Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Mass-Shifting Phenomenon of Truncated Multivariate Normal Priors

Shuang Zhou, Pallavi Ray, Debdeep Pati & Anirban Bhattacharya

To cite this article: Shuang Zhou, Pallavi Ray, Debdeep Pati & Anirban Bhattacharya (11 Nov 2022): A Mass-Shifting Phenomenon of Truncated Multivariate Normal Priors, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2129059

To link to this article: https://doi.org/10.1080/01621459.2022.2129059

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2022.2129059

Taylor & Francis
Taylor & Francis Group

A Mass-Shifting Phenomenon of Truncated Multivariate Normal Priors

Shuang Zhou[a], Pallavi Ray[b], Debdeep Pati[c], and Anirban Bhattacharya[c]

[a]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ; [b]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN; [c]Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

We show that lower-dimensional marginal densities of dependent zero-mean normal distributions truncated to the positive orthant exhibit a *mass-shifting* phenomenon. Despite the truncated multivariate normal density having a mode at the origin, the marginal density assigns increasingly small mass near the origin as the dimension increases. The phenomenon accentuates with stronger correlation between the random variables. This surprising behavior has serious implications toward Bayesian constrained estimation and inference, where the prior, in addition to having a full support, is required to assign a substantial probability near the origin to capture flat parts of the true function of interest. A precise quantification of the mass-shifting phenomenon for both the prior and the posterior, characterizing the role of the dimension as well as the dependence, is provided under a variety of correlation structures. Without further modification, we show that truncated normal priors are not suitable for modeling flat regions and propose a novel alternative strategy based on shrinking the coordinates using a multiplicative scale parameter. The proposed shrinkage prior is shown to achieve optimal posterior contraction around true functions with potentially flat regions. Synthetic and real data studies demonstrate how the modification guards against the mass shifting phenomenon while retaining computational efficiency. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2020
Accepted September 2022

KEYWORDS
Bayesian; Basis expansion; Comparison inequality; Constrained estimation; Gaussian process; Shrinkage

1. Introduction

Let $p(\cdot)$ denote the density of a non-singular $\mathcal{N}_N(\mathbf{0}, \Sigma)$ distribution truncated to the nonnegative orthant in $\mathbb{R}^N$,

$$p(\theta) \propto e^{-\theta^{\mathrm{T}}\Sigma^{-1}\theta/2}\, \mathbb{1}_{\mathcal{C}}(\theta),$$
$$\mathcal{C} = [0,\infty)^N := \left\{\theta \in \mathbb{R}^N \,:\, \theta_1 \geq 0, \ldots, \theta_N \geq 0\right\}. \quad (1.1)$$

The positive definite matrix Σ will henceforth be referred to as the scale matrix associated with the truncated multivariate normal vector θ. The density p is clearly unimodal with its mode at the origin. However, for certain classes of non-diagonal Σ, we surprisingly observe that the lower-dimensional marginal distributions increasingly shift mass away from the origin as N increases. This observation is quantified in Theorem 2, where we provide nonasymptotic estimates for marginal probabilities of events of the form $\{\theta_1 \leq \delta\}$, for $\delta > 0$. En-route to the proof, we derive a novel Gaussian comparison inequality in Lemma S1 in the supplementary materials. An immediate implication of this mass-shifting phenomenon is that corner regions of the support $\mathcal{C}$, where a subset of the coordinates take values close to zero, increasingly become low-probability regions under $p(\cdot)$ as dimension increases. From a statistical perspective, this helps explain a paradoxical behavior in Bayesian constrained regression empirically observed in Curtis and Ghosh (2011) and Neelon and Dunson (2004), where truncated normal priors led to biased posterior inference when the underlying function had flat regions.

A common approach toward Bayesian constrained regression expands the function in a flexible basis which facilitates representation of the functional constraints in terms of simple constraints on the coefficient space, and then specifies a prior distribution on the coefficients obeying the said constraints. In this context, the multivariate normal distribution subject to linear constraints arises as a natural conjugate prior in Gaussian models and beyond. Various basis, such as Bernstein polynomials (Curtis and Ghosh 2011), regression splines (Cai and Dunson 2007; Meyer, Hackstadt, and Hoeting 2011), penalized spines (Brezger and Steiner 2008), cumulative distribution functions (Bornkamp and Ickstadt 2009), restricted splines (Shively, Walker, and Damien 2011), and compactly supported basis (Maatouk and Bay 2017) have been employed in the literature. For numerical illustrations in this article, we shall use the formulation of Maatouk and Bay (2017) where various restrictions such as boundedness, monotonicity, convexity, etc were equivalently translated into nonnegativity constraints on the coefficients under an appropriate basis expansion. They used a truncated normal prior as in (1.1) on the coefficients, with Σ induced from a parent Gaussian process on the regression function; see Appendix A for more details.

CONTACT Anirban Bhattacharya anirbanb@stat.tamu.edu Department of Statistics, Texas A&M University, College Station, TX, 77843.
This work was done by Pallavi Ray as a former Ph.D. student at Department of Statistics, Texas A&M University.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

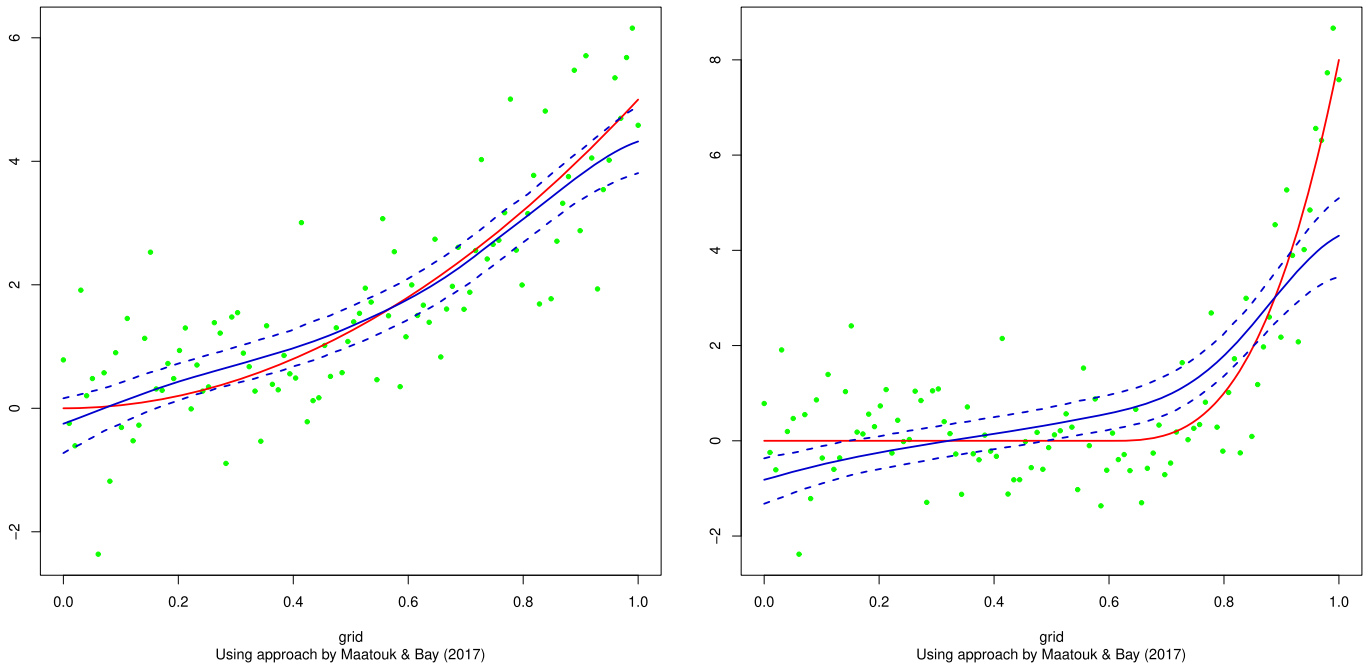

Figure 1. Monotone function estimation using the basis of Maatouk and Bay (2017) and a joint truncated normal prior $p(\cdot)$ on the coefficients. Red solid curve corresponds to the true function, blue solid curve is the posterior mean, the region within two dotted blue curves represents a pointwise 95% credible interval, and the green dots are observed data points. Left panel: true function is strictly monotone. Right panel: true function is monotone with a near-flat region.

Often the shape constraints enforce certain dependence structure in the joint prior for the coefficients to allow the posterior to borrow information from adjacent intervals (see e.g., Neelon and Dunson 2004). A multivariate normal prior truncated to the appropriate constraint set on the basis coefficients is used in such cases. For instance, with the Bernstein polynomials in Curtis and Ghosh (2011), the differences of coefficients are endowed with independent truncated normal prior, inducing dependence across the coefficients. In Neelon and Dunson (2004), a piecewise linear model with univariate truncated normal prior is used with autocorrelated means. In Maatouk and Bay (2017), a multivariate normal prior restricted to a set of linear constraints is an appropriate choice that ensures smoothness along the curve. We regard this as a prototype for the subsequent theoretical and empirical investigations.

To motivate our theoretical investigations, the two panels in Figure 1 depict the estimation of two different monotone smooth functions on $[0, 1]$ based on 100 samples using the basis of Maatouk and Bay (2017) and a joint prior $p(\cdot)$ as in (1.1) on the $N = 50$ dimensional basis coefficients. The same prior scale matrix Σ was employed across the two settings; the specifics are deferred to Section 3 and Appendix A. Observe that the function in the left panel is strictly monotone, while the one on the right panel is relatively flat over a region. While the point estimate (posterior mean) as well as the credible intervals look reasonable for the function in the left panel, the situation is significantly worse for the function in the right panel. The posterior mean incurs a large bias, and the pointwise 95% credible intervals fail to capture the true function for a substantial part of the input domain, suggesting that the entire posterior distribution is biased away from the truth. This behavior is perplexing; we are fitting a well-specified model with a prior that has full support[1] on the parameter space, which under mild conditions implies good first-order asymptotic properties (Ghosal, Ghosh, and van der Vaart 2000) such as posterior consistency. However, the finite sample behavior of the posterior under the second scenario clearly suggests otherwise.

Functions with flat regions as in the right panel of Figure 1 routinely appear in many applications; for example, dose–response curves are assumed to be nondecreasing with the possibility that the dose-response relationship is flat over certain regions (Neelon and Dunson 2004). A similar biased behavior of the posterior for such functions under truncated normal priors was observed by Neelon and Dunson (2004) while using a piecewise linear model, and also by Curtis and Ghosh (2011) under a Bernstein polynomial basis. However, a clear explanation behind such behavior as well as the extent to which it is prevalent has been missing in the literature, and the mass-shifting phenomenon alluded before offers an explanation. Under the basis of Maatouk and Bay (2017), a subset of the basis coefficients are required to shrink close to zero to accurately approximate functions with such flat regions. However, the truncated normal posterior pushes mass away from such corner regions, leading to the bias. Importantly, our theory also suggests that the problem would not disappear and would rather get accentuated in the large sample scenario if one follows standard practice of scaling up the number of basis functions with increasing sample size, since the mass-shifting gets more pronounced with increasing dimension. To illustrate this point, Figure S2 in Section S2.3 in the supplementary materials shows the estimation of the same function in the right

[1] The prior probability assigned to arbitrarily small Kullback–Leibler neighborhoods of any point is positive.

panel of Figure 1, now based on 500 samples and $N = 50$ and $N = 250$ basis functions in the left and right panel respectively. Increasing the number of basis functions indeed results in a noticeable increase in the bias as clearly seen from the insets which zoom into two disjoint regions of the covariate domain. A similar story holds for the basis of Curtis and Ghosh (2011) and Neelon and Dunson (2004).

One of our main contributions is to rigorously study the mass-shifting phenomenon of the marginal posterior distribution in a Bayesian shape-restricted inference problem. We show that the root of the problem lies in poor marginal posterior concentration around functions having flat regions, a phenomenon caused by the combined effects of truncation and dependence in truncated multivariate normal priors. For a general truncated normal distribution, we have found that the mass-shifting phenomenon occurs if the mode of the distribution lies near the boundary of the truncation region. A similar result is shown for the truncated normal posterior distribution in the context of Bayesian constrained inference.

Curtis and Ghosh (2011) and Neelon and Dunson (2004) both used point-mass mixture priors as remedy, which is a natural choice under a nondecreasing constraint. However, such mixture priors become somewhat cumbersome under the nonnegativity constraint in (1.1). As a simple remedy, we suggest introducing a multiplicative scale parameter for each coordinate a priori and further equipping it with a prior mixing distribution which has positive density at the origin and heavy tails; a default candidate is the half-Cauchy density (Carvalho, Polson, and Scott 2010; Polson and Scott 2012). In contrast to independent point-mass mixture priors, the resulting prior is more appealing as it retains the correlation between significant coefficients due to the dependence structure. The proposed prior shrinks more aggressively toward the origin, and we rigorously establish its de-biasing property that rectifies the mass-shifting issue, with empirical evidence of its superior performance over the truncated normal prior. Moreover, we offer theoretical justification toward prediction accuracy and parameter recovery over the class of all nondecreasing functions which may contain a flat region. In particular, the best obtainable posterior contraction rate is achieved adaptively regardless of the presence or absence of flat regions. Simulations studies are performed to compare the proposed method with the current state-of-the-art approaches under various choices of shape constraints. Multiple real applications provide further support to our argument that the proposed prior is robust to the presence or absence of flat regions in the true function. Proofs of all theorems and technical results are deferred to supplementary materials, which also contains additional simulations and figures, and implementation details.

2. Mass-Shifting Phenomenon of Truncated Normal Distributions

2.1. Marginal Densities of Truncated Normal Distributions

Our main focus is studying the properties of marginal densities of truncated normal distributions described in Equation (1.1) and quantifying how they behave with increasing dimensions. We begin by introducing some notations. We use $\mathcal{N}(\gamma, \Omega)$ to denote the d-dimensional normal distribution with mean $\gamma \in \mathbb{R}^d$ and positive definite covariance matrix Ω; also let $\mathcal{N}(x; \gamma, \Omega)$ denote its density evaluated at $x \in \mathbb{R}^d$. We reserve the notation $\Sigma_d(\rho)$ to denote the $d \times d$ compound-symmetry correlation matrix with diagonal elements equal to 1 and off-diagonal elements equal to $\rho \in (0, 1)$,

$$\Sigma_d(\rho) = (1 - \rho)\mathrm{I}_d + \rho \mathbf{1}_d \mathbf{1}_d^{\mathrm{T}}, \tag{2.1}$$

with $\mathbf{1}_d$ the vector of ones in $\mathbb{R}^d$ and I_d the $d \times d$ identity matrix.

For a subset $\mathcal{C} \subset \mathbb{R}^N$ with positive Lebesgue measure, let $\mathcal{N}_{\mathcal{C}}(\gamma, \Omega)$ denote a $\mathcal{N}(\gamma, \Omega)$ distribution truncated onto $\mathcal{C}$, with density

$$\widetilde{p}(\theta) = m_{\mathcal{C}}^{-1}\, \mathcal{N}(\theta; \gamma, \Omega)\, \mathbb{1}_{\mathcal{C}}(\theta), \tag{2.2}$$

where $m_{\mathcal{C}} = \mathbb{P}(X \in \mathcal{C})$ for $X \sim \mathcal{N}(\gamma, \Omega)$ is the constant of integration and $\mathbb{1}_{\mathcal{C}}(\cdot)$ the indicator function of the set $\mathcal{C}$. We throughout assume $\mathcal{C}$ to be the positive orthant of $\mathbb{R}^N$ as in Equation (1.1), namely, $\mathcal{C} = [0, \infty)^N$; a general $\mathcal{C}$ defined by linear inequality constraints can be reduced to rectangular constraints using a linear transformation—see, for example, section 2 of Botev (2017). The dimension N will be typically evident from the context.

Our investigations were originally motivated by the following observation. Consider $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}, \Sigma_2(\rho))$ for $\rho \in (0, 1)$. Then, the marginal distribution of θ_1 has density proportional to $e^{-\theta_1^2/2}\, \Phi\{\rho\theta_1/(1 - \rho^2)^{1/2}\}$ on $(0, \infty)$, where Φ denotes the $\mathcal{N}(0, 1)$ cumulative distribution function. This distribution is readily recognized as a skew normal density (Azzalini and Valle 1996) truncated to $(0, \infty)$. Interestingly, the marginal of θ_1 has a strictly positive mode, while the joint distribution of θ had its mode at $\mathbf{0}$. Cartinhour (1990) noted that the truncated normal family is not closed under marginalization for nondiagonal Σ, and derived a general formula for the univariate marginal as the product of a univariate normal density with a skewing factor. In Proposition 1, we generalize the result in Cartinhour (1990) for any lower-dimensional marginal density. We write the scale matrix Σ_N in block form as $\Sigma_N = [\Sigma_{k,k},\ \Sigma_{N-k,k}; \Sigma_{k,N-k},\ \Sigma_{N-k,N-k}]$.

Proposition 1. Suppose $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Sigma_N)$. The marginal density $\widetilde{p}_{k,N}$ of $\theta^{(k)} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ is

$$\widetilde{p}_{k,N}(\theta_1, \ldots, \theta_k) = (2\pi)^{-k/2}\, m_{\mathcal{C}}^{-1}\, e^{-\frac{1}{2}\theta^{(k)\mathrm{T}} \Sigma_{k,k}^{-1} \theta^{(k)}} \mathbb{P}(\widetilde{X}_{N-k} \le \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}\, \theta^{(k)}) \prod_{i=1}^{k} \mathbb{1}_{[0,\infty)}(\theta_i),$$

where $\widetilde{X}_{N-k} \sim \mathcal{N}(\mathbf{0}_{N-k}, \widetilde{\Sigma}_{N-k,N-k}^{-1})$ with $\widetilde{\Sigma}_{N-k,N-k} = (\Sigma_{N-k,N-k} - \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}\, \Sigma_{k,N-k})^{-1}$, and the $\le$ symbol is to be interpreted elementwise. Here, the constant $m_{\mathcal{C}} = \mathbb{P}(X \in \mathcal{C})$ for $X \sim \mathcal{N}(\mathbf{0}_N, \Sigma_N)$.

When $k = 1$, Proposition 1 implies

$$\widetilde{p}_{1,N} \propto e^{-\theta_1^2/(2\Sigma_{1,1})} \mathbb{P}(\widetilde{X}_{N-1} \le \Sigma_{N-1,1}\, \theta_1 / \Sigma_{1,1}) \mathbb{1}_{[0,\infty)}(\theta_1). \tag{2.3}$$

Let $\mathcal{S}_N$ denote the set of $N \times N$ covariance matrices whose correlation coefficients are all nonnegative. The map $\theta_1 \mapsto e^{-\theta_1^2/(2\Sigma_{1,1})}$ is decreasing and when $\Sigma_N \in \mathcal{S}_N$, $\theta_1 \mapsto \mathbb{P}(\widetilde{X}_{N-1} \leq \Sigma_{N-1,1}\,\theta_1/\Sigma_{1,1})$ is increasing, on $(0,\infty)$. Thus, if $\Sigma_N \in \mathcal{S}_N$ and it contains nonzero off-diagonal elements, $\widetilde{p}_{1,N}$ is unimodal with a strictly positive mode.

As another special case, suppose $\Sigma_N = \Sigma_N(\rho)$ for some $\rho \in (0,1)$ and let $k = N-1$. We then have,

$$\widetilde{p}_{N-1,N} \propto e^{-\theta^{(N-1)^{\mathrm{T}}}\Sigma_{N-1}^{-1}(\rho)\,\theta^{(N-1)}}\,\Phi(a^{\mathrm{T}}\theta^{(N-1)})\prod_{i=1}^{N-1}\mathbb{1}_{[0,\infty)}(\theta_i),$$

with $a = C_\rho\big(\sum_{i=1}^{N-1}\theta_i\big)\,\mathbf{1}_{N-1}$, where C_ρ is a positive constant. This density can be recognized as a multivariate skew-normal distribution (Azzalini and Valle 1996) truncated to the nonnegative orthant.

2.2. Mass-Shifting Phenomenon of Marginal Densities

While the results in the previous section imply that the marginal distributions shift mass away from the origin, they do not precisely characterize the severity of its prevalence. In this section, we show that under appropriate conditions, the univariate marginals assign increasingly smaller mass to a fixed neighborhood of the origin with increasing dimension. In other words, the skewing factor noted by Cartinhour (1990) begins to dominate when the ambient dimension is large. In addition to the dimension, we also quantify the amount of dependence in Σ_N contributing to this mass-shifting. On the other hand, when the mode of a truncated multivariate normal distribution lies in the interior of truncation region and not on the boundary, the skewness is less pronounced. One of our main interests is to understand the relation between the magnitude of the mode and the severity of the mass shifting phenomenon. To the best of our knowledge, such results have not been observed or quantified in the literature.

In the following, we consider a truncated multivariate normal distribution $\mathcal{N}_{\mathcal{C}}(\boldsymbol{\mu}_N, \Sigma_N)$ where the center $\boldsymbol{\mu}_N$ is assumed to be coordinate-wise nonnegative and Σ_N is a banded nonnegative scale matrix with equal variances. Such a banded structure arises naturally in statistical applications as will be discussed in the next section. Define the space of K-banded nonnegative and equal-variance scale matrices as

$$\mathcal{B}_{N,K} = \Big\{\Sigma_N = (\sigma_{ij}) \in \mathcal{S}_N \,:\, \sigma_{ii} \equiv \sigma^2\ \forall\, i, \text{for some } \sigma^2 > 0;\ \sigma_{ij} \in (0,\sigma^2), \forall\, |i-j| < K,\ \ \sigma_{ij} = 0,\ \forall\, |i-j| \geq K\Big\}, \tag{2.4}$$

for $2 \leq K \leq N-1$. In the sequel, the bandwidth K is allowed to increase with N.

For any $\Sigma_N \in \mathcal{B}_{N,K}$, it can be viewed as a scaling of a positive correlation matrix with σ^2 as the scale. We let $\rho_{ij} = \sigma_{ij}/\sigma^2$ for $1 \leq i,j \leq N$ and then define $\rho_{\max} = \max_{i\neq j,|i-j|<K}\{\rho_{ij}\}$ and $\rho_{\min} = \min_{i\neq j,|i-j|<K}\{\rho_{ij}\}$ as the maximum and minimum off-diagonal elements within the band separately to ensure $\rho_{\max}, \rho_{\min} \in (0,1)$.

We assume $\boldsymbol{\mu}_N = \{\mu_j\}$ with $\mu_j \geq 0$ for $j = 1,\ldots,N$ and denote $\mu^* = \max_{1\leq j\leq N}\{\mu_j\}$. For $\theta \sim \mathcal{N}_{\mathcal{C}}(\boldsymbol{\mu}_N, \Sigma_N)$ with $\mathcal{C} = [0,\infty)^N$, let $\alpha_{N,\delta} = \mathbb{P}(\theta_1 \leq \delta)$. With these definitions and assumptions in place, we are ready to state the following key theorem.

Theorem 2 (Strictly banded case). Let $\Sigma_N \in \mathcal{B}_{N,K}$ be such that $(\rho_{\min}, \rho_{\max}) \in \mathcal{Q}$, where

$$\mathcal{Q} = \left\{(u,v) \in (0,1)^2 : u \leq v,\ \frac{u}{2(1-u)} \geq v\right\}.$$

Fix $\beta \in [0,1)$. For any $\boldsymbol{\mu}_N$ satisfying $\mu^* \leq C_{\rho_{\min},\rho_{\max}}\beta\, G_\alpha(\rho_{\min},\rho_{\max})(\log K)^{1/2}$, there exists a constant K_0 such that whenever $K \geq K_0$, we have for any $\delta > 0$,

$$\alpha_{N,\delta} \leq C'_{\rho_{\min},\rho_{\max}}\,(\delta/\sigma)\,(\log K)^{1/2}K^{-(1-\beta)G_\alpha(\rho_{\min},\rho_{\max})},$$

where $G_\alpha(\rho_{\min},\rho_{\max}) = (1-\alpha)/\rho_{\max} - 2(1-\rho_{\min})/\rho_{\min}$ for some constant $\alpha \in (0,1)$, and $C_{\rho_{\min},\rho_{\max}}$, $C'_{\rho_{\min},\rho_{\max}}$ are positive constants free of K, N.

In particular, if we consider a sequence of K_N-banded scale matrices $\Sigma_N \in \mathcal{B}_{N,K_N}$ with $K_N \to \infty$ as $N \to \infty$, then under the conditions of Theorem 2, $\lim_{N\to\infty}\alpha_{N,\delta} = 0$ for any fixed $\delta > 0$. Theorem 2, being non-asymptotic in nature, additionally characterizes the rate of decay of $\alpha_{N,\delta}$. Simulations illustrating the conclusion of Theorem 2 can be found in Section S2.1 of the supplementary materials, where the univariate marginal density $\widetilde{p}_{1,N}$ is displayed under different values of the dimension N and the bandwidth K. To contrast the conclusion of Theorem 2 with two closely related cases, consider first the case when $\theta \sim \mathcal{N}(\mathbf{0}_N, \Sigma_N)$. For any N, the marginal distribution of θ_1 is always $\mathcal{N}(0,1)$, and hence $\alpha_{N,\delta}$ does not depend on N. Similarly, if $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Sigma_N)$ with Σ_N a diagonal correlation matrix, then for any $N \geq 1$, the marginal distribution of θ_1 is $\mathcal{N}(0,1)$ truncated to $(0,\infty)$ and $\alpha_{N,\delta}$ again does not depend on N. In particular, in both these cases, $\alpha_{N,\delta} \asymp \delta$ for δ small. Here we denote $a_n \asymp b_n$ for positive sequences a_n, b_n if $0 < \liminf a_n/b_n < \limsup a_n/b_n < \infty$. However, when a combination of dependence and truncation is present, an additional $(\log K)^{1/2}\,K^{-G_\alpha(\rho_{\min},\rho_{\max})}$ penalty (obtained by setting $\beta = 0$) is incurred. When $\mu^* > 0$, the theorem reveals an inverse relationship between the allowable upper bound on μ^* and $\alpha_{N,\delta}$, implying that the mass shifting phenomenon is mitigated when the mode of the truncated multivariate normal shifts to the right.

For the conclusion of Theorem 2 to hold, our current proof technique requires $(\rho_{\min},\rho_{\max})$ to lie in the region $\mathcal{Q}$, which is pictorially represented by the black shaded region in Figure S3 in Section S2.3 of the supplementary material. Fixing α and $\rho_{\max}$, one can see that a larger value of $\rho_{\min}$ leads to a greater value of G_α and thus a more severe mass-shifting problem. As a special case, if all the nonzero correlations are the same, that is, $\rho_{\min} = \rho_{\max}$, then the condition simplifies to $\rho_{\min} > 0.5$. More generally, if we write $\rho_{\min} = \kappa\rho_{\max}$ for some $\kappa \in (0,1]$, then the condition reduces to $\rho_{\min} \geq 1 - \kappa/2$.

Remark 1. For any fixed N, the marginal density of θ_1 evaluated at the origin, $\widetilde{p}_{1,N}(0) = \lim_{\delta\to 0}\alpha_{N,\delta}/\delta$. Theorem 2 thus implies in particular that $\lim_{N\to\infty}\widetilde{p}_{1,N}(0) = 0$, when $K = K_N \to \infty$ as $N \to \infty$. Also, for any fixed $1 \leq k \leq N$, if we denote $\beta_{N,k,\delta} = \mathbb{P}(\theta_1 \leq \delta, \ldots, \theta_k \leq \delta)$, it is immediate that $\beta_{N,k,\delta} < \alpha_{N,\delta}$, and

hence $\lim_{N\to\infty}\beta_{N,k,\delta}=0$, meaning the probability of a corner region is vanishingly small for large N.

Theorem 2 assumed a strictly banded assumption on the scale matrix which may be restrictive in real applications. We now generalize the result to "approximately" banded scale matrices, which allow long range dependency between the variables. For an arbitrary nonnegative scale matrix Σ_N, we say it is "approximately" banded if there exists some integer $2\le K\le N-1$ such that there exists a matrix $\Sigma'_N\in\mathcal{B}_{N,K}$ satisfying $||\Sigma'_N-\Sigma_N||\le\varepsilon(N,K)||\Sigma_N||$ for some sufficiently small constant $\varepsilon(N,K)>0$ that may depend on N,K. We denote the operator norm of a matrix A by $||A||=\{\lambda_{\max}(A^{\mathrm{T}}A)\}^{1/2}$ where $\lambda_{\max}(A)$ denotes its largest eigenvalue.

As our analysis relies on Theorem 2, the assumptions are directly applied to the banded approximating matrix Σ'_N. Akin to Theorem 2, we assume $\Sigma'_N=(\sigma'_{ij})\in\mathcal{B}_{N,K}$ and denote the variance by σ'^2. We then denote by $\rho'_{\min},\rho'_{\max}\in(0,1)$ the minimum and maximum values within the band after scaling Σ'_N by σ'^2, respectively. And for positive sequences a_n,b_n, we denote $a_n\lesssim b_n$ if $a_n\le Cb_n$ for some universal fixed constant $C>0$, similarly we define $a_n\gtrsim b_n$.

Theorem 3 (Approximately banded case). Let $\theta\sim\mathcal{N}_{\mathcal{C}}(\boldsymbol{\mu}_N,\Sigma_N)$, where Σ_N is a positive definite scale matrix. Assume there exists an integer $2\le K\le N-1$ such that one can construct a K-banded matrix $\Sigma'_N\in\mathcal{B}_{N,K}$ that satisfies $||\Sigma_N-\Sigma'_N||\lesssim(N\log K)^{-1}||\Sigma_N||$, and assume $(\rho'_{\min},\rho'_{\max})\in\mathcal{Q}$. Fix $\beta\in[0,1)$. For any $\boldsymbol{\mu}_N$ satisfying $\mu^*\le C_{\rho'_{\min},\rho'_{\max}}\beta G_\alpha(\rho'_{\min},\rho'_{\max})(\log K)^{1/2}$, there exists some integer $K_0>0$ such that for $K>K_0$, for any fixed $\delta>0$ and some $\alpha\in(0,1)$,

$$\alpha_{N,\delta}\le C''_{\rho'_{\min},\rho'_{\max}}\,\delta\,(\log K)^{1/2}K^{-(1-\beta)G_\alpha(\rho'_{\min},\rho'_{\max})},$$

where $\mathcal{Q},G_\alpha(\rho'_{\min},\rho'_{\max})$ and $C_{\rho'_{\min},\rho'_{\max}}$ are defined in Theorem 2, and $C''_{\rho'_{\min},\rho'_{\max}}$ is a positive constant that is independent of K,N.

Theorem 3 states the upper bound of $\alpha_{N,\delta}$ remains the same as in Theorem 2 even if the scale matrix deviates slightly from a banded structure. As Theorem 2 holds for any $K>K_0$, it guarantees that the conclusion holds for Σ_N so long as it can be approximated by a wide enough banded matrix. To obtain such banded approximating matrix for general matrices, one may adopt commonly used techniques in matrix approximation (e.g., Bickel and Lindner 2012; Yoo and Ghosal 2016). Indeed in Section 3, we shall discuss an example where the induced posterior scale matrix can be approximated by a banded matrix. Theorem 3 can be also applied to "approximately" banded scale matrices with unequal variances. A similar result for the *unequal-variance* case is deferred to Corollary S1 in Section S2.2 of the supplementary materials. This generalization allows us to apply our mass-shifting theory to a truncated normal posterior induced from Bayesian isotonic regression estimation problem, since the banded approximating matrix of the associated posterior scale matrix may not have equal variances in general. More discussion will be given in Section 3 in the context of Bayesian constrained regression estimation.

3. Connections with Bayesian Constrained Inference

In this section, we connect the theoretical findings in the previous section to posterior inference in Bayesian constrained regression models. We work under the setup of a Gaussian regression model,

$$y_i=f(x_i)+\epsilon_i,\quad \epsilon_i\sim\mathcal{N}(0,\sigma^2),\quad i=1,\dots,n,\tag{3.1}$$

where we assume $x_i\in[0,1]$ for simplicity. We are interested in the situation when the regression function f is constrained to lie in some space $\mathcal{C}_f$ which is a subset of the space of all continuous functions on $[0,1]$, determined by linear restrictions on f and possibly its higher-order derivatives. Common examples include bounded, monotone, convex, and concave functions.

As discussed in the introduction, a general approach is to expand f in some basis $\{\phi_j\}$ as $f(\cdot)=\sum_{j=1}^N\theta_j\phi_j(\cdot)$ so that the restrictions on f can be posed as linear restrictions on the vector of basis coefficients $\theta\in\mathbb{R}^N$, with the parameter space $\mathcal{C}$ for θ of the form $\mathcal{C}=\{\theta\in\mathbb{R}^N:A\theta\ge b\}$. For example, when $\mathcal{C}_f$ corresponds to monotone increasing functions, the set $\mathcal{C}$ is of the form $\{\theta_1\le\theta_2\le\cdots\le\theta_N\}$ under the Bernstein polynomial basis (Curtis and Ghosh 2011) and $[0,\infty)^N$ under the integrated triangular basis of Maatouk and Bay (2017). For sake of concreteness, we shall henceforth work with $\mathcal{C}=[0,\infty)^N$. Under such a basis representation, the model (3.1) can be expressed as

$$Y=\Phi\theta+\epsilon,\quad \epsilon\sim\mathcal{N}(\mathbf{0}_n,\mathrm{I}_n),\quad \theta\in\mathcal{C},\tag{3.2}$$

where $Y=(y_1,\dots,y_n)^{\mathrm{T}}$ and $\Phi=\{\phi_j(x_i)\}_{ij}$ is an $n\times N$ basis matrix.

The truncated normal prior $\theta\sim\mathcal{N}_{\mathcal{C}}(\mathbf{0}_N,\Omega_N)$ is conjugate, with the posterior $\theta|Y\sim\mathcal{N}_{\mathcal{C}}(\boldsymbol{\mu}_N,\Sigma_N)$, with $\mu_N=\Sigma_N\Phi^{\mathrm{T}}Y$ and $\Sigma_N=(\Omega_N^{-1}+\Phi^{\mathrm{T}}\Phi)^{-1}$. To motivate their prior choice, Maatouk and Bay (2017) begin with an unconstrained mean-zero Gaussian process prior on f, $f\sim\mathrm{GP}(0,K)$, with covariance kernel K. Since their basis coefficients correspond to evaluation of the function and its derivatives at the grid points (see Appendix A for details), this induces a multivariate zero-mean Gaussian prior $\mathcal{N}(\mathbf{0}_N,\Omega_N)$ on θ provided the covariance kernel K of the parent Gaussian process is sufficiently smooth. Having obtained this unconstrained Gaussian prior on θ, Maatouk and Bay (2017) multiply it with the indicator function $\mathbb{1}_{\mathcal{C}}(\theta)$ of the truncation region to obtain the truncated normal prior.

We are now in a position to connect the posterior bias in Figure 1 and Figure S2 (in the supplementary materials) to the mass-shifting phenomenon characterized in the preceding section. Consider an extreme scenario where the true function $f_0(x)\equiv0$ for all $x\in[0,1]$. Expanding f_0 on a proper basis such as the basis (M) in (A.1), one can express $[f_0(x_1),\dots,f_0(x_n)]^{\mathrm{T}}=\Phi\theta_0$ where the pseudo-true parameter $\theta_0=\mathbf{0}_N$ (or approximately). Given the posterior $\theta|Y\sim\mathcal{N}_{\mathcal{C}}(\boldsymbol{\mu}_N,\Sigma_N)$, a draw from the posterior can be represented as

$$\theta=\boldsymbol{\mu}_N+\theta_c,\quad \theta_c\sim\mathcal{N}_{\mathcal{C}}(\mathbf{0}_N,\Sigma_N).\tag{3.3}$$

Under mild assumptions, $\boldsymbol{\mu}_N$ concentrates near the origin with high probability under the true data generating distribution. The mass-shifting phenomenon pushes θ_c away from the origin, resulting in the bias. On the other hand, when the true function is strictly monotone as in the left panel of Figure 1, all

the entries of $\boldsymbol{\mu}_N$ are bounded away from zero, which masks the effect of the shift in θ_c.

In strict technical terms, our theory is not directly applicable to θ_c since the scale matrix Σ_N is a dense matrix in general. However, we show below that Σ_N is approximately banded under mild conditions. Figure S4 in Section S2.3 of the supplementary materials shows image plots of Σ_N for three choices of N using the basis of Maatouk and Bay (2017) and sample size $n = 500$. In all cases, Σ_N is seen to have a near-banded structure.

We make this empirical observation concrete below. We first state the assumptions on the basis matrix Φ and prior scale matrix Ω_N that allow the construction of a strictly banded matrix approximation.

Assumption 1. We assume the basis matrix Φ is such that the matrix $\Phi^{\mathrm{T}}\Phi$ is q-banded for some $2 \le q \le N$; also there exist constants $0 < C_1 < C_2 < \infty$ such that $C_1\,(n/N)\,\mathrm{I}_N \le \Phi^{\mathrm{T}}\Phi \le C_2\,(n/N)\,\mathrm{I}_N$.

One example of a basis satisfying Assumption 1 is a B-Spline of fixed order q denoted as $B_{N,q}(x)$ with $N = J + q$ over quasi-uniform knot points of number $J > 0$; see, for example, Yoo and Ghosal (2016). The basis (M) introduced in (A.1) also satisfies Assumption 1 under some mild conditions on the grid points by Lemma 2 in Appendix A. For any square matrix A, let $\lambda_{\min}(A), \lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A, respectively. Now define a uniform class of symmetric positive definite well-conditioned matrices (Bickel and Levina 2008) as

$$\begin{aligned}\mathcal{M}(\lambda_0, \alpha, k) = \Big\{\Omega_N = (\omega_{ij}) \,:\, & \max_j \sum_i \{|\omega_{ij}| \,:\, |i-j| > k\} \\ & \le C\,k^{-\alpha} \text{ for all } k > 0, \text{ and } 0 < \lambda_0 \le \lambda_{\min}(\Omega_N) \\ & \le \lambda_{\max}(\Omega_N) \le 1/\lambda_0\Big\}, \end{aligned} \tag{3.4}$$

for some positive constants $\alpha, \lambda_0 > 0$.

Assumption 2. We assume the prior scale matrix $\Omega_N \in \mathcal{M}(\lambda_0, \alpha, k)$ defined in (3.4).

Given above assumptions, we are now ready to give the approximation result of posterior scale matrix Σ_N to a banded symmetric positive definite matrix. We first introduce few new notations. For positive sequences a_n, b_n, we write $a_n = O(b_n)$ if there exists a global constant $C' > 0$ such that $a_n \le C' b_n$ and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For any $a \in \mathbb{R}$, we denote by $\lfloor a \rfloor$ the greatest integer that is no larger than a.

Proposition 4. If Φ and Ω_N in $\Sigma_N = (\Omega_N^{-1} + \Phi^{\mathrm{T}}\Phi)^{-1}$ satisfy Assumptions 1 and 2, for sufficiently small $\epsilon < \min\{\lambda_0, 1/\lambda_0\}$ and for any integer $n_0 \ge \lfloor \log[\lambda_0(\lambda_0 - \epsilon)/(1 + \lambda_0^2)]/\log\kappa \rfloor$, there exists $r \gtrsim \log(1/\epsilon)$ such that we can find a K-banded, symmetric and positive definite matrix $\widetilde{\Sigma}_N$ with $||\Sigma_N - \widetilde{\Sigma}_N|| \lesssim \delta_{\epsilon,\kappa}$, where $K = \max(n_0^2 r, n_0 q)$ with q defined in Assumption 1, $\delta_{\epsilon,\kappa} = (\epsilon + \kappa^{n_0+1})(N/n)$ and $0 < \kappa < 1$ is a fixed constant.

Remark 2. It is easy to show that $||\Sigma_N|| \asymp N/n$ under Assumptions 1 and 2. Proposition 4 implies one can construct $\widetilde{\Sigma}_N$ with $||\widetilde{\Sigma}_N|| \asymp N/n$. Moreover, under model (3.2), if Assumptions 1 and 2 are satisfied, then by letting $N = o(n)$ and choosing K and n_0 such that $\max\{K, n_0\} \gtrsim (\log N)^t$ for some constant $t > 0$ and for sufficiently large N, one has $||\Sigma_N - \widetilde{\Sigma}_N|| \lesssim (N \log K)^{-1}||\Sigma_N||$.

Proposition 4 states under mild conditions one can always construct a banded positive definite matrix that approximates Σ_N in operator norm. Proposition S1 (in the supplementary materials) guarantees if $\delta_{\epsilon,\kappa}$ is small enough, the marginal probability $\alpha_{N,\delta}$ will not change significantly if Σ_N is replaced by its banded approximate $\widetilde{\Sigma}_N$. These are the key ingredients to translate the mass-shifting results for the prior in Section 2 to the marginal posterior distribution. To this end, we adopt similar notations and assumptions used in Theorem 3. Assumptions regarding the correlation structure are imposed on the banded approximating matrix $\widetilde{\Sigma}_N = (\widetilde{\sigma}_{ij})$. Let $\widetilde{\sigma}^2_{(1)}, \widetilde{\sigma}^2_{(N)}$ denote the smallest and largest variances of $\widetilde{\Sigma}_N$. Without loss of generality, we assume $\widetilde{\sigma}_{11} = \widetilde{\sigma}^2_{(1)}$ for simplicity. We denote by $\widetilde{\sigma}_{\min} = \min_{i\neq j, |i-j|<K}\{\widetilde{\sigma}_{ij}\}$, $\widetilde{\sigma}_{\max} = \max_{i\neq j, |i-j|<K}\{\widetilde{\sigma}_{ij}\}$ the smallest and largest positive off-diagonal entries of $\widetilde{\Sigma}_N$, respectively. By scaling $\widetilde{\Sigma}_N$ by the value of its smallest variance, we let $\widetilde{\rho}_{ij} = \widetilde{\sigma}_{ij}/\widetilde{\sigma}^2_{(1)}$ for all $1 \le i, j \le N$. Further define $\widetilde{\kappa} = \widetilde{\sigma}^2_{(N)}/\widetilde{\sigma}^2_{(1)}$, $\widetilde{\rho}_{\min} = \widetilde{\sigma}_{\min}/\widetilde{\sigma}^2_{(1)}$ and $\widetilde{\rho}_{\max} = \widetilde{\sigma}_{\max}/\widetilde{\sigma}^2_{(1)}$. In addition, we assume $\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max} \in (0, 1)$.

Theorem 5. Let $\theta|Y \sim \mathcal{NC}(\boldsymbol{\mu}_N, \Sigma_N)$ and assume $\widetilde{\Sigma}_N$ obtained from Proposition 4 satisfies $\widetilde{\Sigma}_N \in \mathcal{B}_{N,K}$ and $||\Sigma_N - \widetilde{\Sigma}_N|| \lesssim (N \log K)^{-1}||\Sigma_N||$ for sufficiently large N. Assume $(\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}) \in \mathcal{Q}_{\widetilde{\kappa}}$, where

$$\mathcal{Q}_s = \left\{(u, v, s) \in (0,1)^2 \otimes [1, \infty) : u \le v,\ \frac{u}{2(s-u)} \ge v\right\}.$$

Fix an arbitrary $\delta > 0$.

(a) Recall $\theta_c \sim \mathcal{NC}(\mathbf{0}_N, \Sigma_N)$ defined in Equation (3.3), then there exists a sufficiently large integer K_0 such that for $K > K_0$,

$$\begin{aligned}&\Pi(0 < \theta_{c1} < \delta|Y) \\ &\quad \le C_{\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}}(\delta/\widetilde{\sigma}_{(1)})(\log K)^{1/2} K^{-G_\alpha(\widetilde{\rho}_{\min}/\widetilde{\kappa},\, \widetilde{\rho}_{\max})},\end{aligned}$$

where the function G_α is same in Theorem 2 for some $\alpha \in (0, 1)$. The constant $C_{\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}}$ is free of N, K.

(b) In addition, there exists a sufficiently large integer K_0' such that for $K > K_0'$, with at least $\mathbb{P}_0$-Probability[2] $1 - C_1'(\log K)^{(1-\beta)/2}\exp\{-C_2'(\log K)^{1-\beta}\}$ for some fixed constants $C_1', C_2' > 0$ and $\beta \in (0, 1)$, we have

$$\begin{aligned}&\Pi(0 < \theta_1 < \delta|Y) \\ &\quad \le C'_{\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}}(\delta/\widetilde{\sigma}_{(1)})(\log K)^{1/2} K^{-G_\alpha(\widetilde{\rho}_{\min}/\widetilde{\kappa},\, \widetilde{\rho}_{\max})},\end{aligned}$$

where the function G_α is same as in Theorem 2 for some $\alpha \in (0, 1)$. The positive constant $C'_{\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}}$ is free of N, K.

Remark 3. For any $\widetilde{\kappa} \ge 1$, $(\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max}, \widetilde{\kappa}) \in \mathcal{Q}_{\widetilde{\kappa}}$ implies $\widetilde{\rho}_{\min} > 1/2$. For any fixed $\widetilde{\rho}_{\min} \in (1/2, 1)$, the condition $u/\{2(\widetilde{\kappa} - u)\} \ge v$ defined in $\mathcal{Q}_{\widetilde{\kappa}}$ implies that $\widetilde{\kappa} \le \widetilde{\rho}_{\min} + 1/2$. This leads to $1 \le \widetilde{\kappa} < 3/2$, which indicates the ratio of largest and smallest variances cannot be greater than 3/2. Thus, if $\widetilde{\kappa} \ge 1$, the area $\mathcal{Q}_{\widetilde{\kappa}}$ places a slightly stronger restriction on $(\widetilde{\rho}_{\min}, \widetilde{\rho}_{\max})$ than the one in the equal-variance scenario considered in Theorem 3. The stronger restriction can be viewed as a price to pay to accommodate a more complex dependence structure with unequal variances.

[2] $\mathbb{P}_0$ denotes the true data generating measure.

Theorem 5 characterizes the decaying rate of the marginal posterior probability assigned over a fixed neighborhood of the truth. This result is a combined effect of two key phenomena: the posterior mode $\boldsymbol{\mu}_N$ is close to the origin with high probability and the posterior scale matrix Σ_N is approximately K-banded. The posterior distribution thus inherits the undesirable mass-shifting property as exhibited by Theorem 5. As a corollary of Theorem 5, we now show that the posterior of $\Phi\theta$ fails to contract at the optimal rate toward a true flat function f_0. We use $\mathbb{E}_0(\cdot)$ to denote expectation taken with respect to the true probability density function.

Proposition 6. Assume that $Y \sim \mathcal{N}(\mathbf{0}_n, \mathrm{I}_n)$ and the basis matrix Φ satisfies Assumption 1. For a truncated multivariate normal prior $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Sigma_N)$ with $\mathcal{C} = [0,\infty)^N$ and the prior scale matrix Σ_N satisfying Assumption 2,

$$\mathbb{E}_0\, \Pi\big(||\Phi\theta - f_0|| \lesssim \sqrt{N}|Y\big) \to 0 \quad \text{as } n, N \to \infty.$$

Remark 4. As an intermediate result in the proof of Proposition 6, we observed that $\Pi(\theta|Y)$ also fails to contract at the optimal rate toward the pseudo-true parameter vector θ_0.

4. A De-biasing Remedy based on a Shrinkage Prior

4.1. A Dependent Global-Local Shrinkage Procedure

In this section, we propose a simple modification to the truncated normal prior that can alleviate the issues related to such mass-shifting phenomenon. Among remedies proposed in the literature (Neelon and Dunson 2004; Dunson 2005; Curtis and Ghosh 2011), a discrete point-mass mixture shrinkage prior was employed $\theta_k \sim (1-\pi)\delta_0 + \pi\mathcal{N}_{\mathcal{D}}(\mu_k, \sigma_k^2)$ coordinate-wise on the parameter vector $\{\theta_k\}$. The mass at zero allows positive prior probability to functions having exactly flat regions, and the normal density truncated to set $\mathcal{D}$ incorporates the non-decreasing constraint. For example, Curtis and Ghosh (2011) considered $\mathcal{D} = (0,\infty)$. In contrast to a regular variable selection scenario where coefficients are treated independently, here the coefficients (might be related to function evaluations) are assumed to be a priori correlated to facilitate smoothness of the function estimates. The hyperparameters $\{\mu_k, \sigma_k^2\}$ of truncated normal distributions can also be chosen to encompass certain dependence structure, see, for example, Dunson (2005) and Neelon and Dunson (2004).

Although possible in principle, introduction of such discrete structure (point-masses) while retaining the dependence structure between the coefficients becomes somewhat cumbersome in addition to being computationally burdensome. With such motivation and the additional consideration that in most real scenarios a function is approximately flat in certain regions, we propose a shrinkage procedure as a remedy to replace the coefficients $\theta \in \mathcal{C}$ by $\xi = (\xi_1, \ldots, \xi_N)^{\mathrm{T}}$, where

$$\xi_j = \tau\, \lambda_j\, \theta_j, \quad j = 1, \ldots, N. \tag{4.1}$$

The parameter τ provides global shrinkage toward the origin while the λ_js provide coefficient-specific deviations. We consider default (Carvalho, Polson, and Scott 2010) half-Cauchy priors $\mathcal{C}_+(0,1)$, which has a density proportional to $(1+t^2)^{-1}\mathbb{1}_{(0,\infty)}(t)$, on τ and the λ_js independently. We continue to use a dependent truncated normal prior $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Omega_N)$ which in turn induces dependence among the ξ_js. Our prior on ξ can thus be considered as a dependent extension of the global-local shrinkage priors (Carvalho, Polson, and Scott 2010) widely used in the high-dimensional regression context. We named our prior as DGL-TMVN (dependent global local truncated multivariate normal) prior. Figure S8 in Section S8.1 of the supplementary materials shows prior draws for the first and third components of both θ and ξ, based on which the marginal distributions of the ξ_js are clearly seen to place more mass near the origin while retaining heavy tails.

We investigate the proposed shrinkage procedure in the context of estimating monotone functions as described in (A.1). The procedure can be readily adapted to include various other constraints. Replacing θ by ξ in (M) in (A.1), we can write (3.1) in vector notation as

$$Y = \zeta\, \mathbf{1}_n + \tau\, \Psi\, \Lambda\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathrm{I}_n). \tag{4.2}$$

Here, Ψ is an $n \times N$ basis matrix with ith row Ψ_i^{T} where $\Psi_{ij} = \psi_{j-1}(x_i)$ for $j = 1, \ldots, N$ and the basis functions ψ_j are as in (A.1). Also, $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ and $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$.

The model is parameterized by $\zeta \in \mathbb{R}, \theta = (\theta_1, \ldots, \theta_N)^{\mathrm{T}} \in \mathcal{C}$, $\lambda = (\lambda_1, \ldots, \lambda_N)^{\mathrm{T}} \in \mathcal{C}$, $\sigma \in \mathbb{R}^+$ and $\tau \in \mathbb{R}^+$. We place a flat prior $\pi(\zeta) \propto 1$ on ζ. We place a truncated normal prior $\mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Omega_N)$ on θ independently of ζ, τ and λ, where the prior scale matrix is defined as $\Omega_N = (\Omega_{jj'})$ with $\Omega_{jj'} = k(u_j - u_{j'}), u_j = j/(N-1), j = 0, 1, \ldots, N-1$, and $k(\cdot)$ is the stationary Matérn kernel with smoothness parameter $\nu > 0$ and length-scale parameter $\ell > 0$. To implement the model, we place improper prior $\pi(\sigma^2) \propto 1/\sigma^2$ on σ^2.

To conclude this section, we rigorously justify the de-biasing property of the proposed procedure by examining the marginal posterior probability over the interval $(0, \delta)$ of ξ_1 for any fixed $\delta > 0$. For brevity and consistency, we assume observations $Y \sim \mathcal{N}(\Psi\theta_0, \mathrm{I}_n)$ with $\theta_0 \equiv \mathbf{0}_N$, and consider model (4.2) with $\zeta = 0$ and $\sigma = 1$. Theorem 7 asserts that the mass-shifting phenomenon of marginal posterior distribution is completely alleviated by the proposed prior. The contrast with the conclusion of Theorem 5 is immediately apparent.

Theorem 7. Suppose Assumptions 1 and 2 hold for the basis matrix Ψ and prior scale matrix Ω_N separately. For ξ defined in equation (4.1) with $\lambda_j \overset{\text{i.i.d.}}{\sim} \mathcal{C}_+(0,1)$, choose $\tau_n \asymp n^{-(1+\alpha)}$ for some constant $\alpha > 0$, then for any fixed $\delta > 0$, $\mathbb{E}_0\, \Pi(0 < \xi_1 < \delta|\tau_n, Y) \to 1$, a.s. as $n, N \to \infty$.

4.2. Asymptotic Properties

In this section, we study the asymptotic properties of the resulting posterior distribution, when the underlying true function is allowed to be flat in certain regions. To that end, we consider the following class of continuously differentiable functions

$$\mathcal{F}_+ := \big\{f \in \mathcal{C}[0,1] : f(0) = 0,\, f'(x) > 0\, \forall\, x \in [0, r_0], f'(x) = 0 \\ \forall\, x \in (r_0, 1],\ \text{for some } r_0 \in [0,1],\ \text{and} \\ f'(x)\ \text{is Lipschitz continuous}\big\}.$$

$\mathcal{F}_+$ includes nondecreasing functions which are strictly increasing followed by a flat region, the length of which is controlled by r_0, which is unknown. Specifically, $r_0 = 1$ implies a strictly increasing function and $r_0 = 0$ implies $f \equiv 0$. Assume the true function $f_0 \in \mathcal{F}_+$. Given n observed covariates $\{x_i\}$, we assume the observations are drawn independently from the true model by $Y_i \sim \mathcal{N}(f_0(x_i), \sigma_0^2)$ for $i = 1, \ldots, n$, with the noise level σ_0. Next, we consider a simplified setting of our model (4.2) by dropping the intercept term

$$Y = \Psi\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathrm{I}_n), \tag{4.3}$$

where (θ, σ^2) are the unknown parameters of interest. In general, the model (4.3) can be misspecified for estimating $f_0 \in \mathcal{F}_+$ and hence we shall investigate posterior convergence around the pseudo-true parameter by θ_0, which is obtained by minimizing the Kullback–Leibler divergence between the true data generating distribution and the model. In many cases, it is possible to quantify the gap between the true function and the pseudo-true parameter. For instance, if one adopts the basis (M) in (A.1) to construct Ψ in (4.3), $\{\theta_{0_j}\}$ can be obtained by evaluating f_0 at equally spaced grid points around which the compactly supported triangular basis functions are supported. In this case, we can show $||f_0 - \Psi\theta_0||_\infty \lesssim N^{-1}$ (refer to Lemma 1 of Appendix A) where N denotes the number of basis functions used. For any function f defined on some $\mathcal{X} \subset \mathbb{R}$, we denote $||f||_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

Using the *equivalence property* of basis (M) (refer to Appendix A and Lemma 3 therein), one can convert the constraint $f_0 \in \mathcal{F}_+$ equivalently to the following constraint set $\mathcal{C}_0$ of θ_0,

$$\mathcal{C}_0 := \left\{\theta \in \mathbb{R}^N : \theta_j > 0,\ j \in S_0;\ \theta_j = 0,\ j \in S_0^c,\ S_0 = \{1, \ldots, s_0\}\right\},$$

for some integer $0 < s_0 \leq N$. The set $\mathcal{C}_0$ then contains the indices of all nonzero coordinates corresponding to the increasing portion of f_0. Thus, $f_0 \in \mathcal{F}_+$ is equivalent to $\theta_0 \in \mathcal{C}_0$. Representing $\theta_0 = [\theta_{0_{S_0}}, \mathbf{0}_{S_0^c}]$, the presence of a flat region of f_0 can be equivalently expressed in terms of sparsity of the pseudo-true parameter θ_0. Without the knowledge of the true flatness, our proposed DGL-TMVN prior incorporates the constraint $\theta \in \mathcal{C} = [0, \infty)^N$ through the following hierarchical representation

$$\theta|\tau, \Lambda \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \tau^2 \Lambda\Omega_N\Lambda), \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N), \tag{4.4}$$

$$\lambda_j \overset{\text{iid}}{\sim} \mathcal{C}_+(0, 1), \quad \sigma^2 \sim \mathrm{IG}(a_0, b_0), \tag{4.5}$$

where τ is a global parameter to be chosen later, and (a_0, b_0) are the shape and rate parameters of the inverse-Gamma prior. An application of Bayes' theorem with (4.3) as the likelihood and (4.4)–(4.5) as the prior leads to the posterior distribution of (θ, σ^2) given data Y denoted by $\Pi(\cdot \mid Y)$ and can be expressed as $\Pi\{(\theta, \sigma^2) \in B|Y\} = \int_B P_{n,\sigma}(Y) d\Pi(\theta|\lambda) d\Pi(\lambda) d\Pi(\sigma^2) / \int P_{n,\sigma}(Y) d\Pi(\theta|\lambda) d\Pi(\lambda) d\Pi(\sigma^2)$, where the likelihood is defined as $P_{n,\sigma}(Y) = (2\pi\sigma^2)^{-n/2} \exp\{-\|Y - \Psi\theta\|^2 / 2\sigma^2\}$ and B is a Borel subset of $\mathbb{R}^{N+1}$. In contrast to a regular truncated multivariate normal prior, the marginal distributions of the proposed prior are designed to assign high probability near the origin by letting the global parameter τ to be sufficiently small, thus mitigating the mass-shifting phenomenon associated with the truly insignificant coefficients. On the other hand, the heavy-tailed prior for λ_j's combined with the dependence across the coordinates result in a good estimation of the nonzero coordinates, while ensuring smoothness.

We now state assumptions on the basis matrix $\Psi^{\mathrm{T}}\Psi$ and the pseudo-true parameter θ_0.

(A1) Assume $N = o(n)$, and assume the number of nonzero coordinates s_0 satisfies $s_0 \leq N$ and $s_0 N \log N \lesssim n$.

(A2) Assume for any nonempty subset $S \subset \{1, \ldots, N\}$, there exist constants $0 < k_1 < k_2 < \infty$ such that $k_1(n/N) \leq \lambda_{\min}(\Psi_S^{\mathrm{T}}\Psi_S) \leq \lambda_{\max}(\Psi_S^{\mathrm{T}}\Psi_S) \leq k_2(n/N)$, where Ψ_S is the $n \times |S|$ sub-matrix of Ψ with columns $\{\Psi_j : j \in S\}$.

(A3) Assume the pseudo-true parameter θ_0 satisfies $\max_{j\in S_0}\{|\theta_{0_j}|\} \leq cE_n$, where $c \in (0, 1)$ and E_n is a positive nondecreasing sequence.

Assumptions (A1)–(A3) are commonly assumed for proving optimal recovery results in high dimensional linear models. Assumption (A1) restricts the number of nonzero coefficients $|S_0| = s_0 \lesssim n/(N \log N)$, corresponding to the increasing portion of the true function. However, observe that we refrain from assuming that $s_0 = o(N)$. Akin to the high-dimensional linear regression setting (Narisetty and He 2014; Song and Liang 2017), Assumption (A2) ensures local invertibility over arbitrary directions of the basis matrix and implies $||\Psi^{\mathrm{T}}\Psi|| \lesssim n/N$. Assumption (A2) coincides with the restricted isometry property (Candes, Romberg, and Tao 2006), characterizing the near-orthonormality of the basis matrix $\Psi_S^{\mathrm{T}}\Psi_S$, for any nonempty set S, up to a scaling factor of value $(k_1 + k_2)n/N$. This assumption holds for a wide variety of basis functions, for example, B-Splines (refer to Lemma 8.9 of Yoo and Ghosal (2016) with a mild modification) and the considered basis (M) in (A.1) (see Lemma 2 in Appendix A). A similar assumption over B-Spline basis matrix can also be found in Bai et al. (2020). Assumption (A3) is commonly used on the growth of nonzero coefficients considered in optimal recovery in sparse (multivariate) linear models, refer to Song and Liang (2017), Chakraborty, Bhattacharya, and Mallick (2020), and Wei and Ghosal (2020). We next state the posterior contraction results regarding the parameter recovery and prediction.

Theorem 8. Let $\epsilon_n \asymp \max\{\sqrt{s_0 \log n/n}, 1/N\}$. Suppose Assumptions (A1)–(A3) hold and consider the prior on (θ, σ^2) defined in Equations (4.4) and (4.5). If $\tau \asymp n^{-(1+\alpha)}$ for some constant $\alpha > 0$, then

$$\sup_{f_0\in\mathcal{F}_+} \mathbb{E}_{f_0}\, \Pi\big(\theta : \|\theta - \theta_0\| \geq M_1\sqrt{N}\epsilon_n|Y\big) \to 0 \quad \text{a.s. as } n, N \to \infty, \tag{4.6}$$

$$\sup_{f_0\in\mathcal{F}_+} \mathbb{E}_{f_0}\, \Pi\big(f : \|f - f_0\| \geq M_2\sqrt{n}\epsilon_n|Y\big) \to 0 \quad \text{a.s. as } n, N \to \infty, \tag{4.7}$$

for positive constants M_1, M_2, large enough.

Remark 5. For $s_0 = O(N)$, one may obtain the best rate $\epsilon_n \asymp (n/\log n)^{-1/3}$ by choosing $N \asymp (n/\log n)^{1/3}$ and $s_0 = [\alpha_0 N]$ in the expression for ϵ_n, for some fixed constant $\alpha_0 \in (0, 1]$.

The proof of Theorem 8 extends existing results (Pati et al. 2014; Chakraborty, Bhattacharya, and Mallick 2020) on prior concentration for independent global-local shrinkage priors to its present dependent counterpart, which may be of independent interest. We also adapt testing arguments from high-dimensional regression problems (Song and Liang 2017; Wei and Ghosal 2020) to the present setup. Theorem 8 holds uniformly for all functions in $\mathcal{F}_+$. As the procedure does not require the knowledge of r_0, the proposed model can successfully recover flat regions in the true function. When $s_0 = O(N)$, the posterior contracts at a near minimax rate for isotonic regression problems; refer to Van der Vaart (2000), Chatterjee, Guntuboyina, and Sen (2015), Gao, Han, and Zhang (2020), and Chakraborty and Ghosal (2021) for a Bayesian counterpart. In a sparse linear regression setting with covariate matrix X, the eigen values of $X_S^{\mathrm{T}} X_S$ are typically allowed to grow at $O(n)$ (refer to Theorem 2 in Castillo et al. (2015)) for optimal recovery. On the other hand, to reflect the inherent smoothness, the eigenvalues of the basis matrix $\Psi_S^{\mathrm{T}} \Psi_S$ are assumed to grow as $O(n/N)$ in Assumption (A2). This accounts for the extra factor $\sqrt{N}$ in the posterior contraction rate. It is easy to see that the obtained rate matches the Theorem of Castillo et al. (2015) by letting $\max_i\{(X^{\mathrm{T}}X)_{ii}\} \asymp n/N$ which is the best possible attainable rate in a minimax sense. Overall, our result provides a framework for obtaining optimal posterior contraction using a dependent global-local shrinkage prior, which can be more broadly relevant.

Next, define the posterior mean of parameter as $\widehat{\theta} = \int \theta\, \Pi(\theta|Y)\, d\theta$ and consider the Bayes estimate $\widehat{f} = \Psi\widehat{\theta}$. The optimality of the posterior contraction rate in Theorem 8 implies, as a byproduct (Castillo and van der Vaart 2012), that the posterior mean converges at the same rate as sample size goes to infinity, in contrast to the observed bias for truncated multivariate normal priors in the right panel of Figure 1, and Figure S2 in the supplementary materials. The result is summarized in the following Corollary.

Corollary 9 (Posterior mean). Under the conditions of Theorem 8 and for the ϵ_n defined in Theorem 8, we have $\sup_{f_0 \in \mathcal{F}_+} \mathbb{E}_{f_0} ||\widehat{f} - f_0||^2 \lesssim n\epsilon_n^2$.

4.3. Empirical Illustrations

In this section, we discuss the efficacy of our proposed debiasing approach based on the DGL-TMVN prior and compare its prediction performance with other existing methods, such as the model based on an independent global local shrinkage prior (by setting the scale matrix to be an identity matrix). The data is generated from (3.1) with true $\sigma = 0.5$ and four different choices of the true f, namely,

$$f_1(x) = (5x-3)^3\, \mathbb{1}_{[0.6,1]}(x), \quad f_2(x) = \frac{3}{1+\exp(-10x+2.1)},$$
$$f_3(x) = \sqrt{2}\sum_{l=1}^{100} l^{-1.7}\, \sin(l)\, \cos(\pi(l-0.5)(1-x)), \quad f_4(x) = 5x^2,$$

for $x \in [0,1]$. The function f_1, which is nondecreasing and flat between 0 and 0.6, was used as the motivating example in the introduction. The functions f_2 and f_3 are both approximately flat between 0.7 and 1. In particular, f_3 is decreasing in certain regions which allows us to evaluate the performance of the proposed model under slight model misspecification. Finally, f_4 is considered for testing the performance of the proposed method in recovering strictly monotone functions.

We used the same model set up and prior specifications described in Section 4.1 with $k(\cdot)$ as the stationary Matérn kernel with smoothness parameter $\nu > 0$ and length-scale parameter $\ell > 0$. In the comparisons below when the hyperparameters are not fixed, we place compactly supported priors $\nu \sim \mathcal{U}(0.5,1)$ and $\ell \sim \mathcal{U}(0.1,1)$ on ν and ℓ. The hyperprior and covariance kernel choices are made with utmost care; detailed justifications are deferred to Section S8.5 of the supplementary materials along with the sensitivity study results on the model robustness to different covariance kernels and mild variations of hyperpriors. We also develop a data-augmentation Gibbs sampler which combined with the embedding technique of Ray, Pati, and Bhattacharya (2020) results in an efficient MCMC algorithm to sample from the joint posterior of $(\zeta, \theta, \lambda, \sigma^2, \tau^2, \nu, \ell)$; the details are in Section S8.2 of the supplementary materials.

First, we discuss the improvement due to the shrinkage. We consider a sequence of priors, becoming progressively complex, beginning with a truncated normal prior (TMVN) and gradually adding more structure to eventually arrive at the proposed shrinkage prior (DGL-TMVN). Specifically, four variants of TMVN priors are compared, a detailed elaboration is deferred to Section S8.3 of the supplementary materials. We generate 500 pairs of response and covariates and randomly divide the data into 300 training samples and 200 test samples. For all of the variants above, we set the number of knots $N = 150$. We provide plots of the function fit for four functions along with pointwise 95% credible intervals in Figures S9– S12 in the supplementary materials, and also report the mean squared prediction error (MSPE) at the bottom of the sub-plots. The results show that DGL-TMVN performs the best, both visually and also in terms of MSPE.

We now focus on the performance of the DGL-TMVN prior against that of some potential competitors. Three different priors are compared: our proposed DGL-TMVN prior (DGL for short), the TMVN prior with updating hyperparameters and incorporated with an independent global-local shrinkage prior (IGL for short), which can be considered as the continuous version of the independent univariate point-mass mixture priors, and the TMVN prior with the global shrinkage where we consider the prior on τ as $\pi(\tau^2) \propto 1/\tau^2$ (TMVN for short). The simulation studies are conducted over 25 replicated datasets of size 500 which are randomly split into training set of size 300 and test set of size 200 for each function under the same setting as the previous cascading analysis. For each replicate we run a Gibbs sampler of 15,000 iterations with the first 5000 discarded as burn-in. To compare the performance specifically for the flat region and for the increasing region separately, in addition to the average MSPE, we report the average partial MSPEs corresponding to the flat portion (MSPE flat) and to the increasing portion (MSPE incr.), respectively. Finally, we look into the average coverage probability over the true function to evaluate the concentration of resulting posterior distribution toward the true function. Out-of-sample prediction results on the test data are

Table 1. Results of three methods over test samples for f_1, f_2, f_3, and f_4.

Function	Method	MSPE (total)	MSPE (flat)	MSPE (incr.)	Coverage
f_1	dgl	**11.36**(2.62)	**8.13**(1.95)	**14.71**(4.79)	**0.715**
	igl	13.44(2.62)	9.86(1.70)	17.32(5.23)	0.651
	tmvn	65.63(7.21)	14.53(2.59)	102.6(11.16)	0.391
f_2	dgl	**8.29**(1.78)	**7.13**(2.64)	8.56(2.32)	**0.887**
	igl	9.55(1.92)	8.40(2.61)	9.84(2.54)	0.856
	tmvn	8.32(2.11)	8.61(2.91)	**7.94**(2.75)	0.793
f_3	dgl	7.76(1.74)	9.16(2.9)	**6.87**(1.87)	0.918
	igl	**7.72**(1.74)	**8.57**(2.45)	7.18(1.74)	**0.946**
	tmvn	11.36(1.33)	15.27(2.85)	8.97(1.76)	0.765
f_4	dgl	8.67(2.15)	–	8.67(2.15)	0.952
	igl	9.34(2.16)	–	9.34(2.16)	0.969
	tmvn	**5.68**(1.61)	–	**5.68**(1.61)	**0.979**

NOTE: The number of knots $N = \lfloor n/8 \rfloor$. All types of MSPEs $\times 10^2$ (standard deviations $\times 10^2$) and the coverage are averaged over 25 replicates. The best reported result of corresponding error measure in each column is highlighted with bold font.

summarized in Table 1. The plots of the function fit are displayed with zoomed-in inset plots over the flat region in Figures 2 and 3 for functions f_1, f_3, in Figures S5 and S6 in Section S2.3 of the supplementary materials for the functions f_2, f_4, respectively.

Results in Table 1 show employing any shrinkage procedure with truncated normal priors improves the prediction accuracy significantly in terms of all types of MSPEs and the coverage for all four functions, providing strong support on the de-biasing properties of the shrinkage procedures. When the model is not misspecified, as fitting f_1 and f_2, the DGL prior obtains the smallest total MSPEs and that for flat region along with the higher posterior coverage than the IGL prior. This result also agrees with the zoomed-in inset plots in Figures 2 and S5 which show smaller biases are induced by the DGL over the flat regions of f_1 and f_2. For the purpose of prediction, the dependent shrinkage prior performs slightly better than its independent counterpart even the true coefficients are (nearly) zero. The enforced independence in the IGL results in a loss of the necessary amount of smoothness which is amplified when estimating the increasing region of f_1, f_2 where a higher level of dependence among the coefficients is required. We notice that the regular TMVN obtains the lowest MSPE in estimating the increasing region of f_2, as the original correlation structure of the truncated normal prior is adequate to model smooth and strictly increasing functions. Based on the simulation results, it is evident that the DGL prior provides the best tradeoff between estimating the increasing region and estimating the flat region among three priors.

Simulation results for f_3 warrant some attention. When the true function contains some discontinuities such as f_3, the independent shrinkage prior is expected to be favorable for estimating the discontinuous area. This is supported by the empirical observations that the IGL obtains the lowest MSPE overall and for the flat region as well as the highest coverage. The zoomed-in inset plot of IGL in Figure 3 also shows it captures the trend of the true curve with a small shift. However, when a dependent truncated normal prior is employed with a global and local shrinkage procedure the performance is not compromised significantly. The total MSPE of the DGL is very close to that of the IGL and the coverage is comparably good. Figure 3 shows the estimation over the flat region of the DGL is also similar to the IGL and its 95% credible interval contains the true function. We notice that the prediction of the DGL is more biased compared to that of the IGL over the region where the slope of true curve changes drastically. We attribute this partially to the independent local shrinkage parameter corresponding to the significant components weaken the global shrinkage on the correlation between significant and the insignificant coefficients. This observation motivates a variant of DGL prior which is discussed in Section 4.4. Finally, results for f_4 provide additional evidence that performance of DGL is not deteriorated when used to estimate a strictly increasing underlying function. In this case, both the error measures and out-of-sample prediction plots in Figure S6 in Section S2.3 of the supplementary materials indicate that DGL performs comparably to TMVN, and results in smoother predictions compared to IGL.

Additional comparison of the proposed DGL-TMVN prior has been conducted to a very recent state-of-the-art method named `bsar`, which is developed by Lenk and Choi (2017) and implemented in the R package `bsamGP`. For the out-of-sample prediction performance of `bsar`, refer to Figure S13 in Section S8.4 of the supplementary material, based on which it is clear that the performance of global-local shrinkage procedure is comparable with that of `bsar`. It is important to point out that bsar is also a shrinkage based method that allows for exact zeros in the coefficients in a transformed Gaussian process prior through a spike and slab specification. Based on the above empirical studies, we conclude DGL prior is most robust for estimating various nondecreasing functions with potential flat regions compared to other priors under consideration.

We conclude this section with a brief discussion on the mixing behavior and computational efficiency of the Gibbs samplers based on the compared priors. Boxplots of the effective sample sizes (ESS) of the MCMC samples of estimated function values based on 200 test points are displayed in Figure S22 in Section S8.6 of the supplementary materials for each function and model. The reported ESS values are averages over 25 replicates. Moreover, the averaged Monte Carlo standard errors (MCSE) in estimating function values over test points for each prior are compared with the averaged standard deviation of test response samples over 25 replicates in Table S2 of Section S8.6 of the supplementary materials. All diagnostic measurements indicate the MCMC chains of each model discussed here converged well. The computing time of running the Gibbs sampler of 15,000 iterations under the same setting for one replicated dataset of the function f_1 based on DGL, IGL and TMVN are 39.26, 13.68, and 47.72 sec on a machine with a 8-Core processor and 32GB RAM. MCMC algorithms related to DGL and TMVN are efficient due to implementing the embedding technique of Ray, Pati, and Bhattacharya (2020) that avoids drawing samples from truncated multivariate normal distribution.

4.4. *A Variant of the* dgl-tmvn *Prior*

As discussed in previous section, the amount of shrinkage on the correlation between the significant and the nonsignificant coefficients is slightly weakened by the local shrinkage parameters corresponding to the significant coefficients. This motivates us to place $\tau \sim \mathcal{C}_+(0, \delta_\tau)$, where $\delta_\tau < 1$ is the scale parameter for the half-Cauchy prior, to facilitate stronger global shrinkage. Such modification will not affect the local shrinkage parameters significantly as it has been shown that a default Cauchy

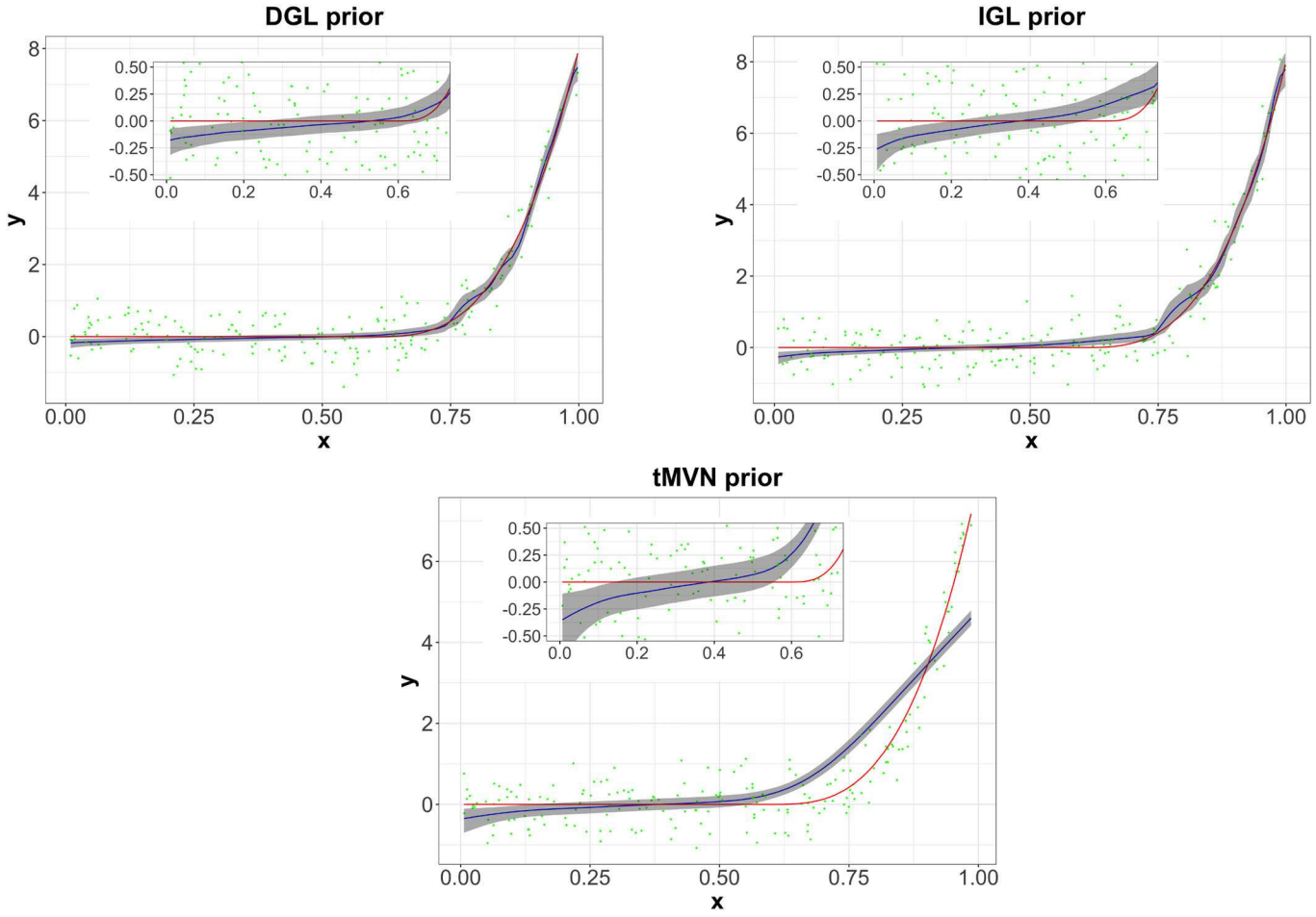

Figure 2. Out-of-sample prediction on f_1 with zoomed-in inset plot for $x \in [0, 0.6]$. The number of knots $N = \lfloor n/8 \rfloor$. Green points are the test samples, the red curve is the true function, the blue curve is the posterior mean and the gray shaded area is the 95% pointwise prediction interval.

prior on λ_js ensures that the global shrinkage on significant coefficients tends to be negligible regardless of how small the global parameter is (Polson and Scott 2010). Based on these properties, we expect this version of DGL helps recover the flat region in more complicated scenarios while the estimation over the increasing region remains unaffected. Based on numerous experimentations on the choice of δ_τ we find that $\delta_\tau = 0.5$ provides desired level of global shrinkage. On the contrary, choosing $\delta_\tau > 1$ imposes weaker global shrinkage and allows the prior to enforce stronger dependence among coefficients. For estimating f_4, DGL works slightly better with $\delta_\tau > 1$ than that with a smaller δ_τ. These observations imply the model might be sensitive to hyperprior choices for local-global shrinkage parameters. However, sensitivity studies show the DGL model is robust as long as δ_τ is chosen within a reasonable range. For more details, refer to Section S8.5 of the supplementary materials. A thorough investigation on the theoretically properties of this version of DGL-TMVN is considered as the future work.

5. Application on Real Datasets

In this section, we provide performance illustrations of DGL-TMVN implemented on real-life datasets. We applied the proposed approach to analyze two datasets where either the underlying true function is known to be monotone with a flat region or the data indicate a monotone pattern with certain flat regions. Since there is no strong indication that the DGL prior is sensitive to the choice of covariance kernel function based on the sensitivity study in Section S8.5 of the supplementary materials, we use the Matérn kernel function and we resort to the same model and prior specifications as described in Section 4.3. We used the version of DGL prior described in Section 4.4 to ensure better detection of the monotone trend in addition to capturing the flat region.

We compared the model performance based on DGL and TMVN priors based on model selection criteria such as the Watanabe-Akaike information criteria (WAIC) in addition to visual representations. For datasets which appear to contain certain flat regions, we compare the WAIC values of models fitting the observed flat region as well. We ran our sampler for 15,000 MCMC iterations, first 5000 of which were discarded as burn-in, and every 10th subsequent observation was stored. The same was done for the Gibbs sampler with the TMVN prior. Details on the mixing behavior of MCMC samples of the two models are deferred to Section S8.6 of the supplementary materials. Similar to Section 4.3, the boxplots of ESS and averaged MCSE are reported for each dataset and each model. Results indicate that the MCMC chains for both models converged.

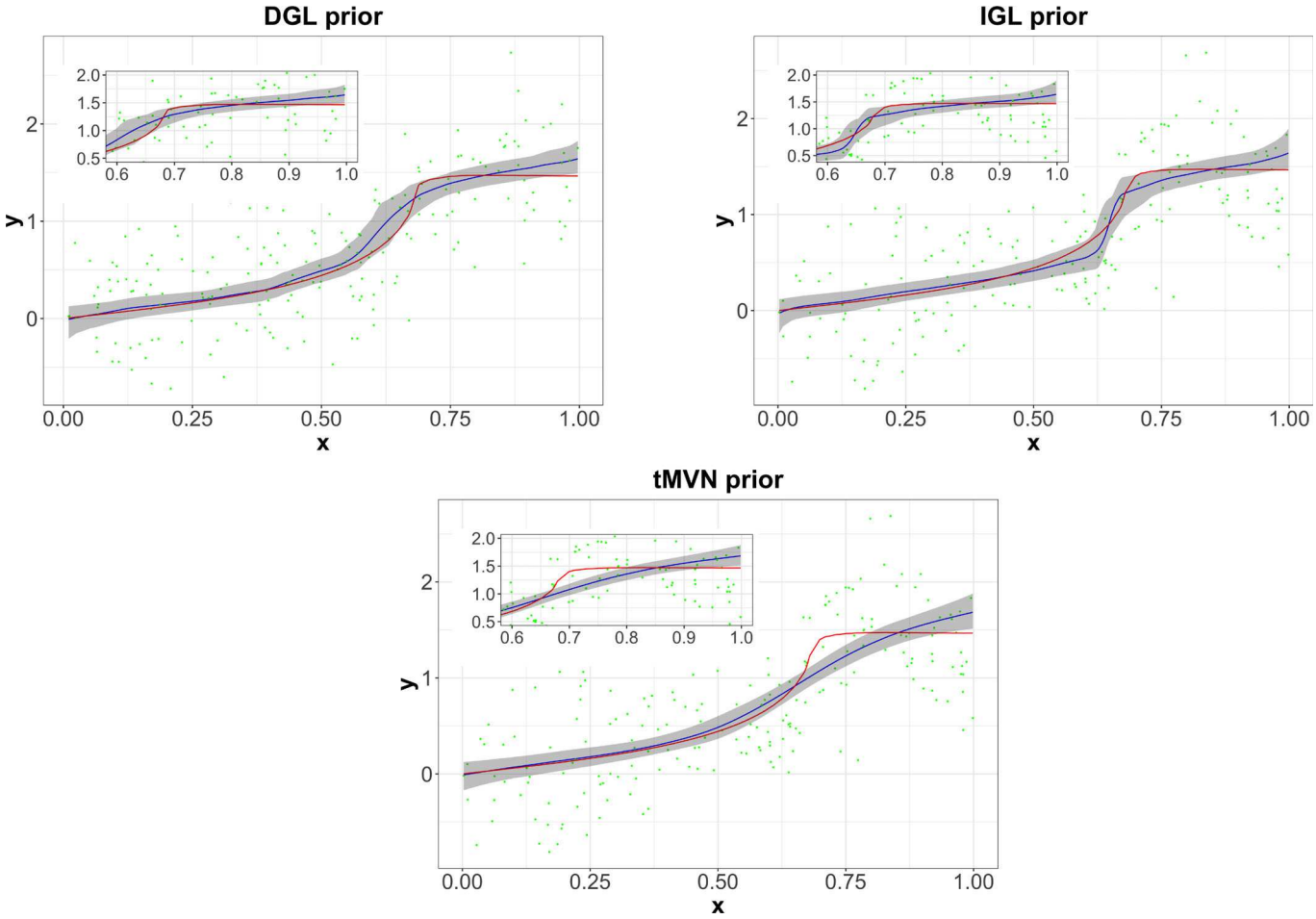

Figure 3. Same as Figure 2 for f_3 with zoomed-in inset plots where $x \in [0.6, 1]$.

5.1. Age and Income Data

We use the age and income data that consist of `age` (in years) and the logarithm of income (`log.income`) on 205 Canadian workers from a 1971 Canadian Census Public Use Tape. This dataset is readily available for public use and accessible through the R package `SemiPar`. Our goal is to estimate logarithm of income as function of age. Data suggest that the true underlying function is monotone nondecreasing with a flat region. We provide performance illustrations of the models with DGL and TMVN priors along with WAIC values for overall fitting and for fitting the observed flat region only. The reported WAIC values in Figure 4 suggest that the model with DGL fits the data better than that with TMVN. This result is also consistent with the model fit result shown in Figure 4. The predictive curve of the model with DGL aligns with the data points well, while the model with TMVN appears to fail to capture the trend for `age` $\leq$ 26 and induces a large bias. For the region with `age` $\geq$ 26, although the fitted curves of both models are approximately flat and similar, we find that the DGL provides a smaller value 274.147 of the WAIC for modeling this region only, compared to a WAIC value of 284.259 for the model with TMVN. This provides some evidence that the model with DGL is more reliable to fit the data over the "flat" region.

5.2. Light Detection and Ranging Data

The light detection and ranging (LiDAR) data have 221 observations from a LiDAR experiment and it contain information on `range` and `logratio`. The predictor `range` is the distance traveled before the light is reflected back to its source and the response variable `logratio` is the logarithm of the ratio of received light from two laser sources. This data is obtained from the R package `HRW`. The data suggest that the true underlying function is monotone nonincreasing with a flat region. Similar to previous analyses, the overall WAIC value of the model with DGL is lower than the model with TMVN, indicating the DGL prior is more appropriate for fitting the LiDAR data. Figure 5 shows the model with TMVN captures the overall trend of the data, however, the predictive curve of DGL seems to align with the data better, specifically over the "flat" region and over the region where the value of `logratio` starts to decrease. To confirm this, we compared the WAIC values of both models for fitting the region with `range` $\leq$ 550. The model with DGL obtains a smaller WAIC value of -328.563 against the value of -310.736 for the model with TMVN. In this sense, the prediction of the model with DGL seems to capture the dynamics between the range and the received lights, for instance, finding a threshold value of the range that begins to affect the ratio of received lights from two sources.

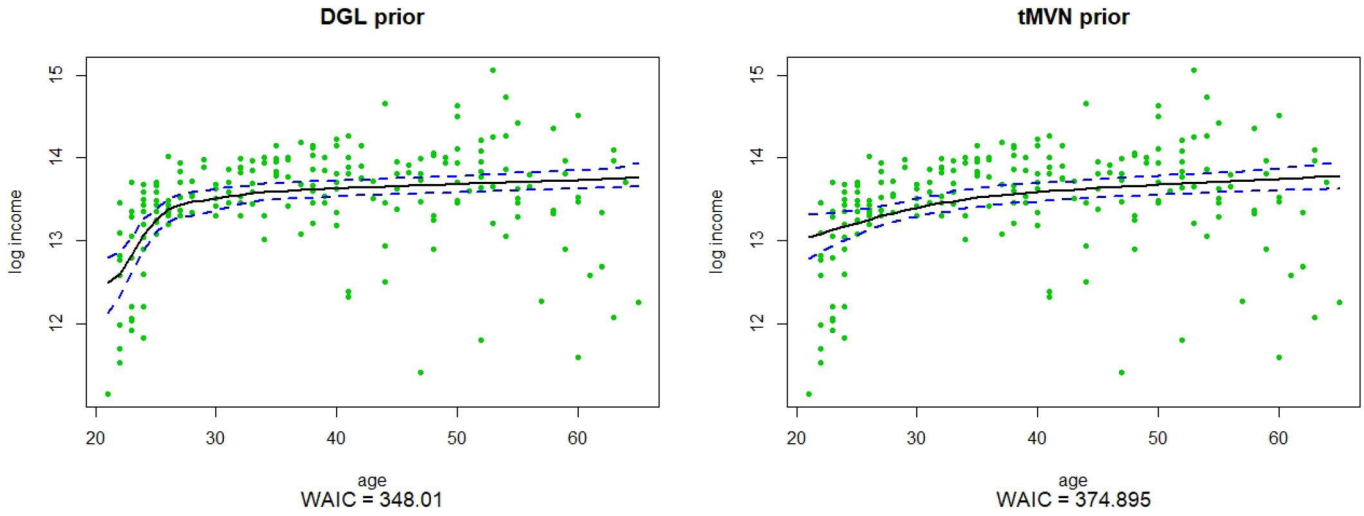

Figure 4. Estimation accuracy of the two competing methods applied on the dysphoria score data. The black solid curve is the posterior mean, the region within two dotted blue curves represent 95% pointwise credible interval and the green dots are the observed data points. The WAIC values corresponding to the methods are shown in the sub-plots.

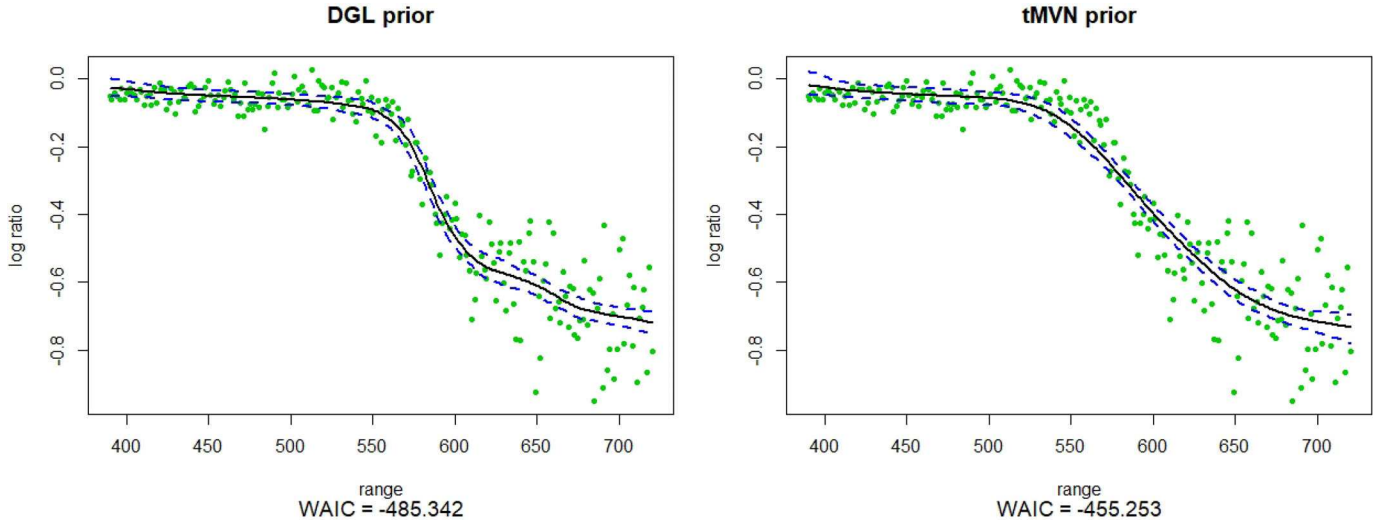

Figure 5. Same as Figure 4 for the LiDAR data.

6. Discussion

A seemingly natural way to define a prior distribution on a constrained parameter space is to consider the restriction of a standard unrestricted prior to the constrained space. The conjugacy properties of the unrestricted prior typically carry over to the restricted case, facilitating computation. Moreover, reference priors on constrained parameters are typically the unconstrained reference prior multiplied by the indicator of the constrained parameter space (Sun and Berger 1998). Despite these various attractive properties, the findings of this article pose a caveat toward routine truncation of priors in moderate to high-dimensional parameter spaces, which might lead to biased inference. This issue gets increasingly severe with increasing dimension due to the concentration of measure phenomenon (Talagrand 1995; Boucheron, Lugosi, and Massart 2013), which forces the prior to increasingly concentrate away from statistically relevant portions of the parameter space. A somewhat related issue with certain high-dimensional shrinkage priors has been noted in Bhattacharya et al. (2016). Overall, our results suggest a careful study of the geometry of truncated priors as a useful practice. Understanding the cause of the biased behavior also suggests natural shrinkage procedures that can guard against such unintended consequences. We note that post-processing approaches based on projection (Lin and Dunson 2014; Sen, Patra, and Dunson 2018; Chakraborty and Ghosal 2021) and constraint relaxation (Duan et al. 2020) do not suffer from this unintended bias. The same is also true for the recently proposed monotone BART (Bayesian Additive Regression Trees) method (Chipman, George, and McCulloch 2010). It would be interesting to explore the presence of similar issues arising from truncations beyond the constrained regression setting. Possible examples include correlation matrix estimation and simultaneous quantile regression. Priors on correlation matrices are often prescribed in terms of constrained priors on scale matrices, and truncated normal priors are used to maintain ordering between quantile functions corresponding

to different quantiles, and this might leave the door open for unintended bias to creep in.

Appendix A: Basis Representation of Maatouk and Bay (2017)

As our example which motivates the main results of this article, we consider the more recent basis sequence of Maatouk and Bay (2017). Let $u_j = j/(N-1), j = 0, 1, \ldots, N-1$ be equally spaced points on $[0,1]$, with spacing $\delta_N = 1/(N-1)$. Let,

$$h_j(x) = h\left(\frac{x-u_j}{\delta_N}\right), \quad \psi_j(x) = \int_0^x h_j(t)\,dt,$$
$$\phi_j(x) = \int_0^x \int_0^t h_j(u)\,dudt,$$

for $j = 0, 1, \ldots, N-1$, where $h(x) = (1-|x|)\,\mathbb{1}_{[-1,1]}(x)$ is the "hat function" on $[-1,1]$. For any continuous function $f : [0,1] \to \mathbb{R}$, the function $\widetilde{f}(\cdot) = \sum_{j=0}^{N-1} f(u_j)\,h_j(\cdot)$ approximates f by linearly interpolating between the function values at the knots $\{u_j\}$, with the quality of the approximation improving with increasing N. With no additional smoothness assumption, this suggests a model for f as $f(\cdot) = \sum_{j=0}^{N-1} \theta_{j+1} h_j(\cdot)$. The basis $\{\psi_j\}$ and $\{\phi_j\}$ take advantage of higher-order smoothness. If f is once or twice continuously differentiable, respectively, then by the fundamental theorem of calculus,

$$f(x) - f(0) = \int_0^x f'(t)dt, \quad f(x) - f(0) - xf'(0) = \int_0^x \int_0^t f''(s)\,dsdt.$$

Expanding f' and f'' in the interpolation basis as in the previous paragraph, respectively, imply the models

$$\underbrace{f(x) = \theta_0 + \sum_{j=0}^{N-1} \theta_{j+1}\psi_j(x),}_{M} \quad \underbrace{f(x) = \theta_0 + \theta^* x + \sum_{j=0}^{N-1} \theta_{j+1}\phi_j(x)\,.}_{C} \tag{A.1}$$

Under the above, the coefficients have a natural interpretation as evaluations of the function or its derivatives at the grid points. For example, under (M), $f'(u_j) = \theta_{j+1}$ for $j = 0, 1, \ldots, N-1$, while under (C), $f''(u_j) = \theta_{j+1}$ for $j = 0, 1, \ldots, N-1$. We provide an approximation result of such basis expansion to a regular differentiable function $f \in \mathcal{C}[0,1]$ in the following lemma.

Lemma 1. For any $f \in \mathcal{C}[0,1]$ and f' is Lipcshitz, for any integer $N > 1$ construct the model denoted by $f_N(\cdot)$ under (M) in (A.1), we have $\|f - f_N\|_\infty \lesssim 1/N$.

Given covariates $\{x_i\}$, construct an $n \times N$ basis matrix $\Psi = (\psi_j(x_i))$ with basis functions $\{\psi_j\}$, and for any nonempty subset of indexes $S \subset \{1, \ldots, N\}$, denote by Ψ_S the $n \times |S|$ sub-matrix with columns $\{\Psi_j : j \in S\}$. Next Lemma bounds eigenvalues of $\Psi_S^{\mathrm{T}}\Psi_S$ under mild conditions.

Lemma 2. For a grid $\{u_j\}$, assume the covariates $\{x_i\}$ satisfy that there exists a constant $c > 0$ such that $\min_{\{j:|x_i-u_j|>0\}}\{|x_i - u_j|\} \geq c\delta_N^{3/2}, i = 1, \ldots, n$. Then for any nonempty set $S \subset \{1, \ldots, N\}$, there exist constants $0 < m_1 < m_2 < \infty$ such that $m_1 n/N \leq \lambda_{\min}(\Psi_S^{\mathrm{T}}\Psi_S) \leq \lambda_{\max}(\Psi_S^{\mathrm{T}}\Psi_S) \leq m_2 n/N$.

(*Equivalence property.*) Maatouk and Bay (2017) showed that under the representation (M) in (A.1), f is monotone nondecreasing *if and only if* $\theta_i \geq 0$ for all $1 \leq i \leq N$. The flat region can be characterized by the corresponding basis coefficients, which is characterized by the following result.

Lemma 3. For any $0 \leq a < b \leq 1$ and some constant $c \in \mathbb{R}$, and for any $x \in [a,b], f(x) \equiv c$ if and only if $\theta_j = 0$, for $j \in S_{[a,b]}$, where $S_{[a,b]}$ is a subset of indexes such that $\cup_{j \in S_{[a,b]}}[u_j, u_{j+1}] \supset [a,b]$ is the shortest interval.

Similarly, under (C), f is convex nondecreasing *if and only if* $\theta_i \geq 0$ for all $i = 1, \ldots, N$. The ability to *equivalently* express various constraints in terms of linear restrictions on the vector $\theta = (\theta_1, \ldots, \theta_N)^{\mathrm{T}}$ is an attractive feature of this basis not necessarily shared by other basis. In either case, the parameter space $\mathcal{C}$ for θ is the nonnegative orthant $[0,\infty)^N$. If f were unrestricted, a GP prior on f would induce a dependent Gaussian prior on θ. The approach of Maatouk and Bay (2017) is to restrict this dependent prior subject to the linear restrictions, resulting in a truncated normal prior.

Supplementary Materials

The supplemental materials contain additional plots, proofs of main results in the manuscript, as well as remaining technical results and additional details and analyses on numerical studies. Code and data (sources) are also included that are needed to replicate the results presented in both simulations and real applications of the manuscript.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Pati and Bhattacharya acknowledge support from NSF DMS (1854731, 1916371). In addition, Bhattacharya acknowledges the NSF CAREER 1653404 award for supporting this project.

References

Azzalini, A., and Valle, A. D. (1996), "The Multivariate Skew-Normal Distribution," *Biometrika*, 83, 715–726. [3,4]

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020), "Spike-and-Slab Group LASSOs for Grouped Regression and Sparse Generalized Additive Models," *Journal of the American Statistical Association*, 117, 184–197. [8]

Bhattacharya, A., Dunson, D. B., Pati, D., and Pillai, N. S. (2016), "Sub-Optimality of Some Continuous Shrinkage Priors," *Stochastic Processes and their Applications*, 126, 3828–3842. [13]

Bickel, P., and Lindner, M. (2012), "Approximating the Inverse of Banded Matrices by Banded Matrices with Applications to Probability and Statistics," *Theory of Probability & Its Applications*, 56, 1–20. [5]

Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [6]

Bornkamp, B., and Ickstadt, K. (2009), "Bayesian Nonparametric Estimation of Continuous Monotone Functions with Applications to Dose–Response Analysis," *Biometrics*, 65, 198–205. [1]

Botev, Z. I. (2017), "The Normal Law under Linear Restrictions: Simulation and Estimation via Minimax Tilting," *Journal of the Royal Statistical Society*, Series B, 79, 125–148. [3]

Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press. [13]

Brezger, A., and Steiner, W. J. (2008), "Monotonic Regression based on Bayesian P–Splines: An Application to Estimating Price Response Functions from Store-Level Scanner Data," *Journal of Business & Economic Statistics*, 26, 90–104. [1]

Cai, B., and Dunson, D. B. (2007), "Bayesian Multivariate Isotonic Regression Splines: Applications to Carcinogenicity Studies," *Journal of the American Statistical Association*, 102, 1158–1171. [1]

Candes, E. J., Romberg, J. K., and Tao, T. (2006), "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Communications on*

Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59, 1207–1223. [8]

Cartinhour, J. (1990), "One-Dimensional Marginal Density Functions of a Truncated Multivariate Normal Density Function," *Communications in Statistics-Theory and Methods*, 19, 197–203. [3,4]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [3,7]

Castillo, I., and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *The Annals of Statistics*, 40, 2069–2101. [9]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian Linear Regression with Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [9]

Chakraborty, A., Bhattacharya, A., and Mallick, B. K. (2020), "Bayesian Sparse Multiple Regression for Simultaneous Rank Reduction and Variable Selection," *Biometrika*, 107, 205–221. [8,9]

Chakraborty, M., and Ghosal, S. (2021), "Convergence Rates for Bayesian Estimation and Testing in Monotone Regression," *Electronic Journal of Statistics*, 15, 3478–3503. [9,13]

Chatterjee, S., Guntuboyina, A., and Sen, B. (2015), "On Risk Bounds in Isotonic and other Shape Restricted Regression Problems," *The Annals of Statistics*, 43, 1774–1800. [9]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [13]

Curtis, S. M., and Ghosh, S. K. (2011), "A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials," *Journal of Applied Statistics*, 38, 961–976. [1,2,3,5,7]

Duan, L. L., Young, A. L., Nishimura, A., and Dunson, D. B. (2020), "Bayesian Constraint Relaxation," *Biometrika*, 107, 191–204. [13]

Dunson, D. B. (2005), "Bayesian Semiparametric Isotonic Regression for Count Data," *Journal of the American Statistical Association*, 100, 618–627. [7]

Gao, C., Han, F., and Zhang, C.-H. (2020), "On Estimation of Isotonic Piecewise Constant Signals," *The Annals of Statistics*, 48, 629–654. [9]

Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000), "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531. [2]

Lenk, P. J., and Choi, T. (2017), "Bayesian Analysis of Shape-Restricted Functions using Gaussian Process Priors," *Statistica Sinica*, 27, 43–69. [10]

Lin, L., and Dunson, D.B. (2014), "Bayesian Monotone Regression using Gaussian Process Projection," *Biometrika*, 101, 303–317. [13]

Maatouk, H., and Bay, X. (2017), "Gaussian Process Emulators for Computer Experiments with Inequality Constraints," *Mathematical Geosciences*, 49, 557–582. [1,2,5,6,14]

Meyer, M. C., Hackstadt, A. J., and Hoeting, J. A. (2011), "Bayesian Estimation and Inference for Generalised Partial Linear Models using Shape-Restricted Splines," *Journal of Nonparametric Statistics*, 23, 867–884. [1]

Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817. [8]

Neelon, B., and Dunson, D. B. (2004), "Bayesian Isotonic Regression and Trend Analysis," *Biometrics*, 60, 398–406. [1,2,3,7]

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014), "Posterior Contraction in Sparse Bayesian Factor Models for Massive Covariance Matrices," *The Annals of Statistics*, 42, 1102–1130. [9]

Polson, N. G., and Scott, J. G. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," *Bayesian Statistics*, 9, 501–538. [11]

——— (2012), "On the Half-Cauchy Prior for a Global Scale Parameter," *Bayesian Analysis*, 7, 887–902. [3]

Ray, P., Pati, D., and Bhattacharya, A. (2020), "Efficient Bayesian Shape-Restricted Function Estimation with Constrained Gaussian Process Priors," *Statistics and Computing*, 30, 839–853. [9,10]

Sen, D., Patra, S., and Dunson, D. (2018), "Constrained Inference through Posterior Projections," arXiv preprint arXiv:1812.05741. [13]

Shively, T. S., Walker, S. G., and Damien, P. (2011), "Nonparametric Function Estimation Subject to Monotonicity, Convexity and other Shape Constraints," *Journal of Econometrics*, 161, 166–181. [1]

Song, Q., and Liang, F. (2017), "Nearly Optimal Bayesian Shrinkage for High Dimensional Regression," arXiv preprint arXiv:1712.08964. [8,9]

Sun, D., and Berger, J. O. (1998), "Reference Priors with Partial Information," *Biometrika*, 85, 55–71. [13]

Talagrand, M. (1995), "Concentration of Measure and Isoperimetric Inequalities in Product Spaces," *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81, 73–205. [13]

Van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [9]

Wei, R., and Ghosal, S. (2020), "Contraction Properties of Shrinkage Priors in Logistic Regression," *Journal of Statistical Planning and Inference*, 207, 215–229. [8,9]

Yoo, W. W., and Ghosal, S. (2016), "Supremum Norm Posterior Contraction and Credible Sets for Nonparametric Multivariate Regression," *The Annals of Statistics*, 44, 1069–1102. [5,6,8]