}_\xi(\mathcal{D})$, there exists $\boldsymbol{\gamma}_j^*$ such that

$$||w_j^* - \boldsymbol{B}_j^{\mathrm{T}}\boldsymbol{\gamma}_j^*||_\infty = \sup_{\boldsymbol{u} \in \mathcal{D}} |w_j^*(\boldsymbol{u}) - \sum_{h=1}^{H_N} B_{jh}(\boldsymbol{u})\gamma_{jh}^*| = O(H_N^{-\xi}),$$

for $j = 1, ..., \tilde{P}$, and $||\boldsymbol{\gamma}^*||_2^2 \prec M_N^{d/(d+2\xi)}$.

(B) N, M_N, H_N satisfy $M_N = o(N)$ and $H_N \asymp M_N^{1/(2\xi+d)}$.

(C) $||\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} - \boldsymbol{I}_{M_N}||_F \leq C'\sqrt{M_N/N}$, for some constant $C' > 0$, for all large N.

(D) The random covariate $x_p(\boldsymbol{u})$ are uniformly bounded for all $\boldsymbol{u} \in \mathcal{D}$, and w.l.g., $|x_p(\boldsymbol{u})| \leq 1$, for all $p = 1, ..., P$ and for all $\boldsymbol{u} \in \mathcal{D}$.

(E) There exists a sequence κ_N such that $||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\alpha}||^2 \asymp \kappa_N||\tilde{\boldsymbol{X}}_N\boldsymbol{\alpha}||^2$, such that $1 \prec N\kappa_N \prec M_N$ for any vector $\boldsymbol{\alpha} \in \mathbb{R}^{N\tilde{P}}$.

Assumption (A) holds for orthogonal Legendre polynomials, Fourier series, B-splines and wavelets (Shen and Ghosal, 2015). Assumption (B) provides an upper bound on the growth of M_N and H_N as a function of N. Assumption (C) is a mild assumption based on the theory of random matrices and occurs with probability at least $1 - e^{-C''M_N}$ when $\boldsymbol{\Phi}$ is constructed using the Gaussian sketching for a constant $C'' > 0$ (see Lemma 5.36 and Remark 5.40 of Vershynin (2010)). Assumption (D) is a technical condition customarily used in functional regression analysis (Bai et al., 2019). Finally, Assumption (E) characterizes the class of feasible compression matrices, roughly explaining how the linear structure of the columns of the original predictor matrix is related to that of the compressed predictor matrix. Such an assumption is reasonable for the set of random compression matrices for a sequence κ_N depending on N, M_N and $\tilde{P}$ (Ahfock et al., 2017).

Let $\boldsymbol{w}(\boldsymbol{u}) = (w_1(\boldsymbol{u}), ..., w_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ and $\boldsymbol{w}^*(\boldsymbol{u}) = (w_1^*(\boldsymbol{u}), ..., w_{\tilde{P}}^*(\boldsymbol{u}))^{\mathrm{T}}$ be the $\tilde{P}$-dimensional fitted and true varying coefficients. Let $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 = \sum_{j=1}^{\tilde{P}} \|w_j - w_j^*\|_2$ denote the sum of integrated L_2 distances between the true and the fitted varying coefficients. Define the set $\mathcal{C}_N = \left\{\boldsymbol{w} : ||\boldsymbol{w} - \boldsymbol{w}^*||_2 > \tilde{C}\theta_N\right\}$, for some constant $\tilde{C}$ and some sequence $\theta_N \to 0$ and $M_N\theta_N^2 \to \infty$. Further suppose $\pi_N(\cdot)$ and $\Pi_N(\cdot)$ are the prior and posterior densities of

w with N observations, respectively. From equation (2), the prior distribution on w is governed by the prior distribution on $\boldsymbol{\gamma}$, so that the posterior probability of $\mathcal{C}_N$ can be expressed as,

$$\Pi_N(\mathcal{C}_N|\boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}) = \frac{\int_{\mathcal{C}_N} f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})\pi_N(\boldsymbol{\gamma})}{\int f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})\pi_N(\boldsymbol{\gamma})},$$

where $f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})$ is the joint density of $\boldsymbol{y}_{\boldsymbol{\Phi},N}$ under model (4). We begin with the following important result from the random matrix theory.

Lemma 1. *Consider the $M_N \times N$ compression matrix $\boldsymbol{\Phi}$ with each entry being drawn independently from $N(0, 1/N)$. Then, almost surely*

$$(\sqrt{N} - \sqrt{M_N} - o(\sqrt{N}))^2/N \leq e_{\min}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}) \leq e_{\max}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}) \leq (\sqrt{N} + \sqrt{M_N} + o(\sqrt{N}))^2/N, \tag{9}$$

when both $M_N, N \to \infty$.

Proof. This is a consequence of Theorem 5.31 and Corollary 5.35 of Vershynin (2010). □

The inequalities in (9) is used to derive the following two results, which we present as Lemma 2 and 3.

Lemma 2. *Let P^* denote the true probability distribution of $\boldsymbol{y}_N$ and $f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)$ denotes the density of $\boldsymbol{y}_{\boldsymbol{\Phi},N}$ (omitting explicit dependence on $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}$) under the true data generating model. Define*

$$\mathcal{A}_N = \left\{\boldsymbol{y} : \int \left\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\right\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma} \leq \exp(-CM_N\theta_N^2)\right\}. \tag{10}$$

Then $P^*(\mathcal{A}_N) \to 0$ as $M_N, N \to \infty$ for any constant $C > 0$.

Proof. See Section S1 in the Supplement. □

Lemma 3. *Let $\boldsymbol{\gamma}^*$ be any fixed vector in the support of $\boldsymbol{\gamma}$ and let $\mathcal{B}_N = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 \leq C_{2w}\theta_N H_N^{1/2}\}$ for some constant $C_{2w} > 0$. Then there exists a sequence ζ_N of random variables depending on $\{\boldsymbol{y}_{\boldsymbol{\Phi},N}, \boldsymbol{X}_{\boldsymbol{\Phi},N}\}$ and taking values in $(0,1)$ such that*

$$\mathbb{E}^*(\zeta_N) \lesssim \exp(-M_N\theta_N^2) \text{ and } \sup_{\boldsymbol{\gamma}\in\mathcal{B}_N^c} \mathbb{E}_{\boldsymbol{\gamma}}(1-\zeta_N) \lesssim \exp(-M_N\theta_N^2), \tag{11}$$

where $\mathbb{E}_{\boldsymbol{\gamma}}$ and $\mathbb{E}^*$ denote the expectations under the distributions $f(\cdot\,|\,\boldsymbol{\gamma})$ and $f^*(\cdot\,|\,\boldsymbol{\gamma}^*)$, respectively.

Proof. See Section S2 in the Supplement. □

We use the above results to establish the posterior contraction result for the proposed model.

Theorem 1. *Under Assumptions (A)-(E), our proposed model (4) satisfies*

$$\max_{j=1,\ldots,\tilde{P}} \sup_{w_j^*\in\mathcal{F}_\xi(\mathcal{D})} \mathbb{E}^*\Pi_N(\mathcal{C}_N \,|\, \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}) \to 0, \text{ as } N, M_N \to \infty,$$

with the posterior contraction rate $\theta_N \asymp M_N^{-\xi/(2\xi+d)}$.

Proof. See Section S3 in the Supplement. □

Since $\theta_N \to 0$ as $N \to \infty$, the model consistently estimates the true varying coefficients under the integrated L_2-norm. Further, data compression decreases the effective sample size from N to M_N, hence, the contraction rate θ_N obtained in Theorem 1 is optimal and adaptive to the smoothness of the true varying coefficients. Our next theorem justifies the two-stage prediction strategy described in Section 2.1.

Theorem 2. *For any input $\boldsymbol{u}_0$ drawn randomly with the density g and corresponding predictors $\tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0)$, let f_u be the predictive density $p(y(\boldsymbol{u}_0)\,|\,\tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0), w(\boldsymbol{u}_0))$ derived from (1) without data compression. Let f^* be the true data generating model (i.e.,* *(1) with $\boldsymbol{w}(\boldsymbol{u}_0)$ fixed at $\boldsymbol{w}^*(\boldsymbol{u}_0)$). Given $\boldsymbol{u}_0$ and $\tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0)$, define $h(f_u, f^*) = \int(\sqrt{f_u} - \sqrt{f^*})^2$ as the Hellinger distance between the densities f_u and f^*. Then*

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{U}}[h(f_u, f^*) \,|\, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{y}_{\boldsymbol{\Phi},N}] \to 0, \;\; \text{as } N, M_N \to \infty, \tag{12}$$

where $\mathbb{E}_{\mathcal{U}}$, $\mathbb{E}$ and $\mathbb{E}^*$ stand for expectations with respect to the density g, the posterior density $\Pi_N(\cdot|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{y}_{\boldsymbol{\Phi},N})$ and the true data generating distribution, respectively.

Proof. See Section S4 in the Supplement. □

The theorem states that the predictive density of the VCM model in (1) is arbitrarily close to the true predictive density even when we plug-in inference on parameters from (4).

4 Simulation Results

4.1 Inferential performance

We empirically validate our proposed approach using (4) for $d = 2$, i.e., for the spatially varying coefficient models. The approach, henceforth abbreviated as *geoS*, is compared with the uncompressed model (3) on some simulated data in terms of inferential performance and computational efficiency. We simulate data by using a fixed set of spatial locations $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ that were drawn uniformly over the domain $\mathcal{D} = [0,1]\times[0,1]$. We set $\tilde{P} = P = 3$ and assume $\boldsymbol{\beta} = 0$, i.e., all predictors have purely space-varying coefficients. We set $\tilde{x}_1(\boldsymbol{u}_i) = 1$, for all $i = 1, \ldots, N$, while the values of $\tilde{x}_j(\boldsymbol{u}_1), \ldots, \tilde{x}_j(\boldsymbol{u}_N)$ for $j = 2, 3$ were set to independently values from $N(0, 1)$. For each $n = 1, \ldots, N$, the response $y(\boldsymbol{u}_n)$ is drawn independently from $N(w_1^*(\boldsymbol{u}_n) + w_2^*(\boldsymbol{u}_n)\tilde{x}_2(\boldsymbol{u}_n) + w_3^*(\boldsymbol{u}_n)\tilde{x}_3(\boldsymbol{u}_n), \sigma^{*2})$ following (3), where σ^{*2} is set to be 0.1. The true space-varying coefficients ($w_j^*(\boldsymbol{u})$s) are simulated from a Gaussian process with mean 0 and covariance kernel $C(\cdot, \cdot; \theta_j)$, i.e., $(w_j^*(\boldsymbol{u}_1), ..., w_j^*(\boldsymbol{u}_N))^{\mathrm{T}}$

is drawn from $N(0, C^*(\theta_j))$, for each $j = 1, \ldots, \tilde{P}$, where $C^*(\theta_j)$ is an $N \times N$ matrix with the (n, n')th element $C(\boldsymbol{u}_n, \boldsymbol{u}_{n'}; \theta_j)$. We set the covariance kernel $C(\cdot, \cdot; \theta_j)$ to be the exponential covariance function given by

$$C(\boldsymbol{u}, \boldsymbol{u}'; \theta_j) = \delta_j^2 \exp\left\{-\frac{1}{2}\left(\frac{||\boldsymbol{u} - \boldsymbol{u}'||}{\phi_j}\right)\right\}, \quad j = 1, 2, 3, \tag{13}$$

with the true values of $\delta_1^2, \delta_2^2, \delta_3^2$ set to $1, 0.8, 1.1$, respectively. We fix the true values of ϕ_1, ϕ_2, ϕ_3 at $1, 1.25, 2$, respectively.

While fitting *geoS* and its uncompressed analogue (3), the varying coefficients are modeled through the linear combination of H basis functions as in (2), where these basis functions are chosen as the tensor-product of B-spline bases of order $q = 4$ (Shen and Ghosal, 2015). More specifically, for $\boldsymbol{u} = (u^{(1)}, u^{(2)})$, the j-th varying coefficient is modeled as

$$w_j(\boldsymbol{u}) = \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} B_{jh_1}^{(1)}(u^{(1)}) B_{jh_2}^{(2)}(u^{(2)}) \gamma_{jh_1h_2} \,, \tag{14}$$

where the marginal B-splines $B_{jh_1}^{(1)}$, $B_{jh_2}^{(2)}$ are defined on sets of H_1 and H_2 knots, respectively. The knots are chosen to be equally-spaced so the entire set of $H = H_1 H_2$ knots is uniformly spaced over the domain $\mathcal{D}$. We complete the hierarchical specification by assigning independent $IG(2, 0.1)$ priors (mean 0.1 with infinite variance) for σ^2 and τ_j^2 for each $j = 1, \ldots, P$.

We implemented our models in the `R` statistical computing environment on a Dell XPS 13 PC with Intel Core i7-8550U CPU @ 4.00GHz processors at 16 GB of RAM. For each of our simulation data sets we ran a single-threaded MCMC chain for 5000 iterations. Posterior inference was based upon 2000 samples retained after adequate convergence was diagnosed using Monte Carlo standard errors and effective sample sizes (ESS) using the `mcmcse` package in `R`. Source codes for these experiments are available from Redacted in blinded version.

Table 1: Results for simulation cases 1 & 2 for the compressed *geoS* and uncompressed models. Mean Squared Error (MSE), length and coverage of 95% CI for the spatially varying coefficients. We also present mean squared prediction error (MSPE), coverage and length of 95% predictive intervals for the competing models. Computational efficiency for the *geoS* with the uncompressed data model is also recorded.

	$N = 5000$, $H = 225$		$N = 10000$, $H = 256$	
	(geoS) $M = 700$	*Uncompressed*	*(geoS)* $M = 1000$	*Uncompressed*
MSE (SVC)	0.0474	0.0168	0.0429	0.0178
95% CI length	0.8368	0.6182	0.7222	0.5531
95% CI Coverage	0.9448	0.9322	0.9153	0.9026
MSPE	0.2574	0.1833	0.2283	0.1605
95% PI length	1.9717	1.5168	1.8613	1.5148
95% PI coverage	0.936	0.925	0.954	0.930
Computation efficiency	2.2050	0.8079	0.9755	0.4356

Table 1 summarizes the estimates of varying coefficients and the predictive performance for *geoS* in comparison to the uncompressed model. We applied these models to data generated with $N = 5000$ (case 1) and $N = 10000$ (case 2). For both cases the compressed dimension is taken to be $M \approx 10\sqrt{N}$ which seems to be effective from empirical considerations in our simulations. We provide further empirical justification for this choice in Section 4.2. Our *geoS* approach compresses the sample sizes to $M = 700$ and $M = 1000$ in cases 1 and 2, respectively. The number of fitted basis functions in cases 1 & 2 are $H = 225, 256$, respectively.

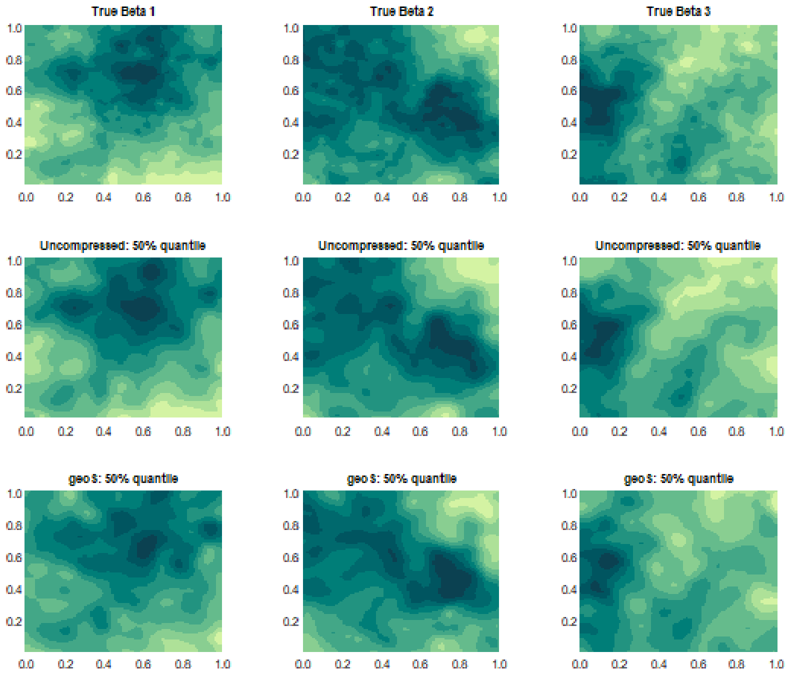


Figure 1: Simulation case 1: $(N, H) = (5000, 225)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0, 1] \times [0, 1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(\boldsymbol{u})$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

Figures 1 and 2 present the estimated varying coefficients by *geoS* and the uncompressed data model for cases 1 and 2, respectively. These figures reveal point estimates that are substantively similar to those from *geoS* and the uncompressed model. The mean squared error of estimating varying coefficients, defined as $\sum_{j=1}^{3}\sum_{n=1}^{N}(\widehat{w}_j(\boldsymbol{u}_n) - w_j^*(\boldsymbol{u}_n))^2/(3N)$ (where $\widehat{w}_j(\boldsymbol{u}_n)$ is the posterior median of $w_j(\boldsymbol{u}_n)$), also confirms very similar point estimates offered by the compressed and uncompressed models (see Table 1). Further, *geoS* offers close to nominal coverage for 95% credible intervals for varying coefficients, with little wider credible intervals compared to uncompressed data model. This can be explained by the smaller sample size for the *geoS* model, though the difference turns out to be minimal. We also carry out predictive inference using *geoS* (Section 2.1). Table 1 presents

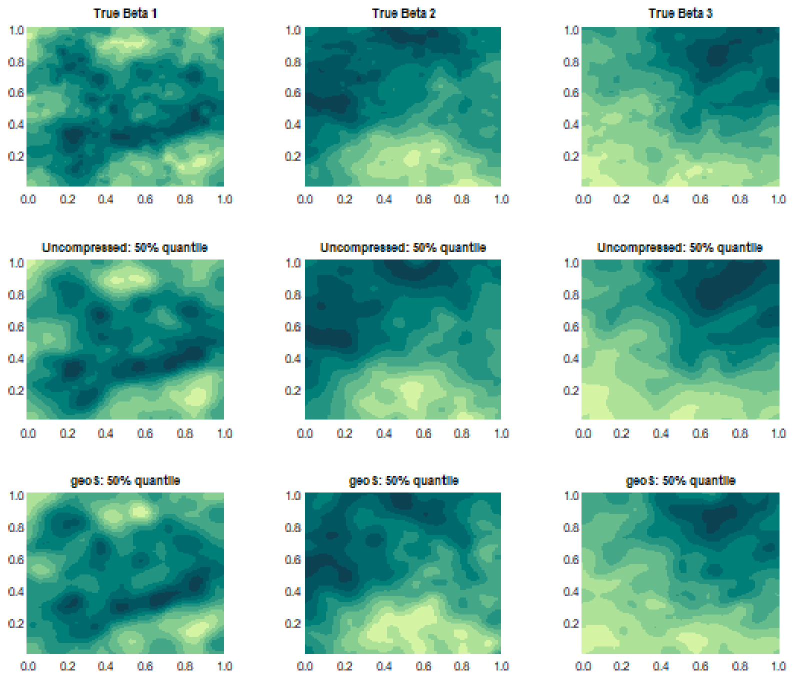
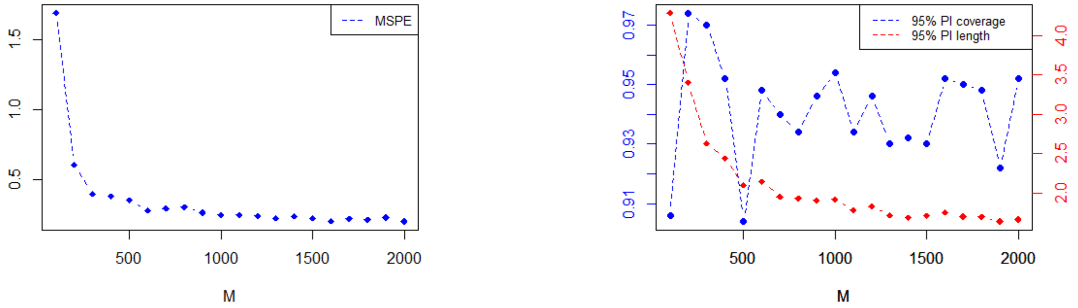


Figure 2: Simulation case 2: $(N, H) = (10000, 256)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0, 1] \times [0, 1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(\boldsymbol{u})$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

mean squared predictive error (MSPE), average length and coverage for the 95% predictive intervals, based on $N^* = 500$ out of the sample observations. We find *geoS* delivers posterior predictive estimates and predictive coverage that are very consistent with the uncompressed model, perhaps with marginally wider predictive intervals than those without compression. Finally, the computational efficiency of both models are computed based on the metric $\log_2(ESS/\text{Computation Time})$, where ESS denotes the effective sample size averaged over the MCMC samples of all parameters. We find *geoS* is almost 270% and 223% more efficient than the uncompressed model for $N = 5,000$ and $N = 10,000$, respectively, while delivering substantively consistent inference on the spatial effects.



(a)

(b)

Figure 3: (a) MSPE, (b) 95% predictive interval coverage and length for different choices of M

4.2 Choice of the dimension of the compression matrix M

We present investigations into the choice of the appropriate compression matrix size M. For simulated data with sample size $N = 10000$, we ran our model for different values of $M = k\sqrt{N}$, $k = 1, \ldots, 20$. Figure 3 shows the variations in point-wise and interval prediction reflected in the $MSPE$ and 95% predicted interval coverage and length, respectively. Unsurprisingly, as M increases the MSPE drops with a diminished rate of decline until the $k \sim 10$. In terms of interval prediction, predictive coverage seems to oscillate within the narrow interval $(0.9, 0.97)$ for all values of M, but the length of the predictive interval improves as M increases and starts to stabilize at around $k \sim 10$. We observe that the choice of $M \sim 10\sqrt{N}$ leads to good performance across various simulations and real data analysis.

5 Vegetation Data Analysis

We implement *geoS* to analyze vegetation data gathered through the Moderate Resolution Imaging Spectroradiometer (MODIS), which resides aboard the Terra and Aqua platforms on NASA spacecrafts. MODIS vegetation indices, produced on 16-day intervals and at multiple spatial resolutions, provide consistent information on the spatial distribution of vegetation canopy greenness, a composite property of leaf area, chlorophyll and canopy structure. The variable of interest will be the Normalized Difference Vegetation Index (NDVI), which quantifies the relative vegetation density for each pixel in a satellite image, by measuring the difference between the reflection in the near-infrared spectrum (NIR) and the red light reflection (RED): $NDVI = \frac{NIR-RED}{NIR+RED}$. High NDVI values, ranging between 0.6 and 0.9 indicate high density of green leaves and healthy vegetation, whereas low values, 0.1 or below, correspond to low or absence of vegetation as in the case of urbanized areas. When analyzed over different locations, NDVI can reveal changes in vegetation due to human activities such as deforestation and natural phenomena such as wild fires and floods.

Our analysis will be focused on geographical data that was mapped on a sinusoidal (SIN) projected grid, located on the western coast of the United States, more precisely zone *h08v05*, between $30°N$ to $40°N$ latitude and $104°W$ to $130°W$ longitude (see Figure 4(a)). The data set, which was downloaded using the `R` package MODIS, comprises $133,000$ observed locations where the response was measured through the MODIS tool over a 16-day period in April, 2016. We retained $N = 113,000$ observations (randomly chosen) for model fitting and held out the rest for prediction. In order to fit (1), we set $y(s_n)$ to be the transformed NDVI ($\log(NDVI) + 1$), $P = \tilde{P} = 2$ and consider the $P \times 1$ vector of predictors that includes an intercept and a binary index of urban area, both with

Table 2: Median and 95% credible interval of β_1, β_2 for geoS and its uncompressed analogue are presented for the Vegetation data analysis. We also present MSPE, coverage and length of 95% predictive intervals for the competing models. Computational efficiency for the two competing models are also provided.

	(geoS) $M = 2300$	*Uncompressed*
β_1	0.222 (0.212, 0.230)	0.229 (0.219, 0.237)
β_2	-0.060 (-0.074, -0.047)	-0.071 (-0.082, -0.059)
$MSPE$	0.00327	0.00276
95% PI length	0.23445	0.22136
95% PI coverage	0.95250	0.95411
Computation efficiency	3.5424	0.46901

fixed effects and spatially varying coefficients, i.e., $\boldsymbol{x}(\boldsymbol{u}_n) = \tilde{\boldsymbol{x}}(\boldsymbol{u}_n) = (1,\ x_2(\boldsymbol{u}_n))^{\mathrm{T}}$, with $x_2(\boldsymbol{u}_n) = \mathbb{1}_U(\boldsymbol{u}_n)$, where U denotes an urban area.

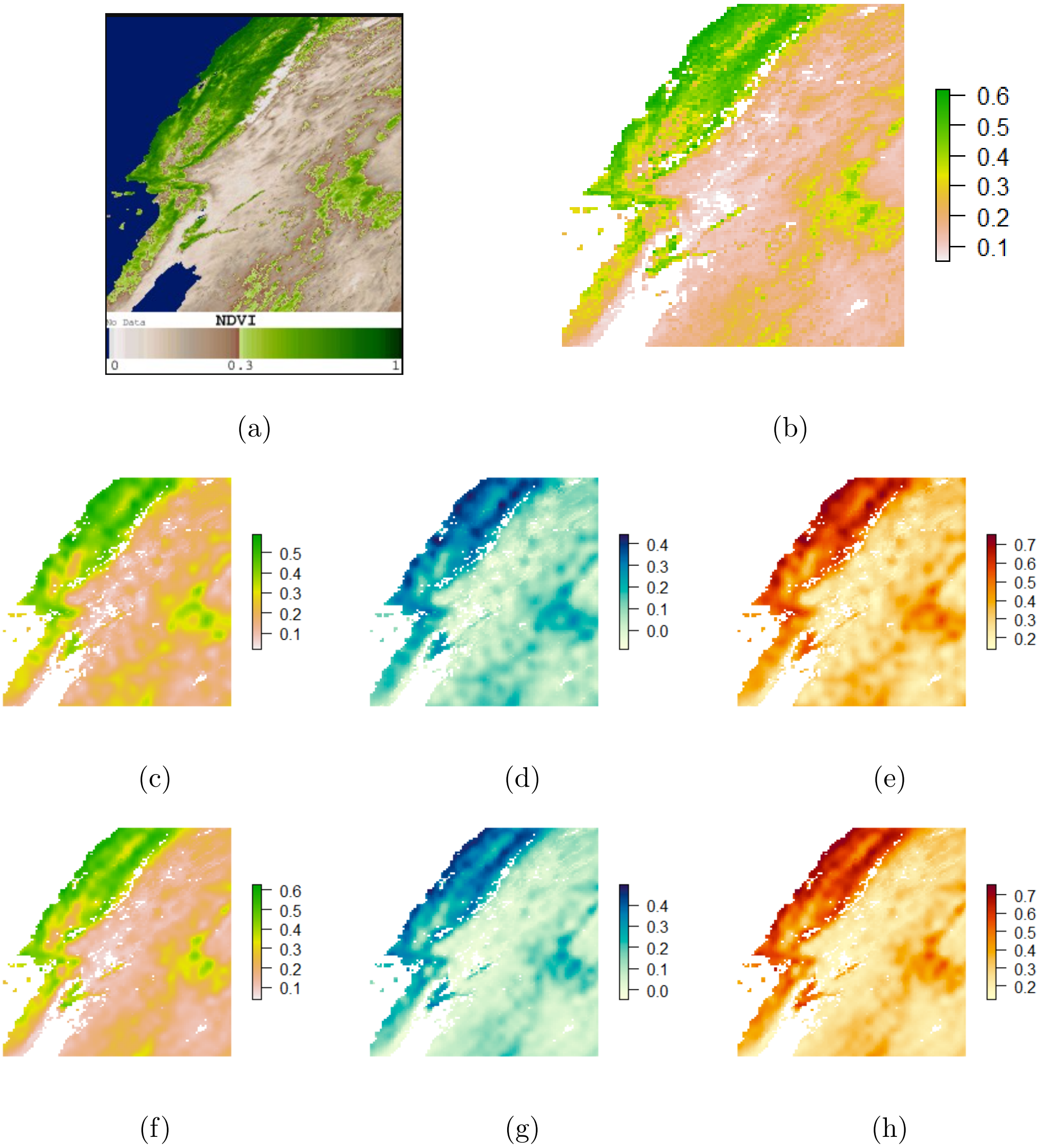
As in Section 4, we fit *geoS* with $M \sim 10\sqrt{N} = 2300$ and its uncompressed counterpart (3), by modeling the varying coefficients through a linear combination of basis functions constructed using the tensor-product of B-splines of order $q = 4$ as in (14). We set $H = H_1H_2 = 39^2 = 1521$ uniformly distributed knots over the domain $\mathcal{D}$, which results in $HP = 3042$ basis coefficients γ_{jh} that are estimated. Specification of priors are identical to the simulation studies for σ^2, and τ_j^2, while β_j is assigned a flat prior for $j = 1, \dots, P$.

We ran an MCMC chain for 5000 iterations and retained 2000 samples for posterior inference after adequate convergence was diagnosed. The posterior mean of β_1 and β_2, along with their estimated 95% credible intervals corresponding to *geoS* and the uncompressed model are presented in Table 2. Additionally, Table 2 offers predictive inference from both competitors based on $N^* = 20,000$ test observations. According to both models there is a global pattern of relatively low vegetation density for areas with positive urban index as the estimated slope coefficient β_2 is negative in the compressed *geoS* and in the uncompressed models. In terms of point prediction and quantification of predictive uncertainty, the two competitors offer practically indistinguishable results, as revealed by Table 2.

Further, Figure 4 shows that the 2.5%, 50% and 97.5% quantiles for the posterior predictive distribution are almost identical for the two competitors across the spatial domain, with the exception of neighborhoods around locations having lower NDVI values. Notably, *geoS* offers nominal coverage for 95% prediction intervals, even with a significant reduction in the sample size from $N = 113,000$ to $M = 2300$. Data sketching to such a scale considerably reduces the computation time, leading to a much higher computation efficiency of *geoS* in comparison with its uncompressed analogue.

6 Summary

We have developed Bayesian sketching for functional response and predictor variables using varying coefficient regression models. The method achieves dimension reduction by compressing the data using a random linear transformation. The approach is different from the prevalent methods for large functional data in that no new models or algorithms need to be developed since those available for existing varying coefficient regression models can be directly applied to the compressed data. We establish attractive concentration properties of the posterior and posterior predictive distributions and empirically demonstrate the effectiveness of this method for analyzing large functional data sets. Access to the values of the response and predictors in the full data are not required at stage of inference, which preserves data confidentiality should that be of concern in the application.



(a) (b)

(c) (d) (e)

(f) (g) (h)

Figure 4: Colored NDVI images of western United States (zone h08v05). (a) Satellite image: MODIS/Terra Vegetation Indices 16-Day L3 Global 1 km SIN Grid - 2016.04.06 to 2016.04.21; (b) True NDVI surface (raw data). Figures (c), (d) & (e) present NVDI predicted 50%, 2.5% and 97.5% quantiles for the *geoS* model. Figures (f), (g) & (h) present NVDI Predicted 50%, 2.5% and 97.5% quantiles for the uncompressed model.

7 Acknowledgments

Redacted in blinded version.

References

Ahfock, D., Astle, W. J., and Richardson, S. (2017). Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*.

Ailon, N. and Chazelle, B. (2009). The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39:302–322.

Bai, R., Boland, M. R., and Chen, Y. (2019). Fast algorithms and theory for high-dimensional bayesian varying coefficient models. *arXiv preprint arXiv:1907.06477*.

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12:583–614.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

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 (Statistical Methodology)*, 70(4):825–848.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042.

Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, 1(3):195–211.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. (2008). Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, 17(2):270–294.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1996.tb00936.x.

Burt, D. R., Rasmussen, C. E., and van der Wilk, M. (2020). Convergence of sparse variational inference in gaussian processes regression. *arXiv preprint arXiv:2008.00323*.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308.

Chen, S., Liu, Y., Lyu, M. R., King, I., and Zhang, S. (2015). Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210.

Clarkson, K. and Woodruff, D. (2017). Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63:1–45.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812.

Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., and Weiss, J. (2020). Vcbart: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*.

Dobriban, E. and Liu, S. (2018). A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische mathematik*, 117(2):219–249.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge university press.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154, https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210X.2010.00060.x.

Finley, A. O., Banerjee, S., and MacFarlane, D. W. (2011). A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, 106(493):31–48.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press.

Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

Guhaniyogi, R. and Dunson, D. B. (2016). Compressed gaussian process for manifold regression. *The Journal of Machine Learning Research*, 17(1):2472–2497.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of agricultural, biological, and environmental statistics*, 18(3):274–298.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2020a). Distributed bayesian varying coefficient modeling using a gaussian process prior. *arXiv preprint arXiv:2006.00783*.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2020b). A divide-and-conquer bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*.

Guhaniyogi, R. and Sansó, B. (2018). Large multi-scale spatial kriging using tree shrinkage priors. *arXiv preprint arXiv:1803.11331*.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Huang, Z. (2018). Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pages 2048–2057. PMLR.

Huang, Z., Li, J., Nott, D., Feng, L., Ng, T.-P., and Wong, T.-Y. (2015). Bayesian estimation of varying-coefficient models with missing data, with application to the singapore longitudinal aging study. *Journal of Statistical Computation and Simulation*, 85(12):2364–2377.

Ji, S., Xue, Y., and Carin, L. (2008). Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6):2346–2356.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, http://dx.doi.org/10.1080/01621459.2015.1123632.

Katzfuss, M. and Guinness, J. (2021). A General Framework for Vecchia Approximations of Gaussian Processes. *Statistical Science*, 36(1):124 – 141.

Kim, M. and Wang, L. (2021). Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics*, 30(1):1–10, https://doi.org/10.1080/10618600.2020.1754225.

Lee, J., Kamenetsky, M. E., Gangnon, R. E., and Zhu, J. (2021). Clustered spatio-temporal varying coefficient regression model. *Statistics in medicine*, 40(2):465–480.

Lemos, R. T. and Sansó, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association*, 104(485):5–18.

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The annals of applied statistics*, 9(2):640.

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv preprint arXiv:1104.5557*.

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(2):579–599, http://dx.doi.org/10.1080/10618600.2014.914946.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association (in press)*.

Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 143–152, USA. IEEE Computer Society.

Shen, W. and Ghosal, S. (2015). Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213.

Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18.

Vaart, A. v. d. and Zanten, H. v. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119.

Van der Vaart, A. W., van Zanten, J. H., et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B*, 50:297–312.

Vempala, S. S. (2005). *The random projection method*, volume 65. American Mathematical Soc.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vidakovic, B. (2009). *Statistical modeling by wavelets*, volume 503. John Wiley & Sons.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486):747–757.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569.

Wheeler, D. C. and Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2):145–166.

Wikle, C. K. (2010). Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 107-118.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

Yuan, X., Llull, P., Brady, D. J., and Carin, L. (2014). Tree-structure bayesian compressive sensing for video. *arXiv preprint arXiv:1410.3080*.

Zhang, L., Mahdavi, M., Jin, R., Yang, T., and Zhu, S. (2013). Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pages 135–157.

Zhou, S., Wasserman, L., and Lafferty, J. D. (2008). Compressed regression. In *Advances in Neural Information Processing Systems*, pages 1713–1720.