Mass-shifting phenomenon of truncated multivariate normal priors

Shuang Zhou*[1], Pallavi Ray†[1], Debdeep Pati‡[1], and Anirban Bhattacharya§[1]

[1]Department of Statistics, Texas A&M University, College Station, Texas, 77843, USA

May 20, 2020

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. A precise quantification characterizing the role of the dimension as well as the dependence is provided. This surprising behavior has serious implications towards 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. 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 empirically shown to guard against the mass shifting phenomenon while retaining computational efficiency.

*shuang@stat.tamu.edu
†pallaviray@stat.tamu.edu
‡debdeep@stat.tamu.edu
§anirbanb@stat.tamu.edu

1 Introduction

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

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

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 non-asymptotic 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 1. 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 Neelon and Dunson [2004] and Curtis and Ghosh [2011], where truncated normal priors led to biased posterior inference when the underlying function had flat regions.

A common approach towards 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 et al., 2011], penalized spines [Brezger and Steiner, 2008], cumulative distribution functions [Bornkamp and Ickstadt, 2009], restricted splines [Shively et al., 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 non-negativity 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.

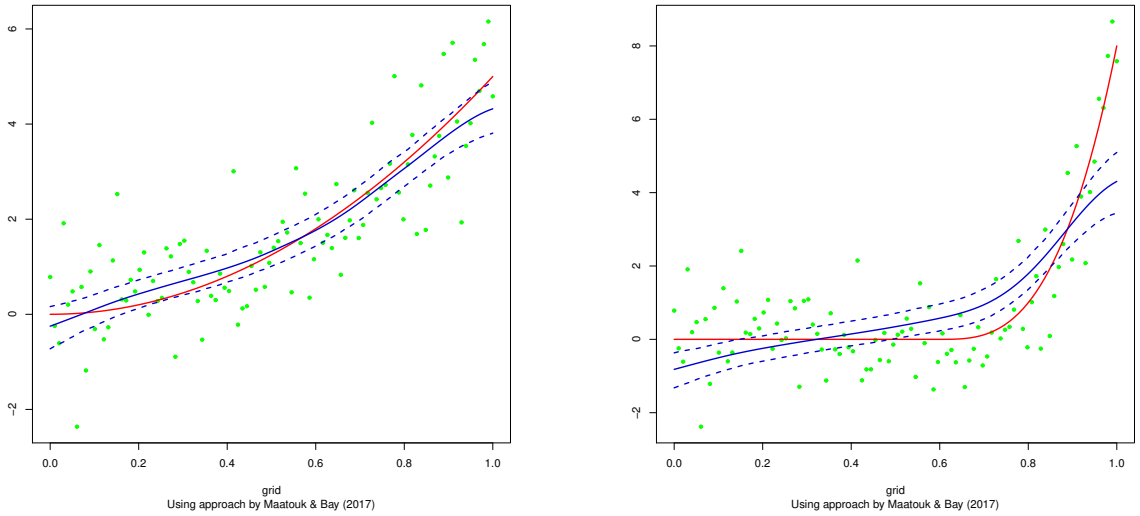

Figure 1: *Monotone function estimation using the basis of Maatouk and Bay [2017] and a joint truncated normal prior $p(\cdot)$ on the coefficients. Red solid curve corresponds to the true function, blue solid curve is the posterior mean, the region within two dotted blue curves represent a pointwise 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.*

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 et al., 2000] such as posterior consistency. However, the finite sample behavior of the

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

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 non-decreasing 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 a clarification. 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 2 shows the estimation of the same function in the right 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 Neelon and Dunson [2004] and Curtis and Ghosh [2011].

Neelon and Dunson [2004] and Curtis and Ghosh [2011] both used point-mass mixture priors as remedy, which is a natural choice under a non-decreasing constraint. However, their introduction becomes somewhat cumbersome under the non-negativity 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 et al., 2010]. The resulting prior shrinks much more aggressively towards the origin, and we empirically illustrate its superior performance over the truncated normal prior. This empirical exercise provides further support to our argument.

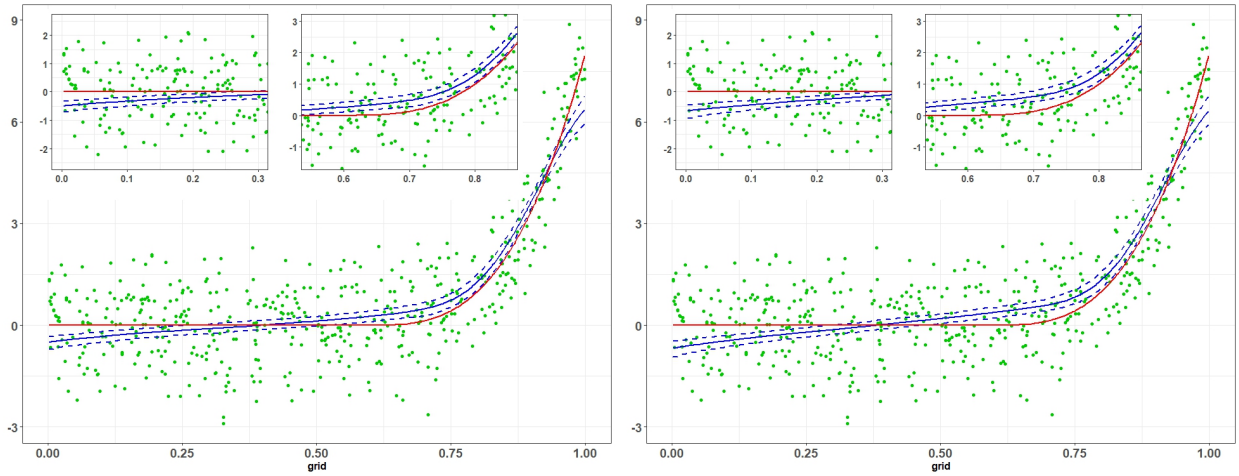

Figure 2: *Monotone function estimation using the basis of Maatouk and Bay [2017] and a joint truncated normal prior on the coefficients. Red solid curve corresponds to the true function, blue solid curve is the posterior mean, the region within two dotted blue curves represent a pointwise 95% credible interval, and the green dots are observed data points corresponding to* $N = 50$ *(left panel) and* $N = 250$ *(right panel).*

2 Mass-shifting phenomenon of truncated normal distributions

2.1 Marginal densities of truncated normal distributions

Our main focus is on 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 notation. 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

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

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, § 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 non-diagonal 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 below, we generalize Cartinhour's result 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*

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

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 non-negative. The map $\theta_1 \mapsto e^{-\theta_1^2/(2\Sigma_{1,1})}$ is decreasing and when $\Sigma_N \in \mathcal{S}_N$, $\theta_1 \mapsto \mathbb{P}(\widetilde{X}_{N-1} \le \Sigma_{N-1,1}\, \theta_1/\Sigma_{1,1})$ is increasing, on $(0, \infty)$. Thus, if $\Sigma_N \in \mathcal{S}_N$, $\widetilde{p}_{1,N}$ is unimodal with a strictly

positive mode.

As another special case, suppose $\Sigma = \Sigma_N(\rho)$ 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 non-negative 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 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. To the best of our knowledge, this has not been observed or quantified in the literature.

We state our results for $\Sigma_N \in \mathcal{B}_{N,K}$, where for $2 \le K \le N-1$, $\mathcal{B}_{N,K}$ denotes the space of K-banded nonnegative correlation matrices,

$$\mathcal{B}_{N,K} = \Big\{\Sigma_N = (\rho_{ij}) \in \mathcal{S}_N \,:\, \rho_{ii} = 1\ \forall\, i,\quad \rho_{ij} = 0\ \forall\ |i-j| \ge K\Big\}. \tag{2.4}$$

While our main theorem below can be proved for other dependence structures, the banded structure naturally arises in statistical applications as discussed in the next section.

Given $\Sigma_N = (\rho_{ij}) \in \mathcal{B}_{N,K}$, define $\rho_{\max} = \max_{i\neq j}\rho_{ij}$ and $\rho_{\min} = \min_{i\neq j, |i-j|\le K}\rho_{ij}$ to be the maximum and minimum correlation values within the band. For $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}, \Sigma_N)$ with $\mathcal{C} = [0,\infty)^N$, let $\alpha_{N,\delta} = \mathbb{P}(\theta_1 \le \delta)$. With these definitions, we are ready to state our main theorem.

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

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

Then, there exists a constant K_0 such that whenever $K \geq K_0$, we have for any $\delta > 0$,

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

where G is a positive rational function of $(\rho_{\min}, \rho_{\max})$, and $C'_{\rho_{\min},\rho_{\max}} > 0$ is a constant free of N.

In particular, if we consider a sequence of K_N-banded correlation 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 $\alpha_{N,\delta}$. To contrast the conclusion of Theorem 2 with two closely related cases, consider first the case when $\theta \sim \mathcal{N}(\mathbf{0}, \Sigma_N)$. 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 $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}, \Sigma_N)$ 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. However, when a combination of dependence and truncation is present, an additional $(\log K)^{1/2} \, K^{-G(\rho_{\min},\rho_{\max})}$ penalty is incurred.

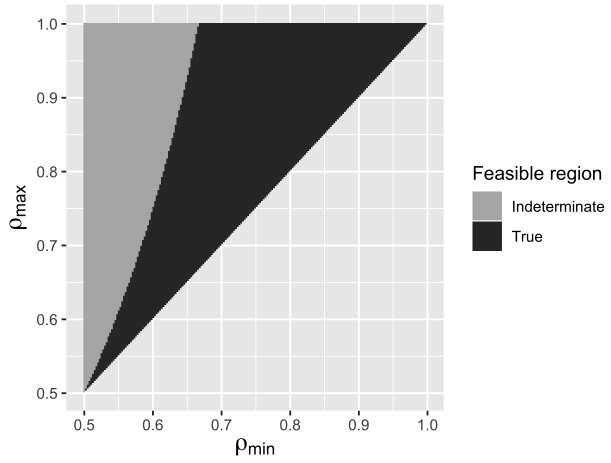

Figure 3: The region shaded in black depicts $\mathcal{Q}$ from the statement of Theorem 2.

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 3. As a special case, if all the non-zero correlations are the same, i.e., $\rho_{\min} = \rho_{\max}$, 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 $\lim_{N\to\infty} \widetilde{p}_{1,N}(0) = 0$. 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.*

We now empirically illustrate the conclusion of the theorem by presenting the univariate marginal density $\widetilde{p}_{1,N}$ for different values of the dimension N and the bandwidth K. The density calculations were performed using the **R** package **tmvtnorm**, which is based on the numerical approximation algorithm proposed in Cartinhour [1990] and subsequent refinements in Genz [1992, 1993], Genz and Bretz [2009]. The left panel of Figure 4 shows that for N fixed at a moderately large value, the

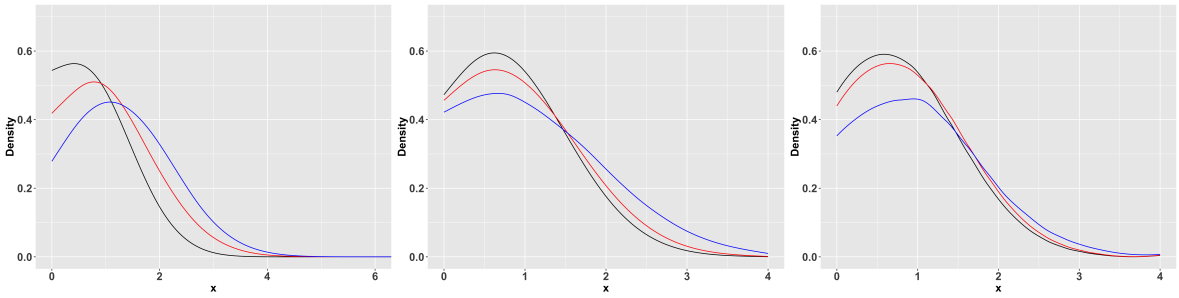

Figure 4: *Left panel shows marginal density functions $\widetilde{p}_{1,N}$ for $K = 2$ (black), $K = 5$ (red) and $K = 20$ (blue) with $N = 100$. Middle panel shows $\widetilde{p}_{1,N}$ for $N = 10$ (black), $N = 50$ (red) and $N = 100$ (blue) with $K = 5$. Right panel shows $\widetilde{p}_{1,N}$ for $(K, N) = (5, 25)$ (black), $(20, 100)$ (red) and $(50, 250)$ (blue).*

probability assigned to a small neighborhood of the origin decreases with increasing K. Also, the mode of the marginal density increasingly shifts away from zero. A similar effect is seen for a fixed K and increasing N in the middle panel and also for an increasing pair (K, N) in the right panel, although the mass-shifting effect is somewhat weakened compared to the left panel. This behavior perfectly aligns with the main message of the theorem that the interplay between the truncation and the dependence brings forth the mass-shifting phenomenon.

The proof of Theorem 2 is non-trivial; we provide the overall chain of arguments in the next subsection, deferring the proof of several auxiliary results to the supplemental document. The marginal probability of $(0, \delta)$ is a ratio of the probabilities of two rectangular regions under a

$\mathcal{N}(\mathbf{0}, \Sigma_N)$ distribution, with the denominator appearing due to the truncation. While there is a rich literature on estimating tail probabilities under correlated multivariate normals using multivariate extensions of the Mill's ratio [Savage, 1962, Ruben, 1964, Sidák, 1968, Steck, 1979, Hashorva and Hüsler, 2003, Lu, 2016], the existing bounds are more suited for numerical evaluation [Cartinhour, 1990, Genz, 1992, Genz and Bretz, 2009] and pose analytic difficulties due to their complicated forms. Moreover, the current bounds lose their accuracy when the region boundary is close to the origin [Gasull and Utzet, 2014], which is precisely our object of interest. Our argument instead relies on novel usage of Gaussian comparison inequalities such as the Slepian's inequality; see Li and Shao [2001], Vershynin [2018a] for book-level treatments. We additionally derive a generalization of Slepian's inequality in Lemma 1, which might be of independent interest. As an important reduction step, we introduce a blocking idea to carefully approximate the banded scale matrix Σ_N by a block tridiagonal matrix to simplify the analysis.

2.3 Proof of Theorem 2

By definition,

$$\alpha_{N,\delta} = \mathbb{P}(\theta_1 \leq \delta) = \frac{\mathbb{P}(0 \leq Z_1 \leq \delta, Z_2 \geq 0, \ldots, Z_N \geq 0)}{\mathbb{P}(Z_1 \geq 0, Z_2 \geq 0, \ldots, Z_N \geq 0)}, \tag{2.5}$$

where $Z \sim \mathcal{N}(\mathbf{0}, \Sigma_N)$. We now proceed to separately bound the numerator and denominator in the above display.

We first consider the denominator in equation (2.5), and use Slepian's lemma to bound it from below. It follows from Slepian's inequality, see comment after Lemma 3 in the supplemental document, that if X, Y are centered d-dimensional Gaussian random variables with $\mathbb{E}(X_i^2) = \mathbb{E}(Y_i^2)$ for all i, and $\mathbb{E}(X_i X_j) \leq \mathbb{E}(Y_i Y_j)$ for all $i \neq j$, then

$$\mathbb{P}(X_1 \geq 0, \ldots, X_d \geq 0) \leq \mathbb{P}(Y_1 \geq 0, \ldots, Y_d \geq 0). \tag{2.6}$$

The Slepian's inequality is a prominent example of a Gaussian comparison inequality originally developed to bound the supremum of Gaussian processes. To apply Slepian's inequality to the present context, we construct another N-dimensional centered Gaussian random vector $S \sim$

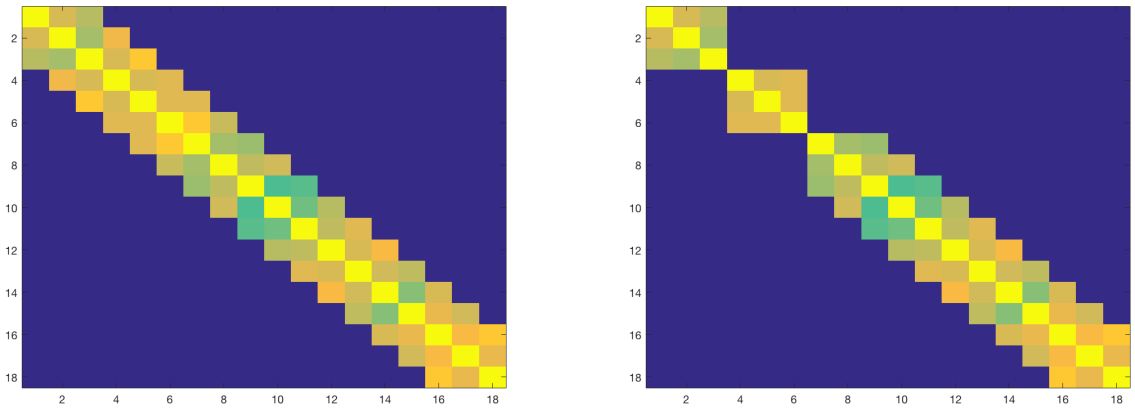

Figure 5: *Left panel: example of* Σ_N *with* $N = 18, K = 3$. *Right panel: the corresponding block approximation* $\widetilde{\Sigma}_N$.

$\mathcal{N}_N(\mathbf{0}, \widetilde{\Sigma}_N)$ such that (i) $S_{[1:K]} \stackrel{d}{=} Z_{[1:K]}$, $S_{[(K+1):2K]} \stackrel{d}{=} Z_{[(K+1):2K]}$ and $S_{[(2K+1):N]} \stackrel{d}{=} Z_{[(2K+1):N]}$, and (ii) the sub-vectors $S_{[1:K]}$, $S_{[(K+1):2K]}$ and $S_{[(2K+1):N]}$ are mutually independent. The correlation matrix $\widetilde{\Sigma}_N$ of S clearly satisfies $(\Sigma_N)_{ij} \geq (\widetilde{\Sigma}_N)_{ij}$ for all $i \neq j$ by construction. Figure 5 pictorially depicts this block approximation in an example with $N = 18$ and $K = 3$. Applying Slepian's inequality, we then have,

$$\begin{aligned}
&\mathbb{P}(Z_1 \geq 0, \ldots, Z_N \geq 0) \geq \mathbb{P}(S_1 \geq 0, \ldots, S_N \geq 0) \\
&\quad = \mathbb{P}(S_{[1:K]} \geq \mathbf{0})\, \mathbb{P}(S_{[(K+1):2K]} \geq \mathbf{0})\, \mathbb{P}(S_{[(2K+1):N]} \geq \mathbf{0}) \\
&\quad = \mathbb{P}(Z_{[1:K]} \geq \mathbf{0})\, \mathbb{P}(Z_{[(K+1):2K]} \geq \mathbf{0})\, \mathbb{P}(Z_{[(2K+1):N]} \geq \mathbf{0}). \qquad (2.7)
\end{aligned}$$

Next, we consider the numerator in equation (2.5). We have,

$$\begin{aligned}
&\mathbb{P}(0 \leq Z_1 \leq \delta, Z_2 \geq 0, \ldots, Z_N \geq 0) \\
&\quad \leq \mathbb{P}\big(0 \leq Z_1 \leq \delta, Z_{[2:K]} \geq \mathbf{0},\, Z_{[(K+1):2K]} \in \mathbb{R}^K,\, Z_{[(2K+1):N]} \geq \mathbf{0}\big) \\
&\quad = \mathbb{P}\big(0 \leq Z_1 \leq \delta, Z_{[2:K]} \geq \mathbf{0}\big)\mathbb{P}\big(Z_{[(2K+1):N]} \geq \mathbf{0}\big). \qquad (2.8)
\end{aligned}$$

The last equality crucially uses $Z_{[1:K]}$ and $Z_{[(2K+1):N]}$ are independent, which is a consequence of Σ_N being K-banded. Taking the ratio of equations (2.7) and (2.8), the term $\mathbb{P}(Z_{[(2K+1):N]} \geq \mathbf{0})$

cancels so that

$$\alpha_{N,\delta} \leq \frac{\mathbb{P}(0 \leq Z_1 \leq \delta, Z_{[2:K]} \geq \mathbf{0})}{\mathbb{P}(Z_{[1:K]} \geq \mathbf{0})\,\mathbb{P}(Z_{[K+1:2K]} \geq \mathbf{0})} = R. \tag{2.9}$$

To bound the terms $\mathbb{P}(Z_{[1:K]} \geq \mathbf{0})$ and $\mathbb{P}(Z_{[K+1:2K]} \geq \mathbf{0})$ in the denominator of R, we resort to another round of Slepian's inequality. Let $Z'' \sim \mathcal{N}(\mathbf{0}, \Sigma_K(\rho_{\min}))$, where recall that $\rho_{\min}$ is the minimum non-zero correlation in Σ_N. Also, recall from equation (2.1) that $\Sigma_K(\rho_{\min})$ denotes the $K \times K$ compound-symmetry correlation matrix with all correlations equal to $\rho_{\min}$. By construction, for any $1 \leq i \neq j \leq K$, $\mathbb{E}(Z_i Z_j), \mathbb{E}(Z_{K+i} Z_{K+j}) \geq \rho_{\min} = \mathbb{E}(Z''_i Z''_j)$. Thus, applying Slepian's inequality as in equation (2.6),

$$\mathbb{P}(Z_{[1:K]} \geq \mathbf{0})\,\mathbb{P}(Z_{[K+1:2K]} \geq \mathbf{0}) \geq \{\mathbb{P}(Z'' \geq \mathbf{0})\}^2.$$

The numerator of equation (2.9) cannot be directly tackled by Slepian's inequality, and we prove the following comparison inequality in the supplemental document.

Lemma 1. *(Generalized Slepian's inequality) Let X, Y be centered d-dimensional Gaussian vectors with $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ for all i and $\mathbb{E}(X_i X_j) \leq \mathbb{E}(Y_i Y_j)$ for all $i \neq j$. Then for any $0 \leq \ell_1 < u_1$ and $u_2, \ldots u_d \in \mathbb{R}$, we have*

$$\mathbb{P}\big(\ell_1 \leq X_1 \leq u_1, X_2 \geq u_2, \ldots, X_d \geq u_d\big) \leq \mathbb{P}\big(\ell_1 \leq Y_1 \leq u_1, Y_2 \geq u_2, \ldots, Y_d \geq u_d\big).$$

Define a random variable $Z' \sim \mathcal{N}(\mathbf{0}, \Sigma_K(\rho_{\max}))$ and use Lemma 1 to conclude that $\mathbb{P}(0 \leq Z_1 \leq \delta, Z_{[2:K]} \geq \mathbf{0}) \leq \mathbb{P}(0 \leq Z'_1 \leq \delta, Z'_{[2:K]} \geq \mathbf{0})$.

Substituting these bounds in equation (2.9), we obtain

$$R \leq R' = \frac{\mathbb{P}(0 \leq Z'_1 \leq \delta, Z'_2 \geq 0, \ldots, Z'_K \geq 0)}{\{\mathbb{P}(Z''_1 \geq 0, \ldots, Z''_K \geq 0)\}^2}. \tag{2.10}$$

The primary reduction achieved by bounding R' by R'' is that we only need to estimate Gaussian probabilities under a compound-symmetry covariance structure. We prove the following inequalities in the supplemental document that provide these estimates.

Lemma 2. *Let $X \sim \mathcal{N}(\mathbf{0}, \Sigma_d(\rho))$ with $\rho \in (0,1)$. Fix $\delta > 0$. Define $\bar{\rho} = (1-\rho)/\rho$. Then,*

$$\begin{aligned}\mathbb{P}(0 \le X_1 < \delta, X_2 \ge 0, \ldots, X_d \ge 0) \\ \le \delta \left\{2(1-\alpha)\bar{\rho}\log(d-1)\right\}^{-1/2} (d-1)^{-(1-\alpha)/\rho} + \exp(-d^\alpha),\end{aligned}$$

for any $\alpha \in (0,1)$. Also,

$$\mathbb{P}(X_1 \ge 0, \ldots, X_d \ge 0) \ge \frac{(2\,\bar{\rho}\log d)^{1/2}}{2\,\bar{\rho}\log d + 1}\, d^{-\bar{\rho}}.$$

A key aspect of the compound-symmetry structure that we exploit is for $X \sim \mathcal{N}(\mathbf{0}, \Sigma_d(\rho))$ with $\rho \in (0,1)$, we can represent $X_i \stackrel{d}{=} \rho^{1/2}\, w + (1-\rho)^{1/2}\, W_i$, where w, W_i's are independent $\mathcal{N}(0,1)$ variables.

Using Lemma 2, we can bound, for any $\alpha \in (0,1)$,

$$\begin{aligned}R' &\le \frac{\delta\left\{2\,\bar{\rho}_{\max}\,(1-\alpha)\log(K-1)\right\}^{-1/2}(K-1)^{-(1-\alpha)/\rho_{\max}} + \exp\{-(K-1)^\alpha\}}{2\,\bar{\rho}_{\min}\log K(2\,\bar{\rho}_{\min}\log K+1)^{-2}K^{-2\,\bar{\rho}_{\min}}} \\ &\le 2^{(1-\alpha)/\rho_{\max}}\,\delta\,\frac{(2\,\bar{\rho}_{\min}\log K+1)^2}{2\,\bar{\rho}_{\min}\,\log K\{2\,\bar{\rho}_{\max}\,(1-\alpha)\log(K-1)\}^{1/2}}\,K^{-\{(1-\alpha)/\rho_{\max}-2\,\bar{\rho}_{\min}\}} \\ &\quad + \exp\{-(K-1)^\alpha\}K^{2\,\bar{\rho}_{\min}}(2\,\bar{\rho}_{\min}\log K+1)^2/(2\,\bar{\rho}_{\min}\log K) \\ &\le C\,\delta\,(\log K)^{1/2}\,K^{-\{(1-\alpha)/\rho_{\max}-2\,\bar{\rho}_{\min}\}} + 4\,\bar{\rho}_{\min}\exp\{-(K-1)^\alpha\}K^{2\,\bar{\rho}_{\min}}\log K, \qquad (2.11)\end{aligned}$$

with $C = 5\bar{\rho}_{\min}/\{(1-\alpha)\bar{\rho}_{\max}\}^{1/2}$. Since $(\rho_{\min}, \rho_{\max}) \in \mathcal{Q}$, we have $\rho_{\min}/\{2(1-\rho_{\min})\} \ge \rho_{\max}$, or equivalently, $2\bar{\rho}_{\min} < 1/\rho_{\max}$. Thus, we can always find $\alpha > 0$ such that $(1-\alpha)/\rho_{\max} - 2\,\bar{\rho}_{\min} > 0$. Fix such an α, and substitute in equation (2.11). The proof is now completed by choosing K_0 large enough so that for any $K > K_0$, the second term in the last line of (2.11) is smaller than the first; this is possible since the second term decreases exponentially while the first does so polynomially in K.

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 usual Gaussian regression

model,

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

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 $\mathcal{C} = \{\theta \in \mathbb{R}^N : A\theta \geq b\}$. For example, when $\mathcal{C}_f$ corresponds to monotone increasing functions, the set $\mathcal{C}$ is of the form $\{\theta_1 \leq \theta_2 \ldots \leq \theta_N\}$ 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

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

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

The truncated normal prior $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}, \Omega_N)$ is conjugate, with the posterior $\theta \mid Y \sim \mathcal{N}_{\mathcal{C}}(\mu_N, \Sigma_N)$, with

$$\mu_N = \Sigma_N \Phi^{\mathrm{T}} Y, \quad \Sigma_N = (\Omega_N^{-1} + \Phi^{\mathrm{T}}\Phi)^{-1}.$$

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 $\mathcal{N}(\mathbf{0}, \Sigma_N)$ 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 Figures 1 and 2 to the mass-shifting phenomenon characterized in the previous section. Since the posterior $\theta \mid Y \sim \mathcal{N}_{\mathcal{C}}(\mu_N, \Sigma_N)$, a draw from the posterior can be represented as

$$\theta = \mu_N + \theta_c, \quad \theta_c \sim \mathcal{N}_{\mathcal{C}}(0, \Sigma_N).$$

Consider an extreme scenario when the true function is entirely flat. In this case, the optimal parameter value $\theta_0 = \mathbf{0}_N$ and under mild assumptions, μ_N is concentrated near the origin with high probability under the true data distribution; see §E of the supplementary material. 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 μ_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 6 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.

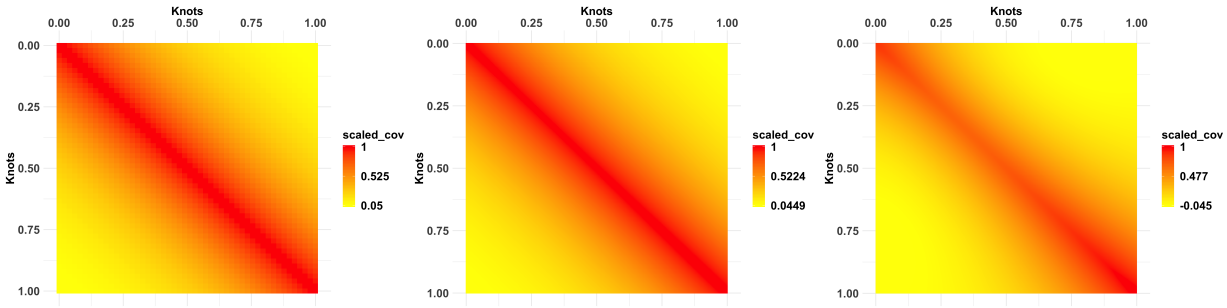

Figure 6: Scaled posterior scale matrix $\widetilde{\Sigma}_N$ of dimension $N = 50$ (left), $N = 250$ (middle) and $N = 500$ (right).

We make this empirical observation concrete below. We first state the assumptions on the basis matrix Φ and prior covariance matrix Ω_N that allows 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*

$2 \leq q \leq N$; *also there exists constants* $0 < C_1 < C_2 < \infty$ *such that*

$$C_1 (n/N) I_N \leq \Phi^{\mathrm{T}} \Phi \leq C_2 (n/N) I_N.$$

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 et al. [2016].

Regarding the prior covariance matrix, we first define a uniform class of symmetric positive definite well-conditioned matrices [Bickel et al., 2008b] as

$$\mathcal{M}(\lambda_0, \alpha, k) = \Big\{ \Omega_N \ : \max_j \sum_i \{|\Sigma_{ij}| \ : \ |i-j| > k\} \leq C\, k^{-\alpha} \text{ for all } k > 0, \\ \text{and } 0 < \lambda_0 \leq \lambda_{\min}(\Omega_N) \leq \lambda_{\max}(\Omega_N) \leq 1/\lambda_0 \Big\}. \tag{3.3}$$

Assumption 2. *We assume the prior covariance matrix* $\Omega_N \in \mathcal{M}(\lambda_0, \alpha, k)$ *defined in* (3.3).

Assumption 2 ensures the covariance matrix is "approximately bandable", which is common in covariance matrix estimation with thresholding techniques [Bickel et al., 2008a,b].

Given above Assumptions, we are now ready to give the approximation result of posterior scale matrix Ω^{-1} to a banded symmetric positive definite matrix.

Proposition 3. *For the posterior scale matrix* $\Sigma_N = (\Omega_N^{-1} + \Phi^{\mathrm{T}}\Phi)^{-1}$ *with* Φ *satisfying Assumption 1 and* Ω_N *satisfying Assumption 2, for sufficiently small* $0 < \epsilon < 1/\lambda_0$ *there exists* $r \gtrsim \log(1/\epsilon)$, *and for sufficiently large* n_0 *we can always find a* $\max(n_0^2 r, n_0 q)$*-banded, symmetric and positive definite matrix* Σ'_N *such that*

$$\|\Sigma_N - \Sigma'_N\| \precsim \delta_{\epsilon,\kappa}, \tag{3.4}$$

where $\delta_{\epsilon,\kappa} = (\epsilon + \kappa^{n_0+1}) \max\{(N/n), (N/n)^2\}$ *and* $0 < \kappa < 1$ *is a fixed constant.*

Proposition 3 states under mild conditions we can always construct a banded positive definite matrix that approximates Σ_N in operator norm. Applying the result in Theorem 2 to a truncated normal distribution with the banded approximation of the posterior scale matrix, the marginal density would present a mass-shifting behavior. If we control the band width K such that the

approximation is close enough to the posterior scale matrix, the marginal posterior distribution would be expected to behave similarly and shift its probability mass away from the origin and the probability mass over the "corner region" will decrease to zero as the dimension N goes to infinity. This helps explain the bias occurred in the posterior mean over the flat area shown in the Figure 1.

4 A de-biasing remedy based on a shrinkage prior

As concluded in previous sections, we view the mass-shifting behavior of the posterior marginals causing the bias in posterior estimation for flat functions. In this section, we will provide empirical evidences that a simple modification to the truncated normal prior can alleviate the issues related to such mass-shifting phenomenon. Among remedies proposed in the literature, Curtis and Ghosh [2011] proposed independent shrinkage priors on the parameter vector $\{u_k\}$ given by a mixture of a point-mass at zero and a univariate normal distribution truncated to the positive real line as an alternative to the truncated normal prior,

$$u_k \sim (1-\pi)\delta_0 + \pi \mathcal{N}_+(\mu, \sigma^2).$$

Similar mixture priors were also previously used by Neelon and Dunson [2004] and Dunson [2005]. The mass at zero allows positive prior probability to functions having exactly flat regions. Although possible in principle, introduction of such 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 towards the origin while the λ_js provide coefficient-specific deviation. We consider default [Carvalho et al., 2010] half-Cauchy priors $\mathcal{C}_+(0,1)$ on τ and the λ_js independently. The $\mathcal{C}_+(0,1)$ distribution has a density proportional to $(1+t^2)^{-1}\mathbb{1}_{(0,\infty)}(t)$.

We continue to use a dependent truncated normal prior $\theta \sim \mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Sigma_N)$ 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 et al., 2010] widely used in the high-dimensional regression context. Figure 9 in §D.1 of the supplementary material shows prior draws for the first and third components of both θ and ξ, based on which the marginal distribution of the ξ_js is clearly seen to place more mass near the origin while retaining heavy tails.

We provide an illustration of 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

$$Y = \xi_0 \mathbf{1}_n + \tau\, \Psi \Lambda \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \mathrm{I}_n). \tag{4.2}$$

Here, Ψ is an $n \times N$ basis matrix with i^{th} 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 parametrized by $\xi_0 \in \mathbb{R}$, $\theta = (\theta_1, \ldots, \theta_N)^{\mathrm{T}} \in \mathcal{C}$, $\lambda = (\lambda_1, \ldots, \lambda_N)^{\mathrm{T}} \in \mathcal{C}$, $\sigma \in \mathbb{R}^+$ and $\tau \in \mathbb{R}^+$. We place a flat prior $\pi(\xi_0) \propto 1$ on ξ_0. We place a truncated normal prior $\mathcal{N}_{\mathcal{C}}(\mathbf{0}_N, \Sigma_N)$ on θ independently of ξ_0, τ and λ with

$$\Sigma_N = (\Sigma_{jj'}), \quad \Sigma_{jj'} = k(u_j - u_{j'}), \quad u_j = j/(N-1), \quad (j = 0, 1, \ldots, N-1)$$

and $k(\cdot)$ is the stationary Matérn kernel with smoothness parameter $\nu > 0$ and length-scale parameter $\ell > 0$. To complete the prior specification, we place improper prior $\pi(\sigma^2) \propto 1/\sigma^2$ on σ^2 and compactly supported priors $\nu \sim \mathcal{U}(0.5, 1)$ and $\ell \sim \mathcal{U}(0.1, 1)$ on ν and ℓ. We develop a data-augmentation Gibbs sampler which combined with the embedding technique of Ray et al. [2019] results in an efficient MCMC algorithm to sample from the joint posterior of $(\xi_0, \theta, \lambda, \sigma^2, \tau^2, \nu, \ell)$; the details are deferred to §D.2 of the supplementary material.

We conduct a small-scale simulation study to illustrate the efficacy of the proposed shrinkage procedure. We consider model (3.1) with true $\sigma = 0.5$ and two different choices of the true f,

namely,

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

for $x \in [0,1]$. The function f_1, which is non-decreasing and flat between 0 and 0.6, was used as the motivating example in the introduction. The function f_2 is also approximately flat between 0.7 and 1, although it is not strictly non-decreasing in this region, which allows us to evaluate the performance under slight model misspecification.

To showcase the improvement due to the shrinkage, we consider a cascading sequence of priors beginning with only a truncated normal prior and gradually adding more structure to eventually arrive at the proposed shrinkage prior. Specifically, the variants considered are

No shrinkage and fixed hyperparameters: Here, we set $\Lambda = \mathrm{I}_N$ and $\tau = 1$ in (4.2), and also fix ν and ℓ, so that we have a truncated normal prior on the coefficients. This was implemented as part of the motivating examples in the introduction. We fix $\nu = 0.75$ and ℓ so that the correlation $k(1)$ between the maximum separated points in the covariate domain equals 0.05.

No shrinkage with hyperparameter updates: The only difference from the previous case is that ν and ℓ are both assigned priors described previously and updated within the MCMC algorithm.

Global shrinkage: We continue with $\Lambda = \mathrm{I}_N$ and place a half-Cauchy prior on the global shrinkage parameter τ. The hyperparameters ν and ℓ are updated.

Global-local shrinkage: This is the proposed procedure where the λ_js are also assigned half-Cauchy priors and the hyperparameters are updated.

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 along with pointwise 95% credible intervals in Figures 7 and 8 respectively, and also report the mean squared prediction error (MSPE) at the bottom of the sub-plots. As expected, only using the truncated normal prior leads to a large bias in the flat region. Adding some global structure to the truncated normal prior, for instance, updating the GP hyperparamaters and adding a global shrinkage term improves estimation around the flat region, which however still lacks the flexibility to transition from the flat region to the strictly increasing region. The global-local shrinkage performs the best, both visually and also in terms of MSPE.

Additionally, the shrinkage procedure performed at least as good as bsar, a very recent state-of-the-art method, 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 10 in §D.3 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.

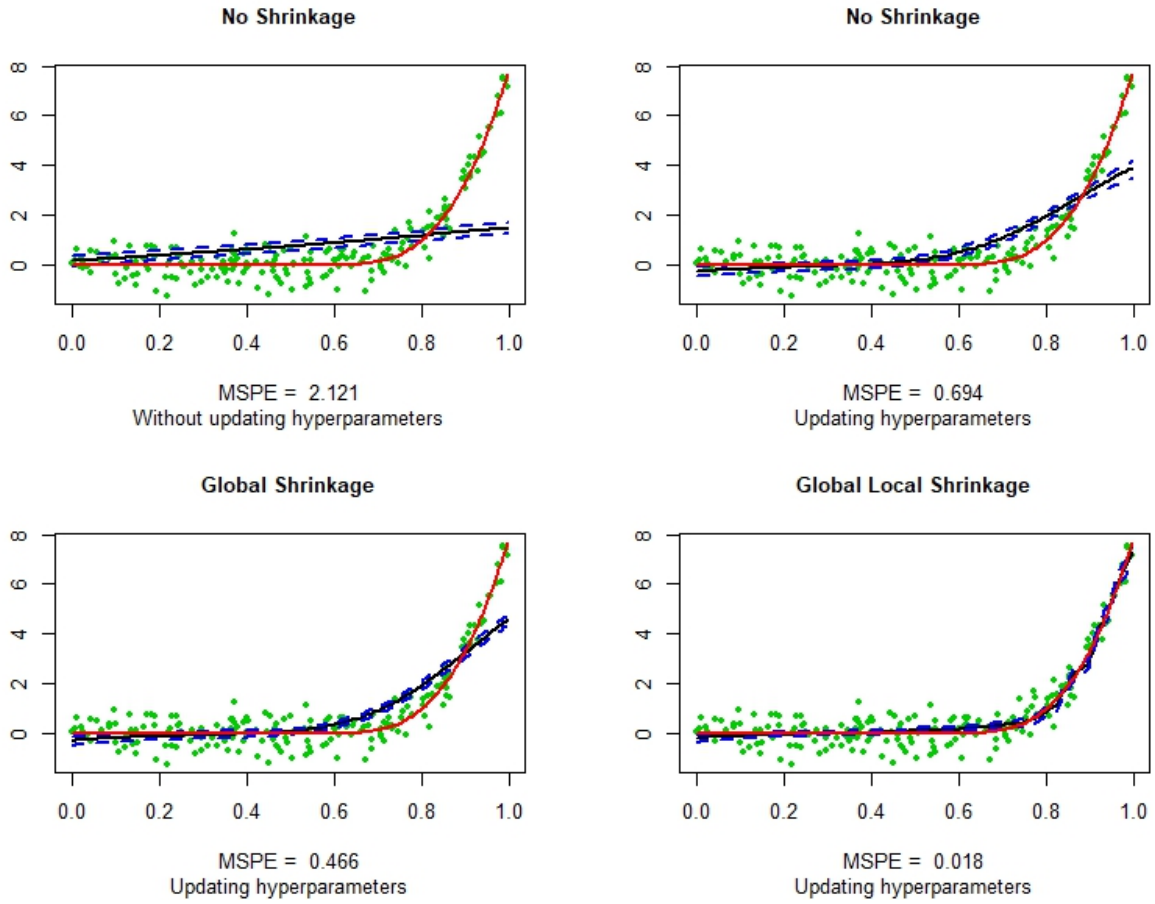

Figure 7: *Out-of-sample prediction accuracy for f_1 using the four variants. Red solid curve corresponds to the true function, black solid curve is the mean prediction, the region within two dotted blue curves represent 95% pointwise prediction Interval and the green dots are* 200 *test data points. MSPE values corresponding to each of the method are also shown in the plots.*

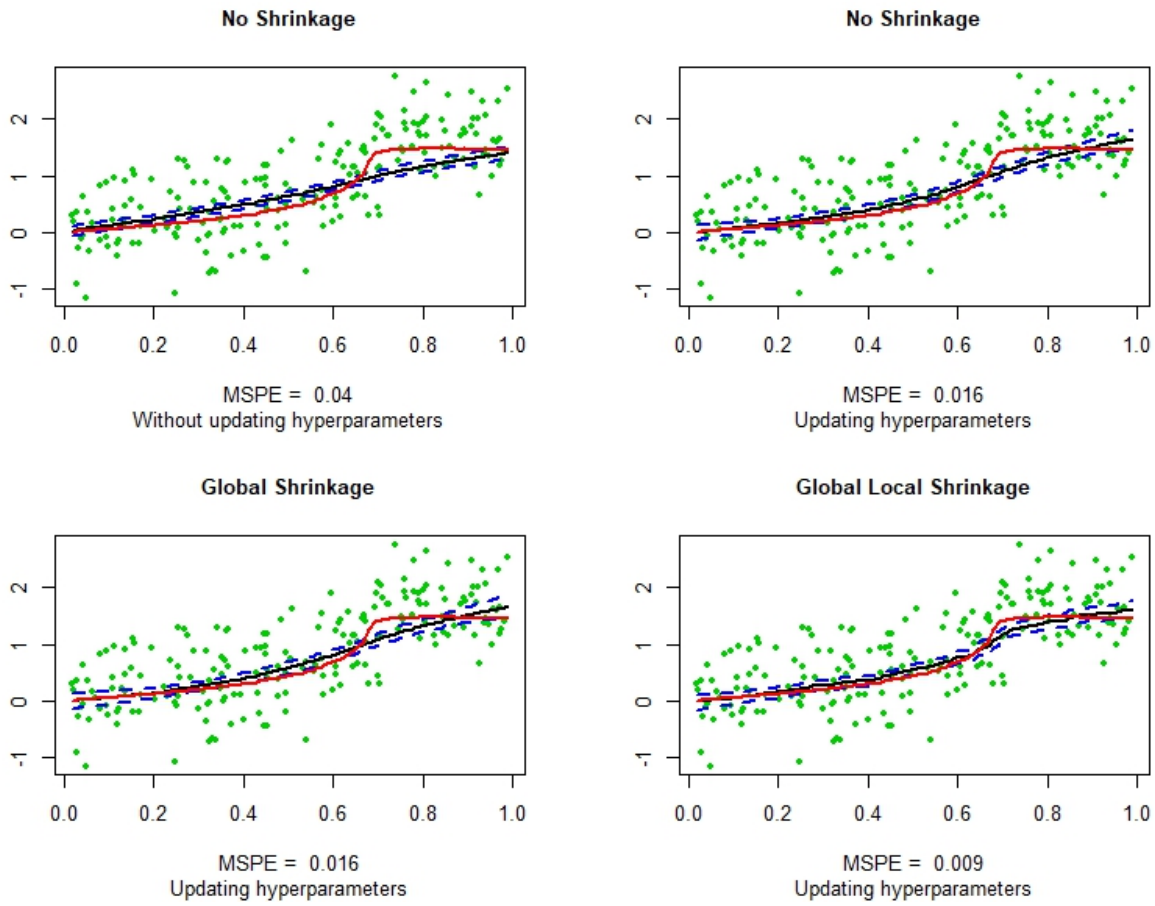

Figure 8: *Same as Figure 7, now for the function* f_2.

5 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 towards 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 et al., 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] 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) [Chipman et al., 2010] method.

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 covariance 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 paper, 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,

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

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 $j = 0, 1, \ldots, N-1$.

Maatouk and Bay [2017] showed that under the representation (M) in (A.1), f is monotone non-decreasing *if and only if* $\theta_i \geq 0$ for all $i = 1, \ldots, N$. Similarly, under (C), f is convex non-decreasing *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 non-negative 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 Material

In this supplementary document, we first collect all remaining technical proofs in the first two sections. §D provides additional details on prior illustration, posterior computation, and posterior performance. In §E we formulate the concentration property of the posterior center μ_N. Several auxiliary results used in the proofs are listed in §F.

B Proofs of auxiliary results in the proof of Theorem 2

In this section, we provide proofs of Lemma 1 and Lemma 2 that were used to prove Theorem 2 in the main manuscript. For any N-dimensional vector $a = [a_1, \ldots, a_d]^{\mathrm{T}}$ we denote its sub-vector $a_{[i_1 : i_2]} = [a_{i_1}, \ldots, a_{i_2}]^{\mathrm{T}}$ for any $1 \leq i_1 < i_2 \leq d$. For two vectors a and b of the same length, let $a \geq b$ ($a \leq b$) denote the event $a_i \geq b_i$ ($a_i \leq b_i$) for all i. For two random variables X and Y, We write $X \overset{\mathrm{d}}{=} Y$ if X and Y are identical in distribution.

B.1 Proof of Lemma 1

For random vectors $X \sim \mathcal{N}(\mathbf{0}, \Sigma_X)$ and $Y \sim \mathcal{N}(\mathbf{0}, \Sigma_Y)$, to show $\mathbb{P}(\ell_1 \leq X_1 \leq u_1,\, X_2 \geq u_2, \ldots, X_d \geq u_d) \leq \mathbb{P}(\ell_1 \leq Y_1 \leq u_1,\, Y_2 \geq u_2, \ldots, Y_d \geq u_d)$, it suffices to show

$$\begin{aligned} &\mathbb{P}(Y_1 \geq u_1, Y_2 \geq u_2, \ldots, Y_d \geq u_d) - \mathbb{P}(X_1 \geq u_1, X_2 \geq u_2, \ldots, X_d \geq u_d) \\ &\leq \mathbb{P}(Y_1 \geq \ell_1, Y_2 \geq u_2, \ldots, Y_d \geq u_d) - \mathbb{P}(X_1 \geq \ell_1, X_2 \geq u_2, \ldots, X_d \geq u_d). \end{aligned} \tag{B.1}$$

We define d-dimensional indicator functions $G(x) = \mathbb{1}_{[u_1,\infty)}(x_1)\prod_{j=2}^d \mathbb{1}_{(u_j,\infty)}(x_j)$ and $F(x) = \mathbb{1}_{[\ell_1,\infty)}(x_1)\prod_{j=2}^d \mathbb{1}_{(u_j,\infty)}(x_j)$, then it is equivalent to show

$$\mathbb{E}\{G(Y)\} - \mathbb{E}\{G(X)\} \leq \mathbb{E}\{F(Y)\} - \mathbb{E}\{F(X)\}. \tag{B.2}$$

We now construct non-decreasing approximating functions of G, F with continuous second order derivatives respectively. Let $\nu \in C^2(\mathbb{R})$ be a non-decreasing twice differentiable function with $\nu(t) = 0$ for $t \leq 0$, $\nu(t) \in [0,1]$ for $t \in [0,1]$, and $\nu(t) = 1$ for $t \geq 1$. Also, choose ν so that $\|\nu'\|_\infty < C$ for some universal constant $C > 0$. For $\eta > 0$, we define $m_\eta(x) = \nu(\eta x)$. It is clear that $m_\eta(x)$ approximates $\mathbb{1}_{[0,\infty)}(x)$ for large η. In fact, for any $x \neq 0$, $\lim_{\eta\to\infty} m_\eta(x) = \mathbb{1}_{[0,\infty)}(x)$.

Given the above, let $g_j^\eta(x_j) = \nu\{\eta(x_j - u_j)\}$ for $j = 1,\ldots,d$, and $f_1^\eta = \nu\{\eta(x - \ell_1)\}$, $f_j^\eta = \nu\{\eta(x_j - u_j)\}$ for $j = 2,\ldots,d$. Define

$$g^\eta(x) = \Pi_{j=1}^d g_j^\eta(x_j) \quad \text{and} \quad f^\eta(x) = \Pi_{j=1}^d f_j^\eta(x_j).$$

It then follows that g^η and f^η provide increasingly better approximations of G and F as $\eta \to \infty$. It thus suffices to show

$$\mathbb{E}\{g^\eta(Y)\} - \mathbb{E}\{g^\eta(X)\} \leq \mathbb{E}\{f^\eta(Y)\} - \mathbb{E}\{f^\eta(X)\}, \tag{B.3}$$

for sufficiently large $\eta > 0$ to be chosen later. We henceforth drop the superscript η from g and f for notation brevity.

We proceed to utilize an interpolation technique commonly used to prove comparison inequalities (see Chapter 7 of Vershynin [2018b]). We construct a sequence of interpolating random variables based on the independent random variables X, Y:

$$S_t = (1-t^2)^{1/2}X + tY, \quad t \in [0,1].$$

Specifically, we have $S_0 = X$, $S_1 = Y$, and for any $t \in [0,1]$, $S_t \sim \mathcal{N}(\mathbf{0}, \widetilde{\Sigma}_t)$ where $\widetilde{\Sigma}_t = (1-t^2)\Sigma_X +$

$t^2\Sigma_Y$. For any twice differentiable function h, we have the following identity

$$\mathbb{E}\{h(Y)\} - \mathbb{E}\{h(X)\} = \int_0^1 \frac{d}{dt}\mathbb{E}\{h(S_t)\}\, dt. \tag{B.4}$$

Applying a multivariate version of Stein's lemma (Lemma 7.2.7 in Vershynin [2018b]) to the integrand in (B.4), one obtains

$$\frac{d}{dt}\mathbb{E}\{h(S_t)\} = t\sum_{i,j=1}^{d}\mathbb{E}\bigg[\{\mathbb{E}(Y_iY_j) - \mathbb{E}(X_iX_j)\}\frac{\partial^2 h}{\partial x_i \partial x_j}(S_t)\bigg]. \tag{B.5}$$

To show (B.3), we define the difference $\Delta = [\mathbb{E}\{f(Y)\} - \mathbb{E}\{f(X)\}] - [\mathbb{E}\{g(Y)\} - \mathbb{E}\{g(X)\}]$. We further decompose Δ as

$$\begin{aligned}
\Delta &= [\mathbb{E}\{f(Y)\} - \mathbb{E}\{f(X)\}] - [\mathbb{E}\{g(Y)\} - \mathbb{E}\{g(X)\}] \\
&= \int_0^1 dt \left\{\frac{d}{dt}\mathbb{E}\{f(S_t)\} - \frac{d}{dt}\mathbb{E}\{g(S_t)\}\right\} \\
&= \int_0^1 dt \left\{t\sum_{i,j=1}^{d}\mathbb{E}\bigg[\{\mathbb{E}(Y_iY_j) - \mathbb{E}(X_iX_j)\}\bigg(\frac{\partial^2 f}{\partial x_i \partial x_j}(S_t) - \frac{\partial^2 g}{\partial x_i \partial x_j}(S_t)\bigg)\bigg]\right\} \\
&= 2\int_0^1 dt \left\{t\sum_{j=2}^{d}\mathbb{E}\bigg[\{\mathbb{E}(Y_1Y_j) - \mathbb{E}(X_1X_j)\}\bigg(\frac{\partial^2 f}{\partial x_1 \partial x_j}(S_t) - \frac{\partial^2 g}{\partial x_1 \partial x_j}(S_t)\bigg)\bigg]\right\} \\
&\quad + \int_0^1 dt \left\{t\sum_{i,j=2}^{d}\mathbb{E}\bigg[\{\mathbb{E}(Y_iY_j) - \mathbb{E}(X_iX_j)\}\bigg(\frac{\partial^2 f}{\partial x_i \partial x_j}(S_t) - \frac{\partial^2 g}{\partial x_i \partial x_j}(S_t)\bigg)\bigg]\right\} \\
&= \Delta_1 + \Delta_2.
\end{aligned}$$

The second equation follows from (B.4) and the third equation follows from (B.5). First we show $\Delta_1 \geq 0$. Since $\mathbb{E}(Y_1Y_j) \geq \mathbb{E}(X_1X_j)$ for all $j > 1$, it suffices to show that for any fixed $t \in [0,1]$ and for any $j = 2, \ldots, d$,

$$D_1 = \mathbb{E}\bigg(\frac{\partial^2 f}{\partial x_1 \partial x_j}(S_t) - \frac{\partial^2 g}{\partial x_1 \partial x_j}(S_t)\bigg) \geq 0.$$

We consider a generic interpolating random variable $S \sim \mathcal{N}(\mathbf{0}, \widetilde{\Sigma})$ by dropping the t-subscript; let

$\phi(s_1, \ldots, s_d)$ denote its probability density function. Then we have

$$\begin{aligned} D_1 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \{f_1'(s_1) f_j'(s_j) - g_1'(s_1) g_j'(s_j)\} \, \Pi_{l \neq 1,j} f_l(s_l) \, \phi(s_1, \ldots, s_d) \, ds_1 \ldots ds_d \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \{f_1'(s_1) - g_1'(s_1)\} \, \phi(s_1, \ldots, s_N) \, ds_1 \right] f_j'(s_j) \, \Pi_{l \neq 1,j} f_l(s_l) \, ds_2 \ldots ds_d. \end{aligned}$$

To guarantee D_1 is non-negative we need the integral over s_1 to be non-negative. Based on the definition of f_1 and g_1, the integral over s_1 can be simplified to

$$\begin{aligned} &\int_{-\infty}^{\infty} \{f_1'(s_1) - g_1'(s_1)\} \, \phi(s_1, \ldots, s_N) \, ds_1 \\ &= \int_{\ell_1}^{\ell_1 + 1/\eta} \left\{ \eta \, \nu'\big(\eta(s_1 - \ell_1)\big) \right\} \phi(s_1, \ldots, s_N) \, ds_1 - \int_{u_1}^{u_1 + 1/\eta} \left\{ \eta \, \nu'\big(\eta(s_1 - u_1)\big) \right\} \phi(s_1, \ldots, s_N) \, ds_1 \\ &= \int_0^{1/\eta} \eta \, \nu'(\eta s_1) \{\phi(s_1 + \ell_1, s_2, \ldots, s_N) - \phi(s_1 + u_1, s_2, \ldots, s_N)\} ds_1. \end{aligned} \tag{B.6}$$

Let us denote the inverse of the covariance matrix $\widetilde{\Sigma}$ as

$$\widetilde{\Sigma}^{-1} = \begin{bmatrix} \widetilde{\Sigma}_{11}^{-1} & \widetilde{\Sigma}_{12}^{-1} \\ \widetilde{\Sigma}_{21}^{-1} & \widetilde{\Sigma}_{22}^{-1} \end{bmatrix},$$

where $\widetilde{\Sigma}_{11}^{-1}$ is a scalar. To check the non-negativity of the last line in (B.6), we now estimate the term

$$\frac{\phi(s_1 + \ell_1, s_2, \ldots, s_d)}{\phi(s_1 + u_1, s_2, \ldots, s_d)} = e^{\{(u_1^2 - \ell_1^2) + 2 s_1 \, (u_1 - \ell_1)\} \, \widetilde{\Sigma}_{11}^{-1}/2 + \, (u_1 - \ell_1) \, \widetilde{\Sigma}_{12}^{-1} \, \tilde{s}_2},$$

where $\tilde{s}_2 = (s_2, \ldots, s_d)^{\mathrm{T}}$. Since $s_j \in [0, 1/\eta]$, we have $s_1 \, (u_1 - \ell_1)\} \, \widetilde{\Sigma}_{11}^{-1} > 0$. We denote $\tilde{\rho} = \max\{\widetilde{\Sigma}_{12}^{-1}\}$ as the largest element of $\widetilde{\Sigma}_{12}^{-1}$. Then, one can choose η large enough such that

$$(u_1 + \ell_1) \, \widetilde{\Sigma}_{11}^{-1} - 2(d-1)\tilde{\rho}/\eta \geq 0,$$

to guarantee $D_1 \geq 0$. For example $\eta = 4(d-1) \, \tilde{\rho} \, \widetilde{\Sigma}_{11}/(u_1 + \ell_1)$ satisfies the above inequality.

Now we show $\Delta_2 \geq 0$. We have $\mathbb{E}(Y_i Y_j) \geq \mathbb{E}(X_i X_j)$ for all $i, j = 2, \ldots, d$. For any $i, j \geq 2$, for

any fixed $t \in [0,1]$, we define

$$D_2 = \mathbb{E}\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(S_t) - \frac{\partial^2 g}{\partial x_i \partial x_j}(S_t) \right) = \mathbb{E}\{(f_1 - g_1) f_i' \, g_j' \, \Pi_{k \neq 1,i,j} f_k\}.$$

Since $f_1 - g_1 \geq 0$, and $f_j' \geq 0$ for all $j > 1$, it follows that $D_2 \geq 0$ and thus $\Delta_2 \geq 0$. Combining with the non-negativity of Δ_1 completes the proof of Lemma B.1.

B.2 Proof of Lemma 2

For $X \sim \mathcal{N}(\mathbf{0}, \Sigma_d(\rho))$ with $\rho \in (0,1)$, we will repeatedly use its equivalent expression

$$X_i = \rho^{1/2} \, w + (1-\rho)^{1/2} \, W_i \quad (i = 1, \ldots, N), \tag{B.7}$$

where w, W_i's are independent standard normal variables.

Proof of the upper bound. We recall $\bar{\rho} = (1-\rho)/\rho$. For any fixed $\delta > 0$ and $\alpha \in (0,1)$, we have

$$\begin{aligned}
&\mathbb{P}(0 \leq X_1 < \delta, X_2 \geq 0, \ldots, X_d \geq 0) \qquad (B.8)\\
&= \mathbb{P}\Big(0 \leq \rho^{1/2} \, w + (1-\rho)^{1/2} \, W_1 \leq \delta, w \geq \bar{\rho}^{1/2} \max_{2 \leq i \leq d} W_i \Big) \\
&= \mathbb{P}\Big(\Big\{ 0 \leq \rho^{1/2} \, w + (1-\rho)^{1/2} \, W_1 \leq \delta, w \geq \bar{\rho}^{1/2} \max_{2 \leq i \leq d} W_i \Big\} \\
&\quad \cup \Big[\max_{i \leq d} W_i \geq \{2(1-\alpha)\log(d-1)\}^{1/2} \Big] \cup \Big[\max_{i \leq d} W_i \leq \{2(1-\alpha)\log(d-1)\}^{1/2} \Big] \Big) \\
&\leq \mathbb{P}\Big[0 \leq \rho^{1/2} \, w + (1-\rho)^{1/2} \, W_1 \leq \delta, w \geq \{2 \, \bar{\rho} \, (1-\alpha)\log(d-1)\}^{1/2} \Big] \\
&\quad + \mathbb{P}\Big[\max_{i \leq d} W_i \leq \{2(1-\alpha)\log(d-1)\}^{1/2} \Big] \\
&= P_1 + P_2.
\end{aligned}$$

First, we estimate P_1 in (B.8). By applying the equivalent expression of X in (B.7), we have

$$
\begin{aligned}
P_1 = \mathbb{P}\Big[W_1 \in \Big\{ -\Big(\frac{\rho}{1-\rho}\Big)^{1/2} w, \delta/(1-\rho)^{1/2} - \Big(\frac{\rho}{1-\rho}\Big)^{1/2} w \Big\} \mid w \geq \{2\,\bar{\rho}\,(1-\alpha)\log(d-1)\}^{1/2}\Big] \\
\mathbb{P}\big[w \geq \{2\,\bar{\rho}\,(1-\alpha)\log(d-1)\}^{1/2}\big] \\
\leq \mathbb{P}\big[W_1 \in \big(-\{2(1-\alpha)\log(d-1)\}^{1/2}, \delta(1-\rho)^{-1/2} - \{2(1-\alpha)\log(d-1)\}^{1/2}\big)\big] \\
\mathbb{P}\big(w \geq \{2\,\bar{\rho}\,(1-\alpha)\log(d-1)\}^{1/2}\big) \\
\leq \delta(2\pi)^{-1/2} \exp\big(-\big[\delta(1-\rho)^{-1/2} - \{2(1-\alpha)\log(d-1)\}^{1/2}\big]^2/2\big) \\
\{2\,\bar{\rho}\,(1-\alpha)\log(d-1)\}^{-1/2} \exp\big(-\{2\,\bar{\rho}\,(1-\alpha)\log(d-1)\}^2/2\big).
\end{aligned}
$$

The last inequality follows from Lemma 5 in Appendix F.

Now we move to estimate the term P_2 in (B.8). We have,

$$
\begin{aligned}
\mathbb{P}\big[\max_{i\leq d} W_i \leq \{2\,(1-\alpha)\log d\}^{1/2}\big] &= \big(1 - \mathbb{P}\big[Z \geq \{2\,(1-\alpha)\,\log d\}^{1/2}\big]\big)^d \\
&\leq \exp\big(- d\,\mathbb{P}\big[Z \geq \{2\,(1-\alpha)\,\log d\}^{1/2}\big]\big) \leq \exp(-d^{\,\alpha}),
\end{aligned}
$$

where $Z \sim \mathcal{N}(0,1)$.

Combining bounds for P_1 and P_2, we obtain

$$
\mathbb{P}(0 \leq X_1 < \delta, X_2 \geq 0, \ldots, X_d \geq 0) \leq \delta\,\{2\bar{\rho}\,(1-\alpha)\log(d-1)\}^{-1/2}\,(d-1)^{-(1-\alpha)/\rho} + \exp(-d^{\,\alpha}),
$$

for any $\alpha \in (0,1)$. We then complete the proof of the upper bound.

Proof of the lower bound. We provide a more general result for the lower bound. We show that for any scalar $a \geq 0$, we have

$$
\mathbb{P}(X \geq a\mathbf{1}_d) \geq \frac{a\rho^{-1/2} + (2\,\bar{\rho}\,\log N)^{1/2}}{\big\{a\rho^{-1/2} + (2\,\bar{\rho}\,\log d)^{1/2}\big\}^2 + 1} \exp\Big[-\frac{1}{2}\Big\{a\rho^{-1/2} + (2\,\bar{\rho}\,\log d)^{1/2}\Big\}^2\Big], \tag{B.9}
$$

where recall that $\mathbf{1}_N$ denotes a N-dimensional vector of ones. By taking $a = 0$ leads to the desired lower bound in Lemma 2.

Now we prove the lower bound in (B.9). First,

$$\begin{aligned}\mathbb{P}(X \geq a\mathbf{1}_d) &= \mathbb{P}\big(\rho^{1/2}\, w + (1-\rho)^{1/2}\, W_i \geq a, \text{ for } i = 1, \ldots, d\big) \\ &= \mathbb{E}\big(\,\mathbb{P}\big[w \geq \rho^{-1/2}\{a - (1-\rho)^{1/2}\, W_i\},\, i = 1, \ldots, d \mid W_1, \ldots, W_d\big]\big) \\ &\overset{\text{(i)}}{=} \mathbb{E}\big(\,\mathbb{P}\big[w \geq \rho^{-1/2}\big\{a + (1-\rho)^{1/2} \max_i W_i\big\} \mid W_1, \ldots, W_N\big]\big) \\ &= \mathbb{E}\bigg\{1 - \Phi\bigg(a\rho^{-1/2} + \bar{\rho}^{1/2} \max_i W_i\bigg)\bigg\},\end{aligned} \tag{B.10}$$

where $W = [W_1, \ldots, W_d]^{\mathrm{T}}$. Here, (i) holds since $-W_i \overset{\mathrm{d}}{=} W_i$ for $i = 1, \ldots, d$ and $\max_{i \leq d}(-W_i) \overset{\mathrm{d}}{=} \max_{i \leq d}(W_i)$.

We now proceed to lower bound the right hand side of the last equation in (B.10). To that end, we define $g(a, b) = 1 - \Phi(a\rho^{-1/2} + \bar{\rho}^{1/2}\, b)$, where $g : \mathbb{R}_+ \times \mathbb{R} \to [0, 1]$. Importantly, g is non-increasing function of a, b for $a, b \in \mathbb{R}$, and g is a convex function of (a, b) for $a, b > 0$. For any fixed $a > 0$, since $g(a, \max_i W_i)$ is non-increasing in $\max_i W_i$, we have $g(a, \max_i W_i) \geq g(a, \max_i |W_i|)$. We then apply Jensen's inequality,

$$\mathbb{E}\Big\{g\Big(a, \max_{1 \leq i \leq d} |W_i|\Big)\Big\} \geq g\Big\{a, \mathbb{E}\Big(\max_{1 \leq i \leq d} |W_i|\Big)\Big\} \geq g\big\{a, (2 \log d)^{1/2}\big\}.$$

The last inequality holds by applying Lemma 4 in Appendix F. To lower bound $g\big\{a, (2 \log d)^{1/2}\big\}$ we apply Lemma 5 in Appendix F. Eventually, we obtain

$$\mathbb{E}\left\{g\Big(a, \max_{1 \leq i \leq d} |W_i|\Big)\right\} \geq \frac{a\rho^{-1/2} + (2\bar{\rho} \log d)^{1/2}}{\big\{a\rho^{-1/2} + (2\bar{\rho} \log d)^{1/2}\big\}^2 + 1} \exp\big[-\big\{a\rho^{-1/2} + (2\bar{\rho} \log d)^{1/2}\big\}^2/2\big].$$

This completes the proof for the lower bound.

C Remaining proofs from the main document

C.1 Proof of the Proposition 1

Now we derive the k-dimensional marginal density function. We denote $\theta^{(k)} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}$ and $\theta^{(N-k)} = (\theta_{k+1}, \ldots, \theta_N)^{\mathrm{T}}$. We partition Σ_N into appropriate blocks as

$$\Sigma_N = \begin{bmatrix} \Sigma_{k,k} & \Sigma_{k,N-k} \\ \Sigma_{N-k,k} & \Sigma_{N-k,N-k} \end{bmatrix}.$$

We also partition its inverse matrix $\widetilde{\Sigma}_N$,

$$\widetilde{\Sigma}_N = \begin{bmatrix} \widetilde{\Sigma}_{k,k} & \widetilde{\Sigma}_{k,N-k} \\ \widetilde{\Sigma}_{k,N-k} & \widetilde{\Sigma}_{N-k,N-k} \end{bmatrix}.$$

Then the k-dimensional marginal $\widetilde{p}_{k,N}(\theta_1, \ldots, \theta_k) =$

$$\begin{aligned}
&\left(\frac{1}{2\pi}\right)^{N/2} \{\det(\Sigma)\}^{-1/2} \int_0^\infty \cdots \int_0^\infty \exp\big\{ -\big(\theta^{(k)^{\mathrm{T}}} \widetilde{\Sigma}_{k,k}\, \theta^{(k)} \\
&\quad - 2\theta^{(k)^{\mathrm{T}}} \widetilde{\Sigma}_{k,N-k}\, \theta^{(N-k)} + \theta^{(N-k)^{\mathrm{T}}} \widetilde{\Sigma}_{N-k,N-k}\, \theta^{(N-k)}\big)/2\big\}\, d\theta^{(N-k)} \\
&= \left(\frac{1}{2\pi}\right)^{k/2} \exp\big\{ -\theta^{(k)^{\mathrm{T}}} \widetilde{\Sigma}_{k,k} \theta^{(k)}/2\big\} \cdot \Pi_{i=1}^k \mathbb{1}_{[0,\infty)}(\theta_i) \left(\frac{1}{2\pi}\right)^{(N-k)/2} \{\det(\widetilde{\Sigma}_{N-k,N-k})\}^{-1/2} \\
&\quad \cdot \int_0^\infty \cdots \int_0^\infty \exp\big\{ -\|\widetilde{\Sigma}_{N-k,N-k}^{\frac{1}{2}} \big(\theta^{(N-k)} - \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}\, \theta^{(k)}\big)\|^2/2\big\} d\theta^{(N-k)} \\
&= \left(\frac{1}{2\pi}\right)^{k/2} \exp\{-\theta^{(k)^{\mathrm{T}}} \widetilde{\Sigma}_{k,k}\, \theta^{(k)}/2\}\, \mathbb{P}(\widetilde{X}_{N-k} \le \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}\, \theta^{(k)}) \cdot \Pi_{i=1}^k \mathbb{1}_{[0,\infty)}(\theta_i).
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\Sigma}_{k,k} &= \Sigma_{k,k}^{-1} + \Sigma_{k,k}^{-1}\, \Sigma_{k,N-k}\, \widetilde{\Sigma}_{N-k,N-k}\, \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}, \\
\widetilde{\Sigma}_{k,N-k} &= \Sigma_{k,k}^{-1}\, \Sigma_{k,N-k}\, \widetilde{\Sigma}_{N-k,N-k}, \\
\widetilde{\Sigma}_{N-k,N-k} &= (\Sigma_{N-k,N-k} - \Sigma_{N-k,k}\, \Sigma_{k,k}^{-1}\, \Sigma_{k,N-k})^{-1},
\end{aligned}$$

and $\widetilde{X}_{N-k} \sim \mathcal{N}_{N-k}(\mathbf{0}_{N-k}, \widetilde{\Sigma}_{N-k,N-k}^{-1})$.

C.2 Proof of Proposition 3

We first introduce some notations that are used in the proof. For a $N \times N$ matrix A, we denote $\lambda_j(A)$ as its jth eigenvalue, and denote $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum of eigenvalues, respectively. For a matrix A, we define its operator norm as $\|A\| = \{\lambda_{\max}(A^{\mathrm{T}}A)\}^{1/2}$. For two quantities a, b, we write $a \asymp b$ when a/b can be bounded from below and above by two finite constants.

We repeatedly apply Newmann series and Lemma 6 in Appendix F to construct the approximation matrix to the posterior scale matrix Σ_N. Under Assumption 2, we have the prior covariance matrix $\Omega_N \in \mathcal{M}(\lambda_0, \alpha, k)$ for some universal constants $\lambda_0, \alpha, k > 0$. Then for any $\epsilon \in (0, \lambda_0/2)$, by choosing $r \geq \log(C/\epsilon)/\alpha$, one can find a r-banded symmetric and positive definite matrix $\Omega_{N,r}$ such that

$$\|\Omega_N - \Omega_{N,r}\| \leq \epsilon. \tag{C.1}$$

Now we let $M = \lambda_{\max}(\Omega_{N,r})$ and $m = \lambda_{\min}(\Omega_{N,r})$. Given (C.1), we have

$$\lambda_0 - \epsilon \leq m \leq M \leq 1/\lambda_0 + \epsilon. \tag{C.2}$$

By choosing $\xi = 2/(M+m)$, simple calculation gives $\|I_N - \xi\,\Omega_{N,r}\| < 1$. We now apply Newmann series to construct a polynomial of $\Omega_{N,r}$ of degree n_1, defined as $\widetilde{\Omega}^{-1} = \xi \sum_{j=0}^{n_1}(I - \xi\,\Omega_{N,r})^j$, for some integer $n_1 > 0$ to be chosen later. Applying Lemma 6 in Appendix F, we have

$$\|\Omega_{N,r}^{-1} - \widetilde{\Omega}^{-1}\| \leq \kappa_0^{n_1+1}/(\lambda_0 - \epsilon), \tag{C.3}$$

where $\kappa_0 = (M-m)/(M+m)$. Applying Lemma 6 we guarantee $\widetilde{\Omega}^{-1}$ is $(n_1\, r)$-banded and positive definite. Combining results in (C.2) and (C.3), we have

$$\lambda_0/(1+\lambda_0\epsilon) - \kappa_0^{n_1+1}/(\lambda_0 - \epsilon) \leq \lambda_{\min}(\widetilde{\Omega}^{-1}) \leq \lambda_{\max}(\widetilde{\Omega}^{-1}) \leq 1/(\lambda_0 - \epsilon) + \kappa_0^{n_1+1}/(\lambda_0 - \epsilon). \tag{C.4}$$

Now we let $\widetilde{\Sigma}^{-1} = \widetilde{\Omega}^{-1} + \Phi^{\mathrm{T}}\Phi$. Under Assumption 1 we have $\widetilde{\Sigma}^{-1}$ is k-banded with $k = \max\{n_1\, r, q\}$.

We then define $\tilde{\lambda}_1 = \lambda_{\max}(\widetilde{\Sigma}^{-1})$ and $\tilde{\lambda}_N = \lambda_{\min}(\widetilde{\Sigma}^{-1})$. Thus, given (C.4), we have

$$C_1\,(n/N) + \lambda_0/(1+\lambda_0\epsilon) - \kappa_0^{n_1+1}/(\lambda_0-\epsilon) \leq \tilde{\lambda}_N \leq \tilde{\lambda}_1 \leq C_2\,(n/N) + 1/(\lambda_0-\epsilon) + \kappa_0^{n_1+1}/(\lambda_0-\epsilon),$$

for constants $0 < C_1 < C_2 < \infty$ in Assumption 2.

We first consider the case where $N/n \to a$ for some constant $a \in (0,1)$, as $n, N \to \infty$. For sufficiently large n, N, we obtain

$$C_1' a + \lambda_0/(1+\lambda_0\epsilon) \leq \tilde{\lambda}_N \leq \tilde{\lambda}_1 \leq C_2'\,a + 1/(\lambda_0-\epsilon), \tag{C.5}$$

for constants C_1', C_2' satisfying $C_1' < C_1$ and $C_2 < C_2'$.

Secondly, we consider the case where $N/n \to 0$ as $n, N \to \infty$. In this case, n/N dominates in the eigenvalues of $\widetilde{\Sigma}^{-1}$. Thus, for sufficiently large n, N, we have

$$C_1\,(n/N) \leq \tilde{\lambda}_N \leq \tilde{\lambda}_1 \leq C_2\,(n/N). \tag{C.6}$$

Now we apply Lemma 6 one more time to construct the approximation matrix to the inverse of $\widetilde{\Sigma}^{-1}$. Again, by taking $\gamma = 2/(\tilde{\lambda}_1 + \tilde{\lambda}_N)$, we have $\|I_N - \gamma\,\widetilde{\Sigma}^{-1}\| < 1$. Now we define $\Sigma' = \gamma \sum_{j=0}^{m_1} (I_N - \gamma\,\widetilde{\Sigma}^{-1})^j$ for some positive integer m_1. Also, it follows

$$\|\widetilde{\Sigma} - \Sigma'\| \leq \tilde{\kappa}^{m_1+1}/\tilde{\lambda}_N, \tag{C.7}$$

where $\tilde{\kappa} = (\tilde{\lambda}_1 - \tilde{\lambda}_N)/(\tilde{\lambda}_1 + \tilde{\lambda}_N)$. By construction Σ' is $(m_1\,k)$-banded.

Now we estimate $\tilde{\kappa}$. For large enough N, n in the first case, we can upper bound

$$\tilde{\kappa} \leq \kappa_1 = \frac{(C_2' - C_1')\,a + 1/(\lambda_0-\epsilon) - \lambda_0/(1+\lambda_0\epsilon)}{(C_2' + C_1')\,a + 1/(\lambda_0-\epsilon) + \lambda_0/(1+\lambda_0\epsilon)}.$$

The inequality holds since the map $x \mapsto (1-x)/(1+x)$ is non-increasing in $x \in (0,1)$. Combing this with the result in (C.5) and taking $x = \tilde{\lambda}_N/\tilde{\lambda}_1$ leads to the expression of κ_1. Based on (C.7), we have $\|\widetilde{\Sigma} - \Sigma'\| \leq \kappa_1^{m_1+1}/\{C_1' a + \lambda_0/(1+\lambda_0\epsilon)\}$. For N, n in the second case, following a similar line of argument, we have $\|\widetilde{\Sigma} - \Sigma'\| \leq \tilde{\kappa}^{m_1+1}\,N/(C_1 n)$ with $\tilde{\kappa} = (C_2 - C_1)/(C_2 + C_1)$.

We recall the posterior scale matrix $\Sigma_N = (\Omega_N^{-1} + \Phi^{\mathrm{T}}\Phi)^{-1}$. Then we have

$$\begin{aligned}\|\Sigma_N - \Sigma'\| &\leq \|\Sigma_N - \widetilde{\Sigma}\| + \|\widetilde{\Sigma} - \Sigma'\| \\ &\leq \|\Sigma_N\|(\|\Omega_N^{-1} - \Omega_{N,r}^{-1}\| + \|\Omega_{N,r}^{-1} - \widetilde{\Omega}^{-1}\|)\|\widetilde{\Sigma}\| + \|\widetilde{\Sigma} - \Sigma'\| \\ &\leq \|\Sigma_N\|\|\widetilde{\Sigma}\|(c_1\,\epsilon + c_2\,\kappa_0^{n_1+1}) + \|\widetilde{\Sigma} - \Sigma'\|\end{aligned}$$

where $c_1 = \|\Omega^{-1}\|\|\Omega_{N,r}^{-1}\|$ and $c_2 = 1/(\lambda_0 - \epsilon)$. The first inequality follows from the triangular inequality and the second inequality follows from the identity $\|A^{-1} - B^{-1}\| = \|A^{-1}\|\|A - B\|\|B^{-1}\|$ for invertible matrices A, B. The last inequality follows from results in (C.1) and (C.3).

For N, n in the first case, $\|\Sigma_N\|$ and $\|\widetilde{\Sigma}\|$ are upper bounded by some constants that are free of n, N given (C.5). Then we obtain

$$\|\Sigma_N - \Sigma'\| \leq C'(\epsilon + \kappa_0^{n_1+1} + \kappa_1^{m_1+1}),$$

where $C' = \max\{c_1, c_2, C_1'\,a + \lambda_0/(1 + \lambda_0\epsilon)\}/\{C_1'\,a + \lambda_0/(1 + \lambda_0\epsilon)\}^2$.

For N, n in the second case, for sufficiently large N, n we have $\|\Sigma_N\| \asymp (N/n)$ given (C.6). Then we have

$$\|\Sigma_N - \Sigma'\| \leq C''\,\{(N/n)^2(\epsilon + \kappa_0^{n_1+1}) + (N/n)\tilde{\kappa}^{m_1+1}\},$$

where $C'' = C_1^{-2}\max\{c_1, c_2, C_1\}$. Letting $\kappa = \max\{\kappa_0, \kappa_1, \tilde{\kappa}\}$, $n_0 = \min\{n_1, m_1\}$, and $\delta_{\epsilon,\kappa} = (\epsilon + \kappa^{n_0+1})\max\{(N/n), (N/n)^2\}$ yields the result in Proposition 3.

D Additional details on the numerical studies

D.1 Prior draws

We consider equation (4.1) and the prior specified in section 4. Prior samples on both θ and ξ of dimension $N = 100$ were drawn. Figure 9 shows prior draws for the first and third components of both θ and ξ.

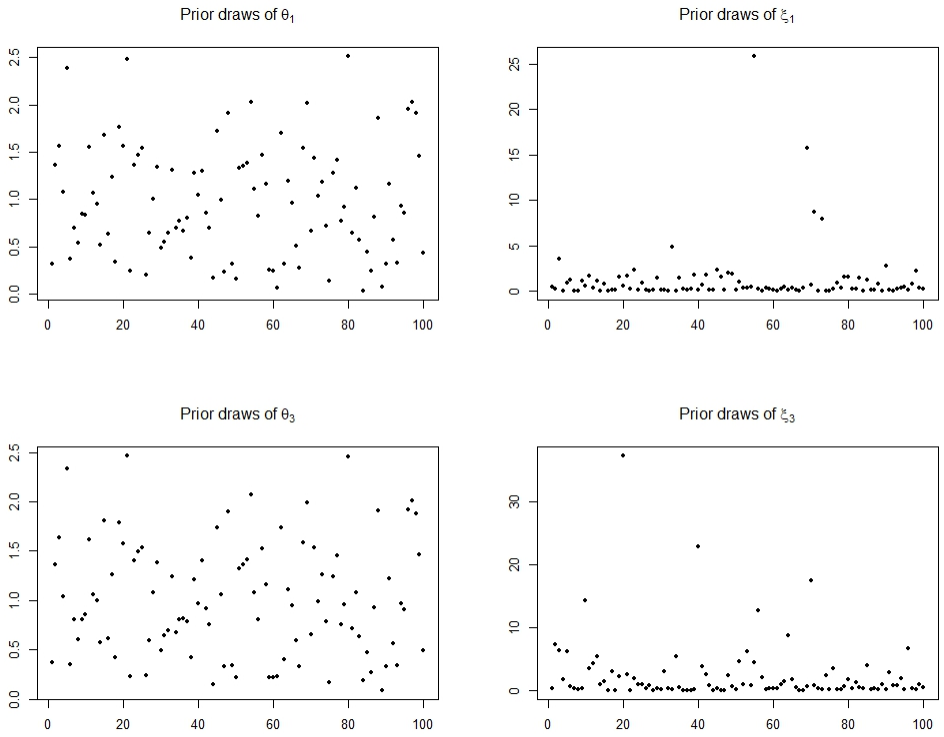

Figure 9: *Showing prior draws from distribution of θ (left panel) and ξ (right panel). Top and bottom panels correspond to first and third components respectively, for both θ and ξ.*

D.2 Posterior Computations

We now consider model (4.2) and the prior specified in section 4. Then the full conditional distribution of θ

$$\pi(\theta \mid Y, \xi_0, \lambda, \tau, \sigma) \propto \exp\left\{ -\frac{1}{2\sigma^2}\|\widetilde{Y} - \Psi\Lambda\theta\|^2 \right\} \exp\left\{ -\frac{1}{2\tau^2}\theta^{\mathrm{T}} K^{-1}\theta \right\} \mathbb{1}_{\mathcal{C}_\theta}(\theta)$$

can be approximated by

$$\begin{aligned}\pi(\theta \mid Y, \xi_0, \lambda, \tau, \sigma) &\propto \exp\left\{ -\frac{1}{2\sigma^2}\|\widetilde{Y} - \Psi_\lambda\,\theta\|^2 \right\} \exp\left\{ -\frac{1}{2\tau^2}\theta^{\mathrm{T}} K^{-1}\theta \right\} \left\{ \prod_{j=1}^{N+1} \frac{e^{\eta\,\theta_j}}{1+e^{\eta\,\theta_j}} \right\} \\ &= \left[\exp\left\{ -\frac{1}{2\sigma^2}\|\widetilde{Y} - \Psi_\lambda\,\theta\|^2 \right\} \left\{ \prod_{j=1}^{N+1} \frac{e^{\eta\,\theta_j}}{1+e^{\eta\,\theta_j}} \right\} \right] \exp\left\{ -\frac{1}{2\tau^2}\theta^{\mathrm{T}} K^{-1}\theta \right\}\end{aligned}$$

where η is a large valued constant, $\widetilde{Y} = Y - \xi_0 1_n$ and $\Psi_\lambda = \Psi\Lambda$. The above is same as equation (5) of Ray et al. [2019] and thus falls under the framework of their sampling scheme. For more details on the sampling scheme and the approximation, one can refer to Ray et al. [2019].

Note that $\lambda_j \sim \mathcal{C}_+(0,1),\ j = 1, \ldots, N$, can be equivalently given by $\lambda_j \mid w_j \sim \mathcal{N}(0, w_j^{-1})\mathbb{1}(\lambda_j > 0)\,,\ w_j \sim \mathcal{G}(0.5, 0.5)\,,\ j = 1, \ldots, N$. Thus the full conditional distribution of λ can be approximated by:

$$\pi(\lambda \mid Y, \xi_0, \theta, w, \tau, \sigma) \propto \left[\exp\left\{ -\frac{1}{2\sigma^2}\|\widetilde{Y} - \Psi_\theta\,\lambda\|^2 \right\} \left\{ \prod_{j=1}^{N+1} \frac{e^{\zeta\,\lambda_j}}{1+e^{\zeta\,\lambda_j}} \right\} \right] \exp\left\{ -\frac{1}{2}\lambda^{\mathrm{T}} W \lambda \right\}$$

where ζ plays the same role as η, $w = (w_1, \ldots, w_N)^{\mathrm{T}}$, $W = \mathrm{diag}(w_1, \ldots, w_N)$, $\Psi_\theta = \Psi\Theta$ and $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_N)$. Thus, λ can be sampled efficiently using algorithm proposed in Ray et al. [2019].

D.3 Performance of bsar

Consider the simulation set-up specified in section 4. Figure 10 shows the out-of-sample prediction performance of bsar, developed by Lenk and Choi [2017], and implemented by the **R** package **bsamGP**.

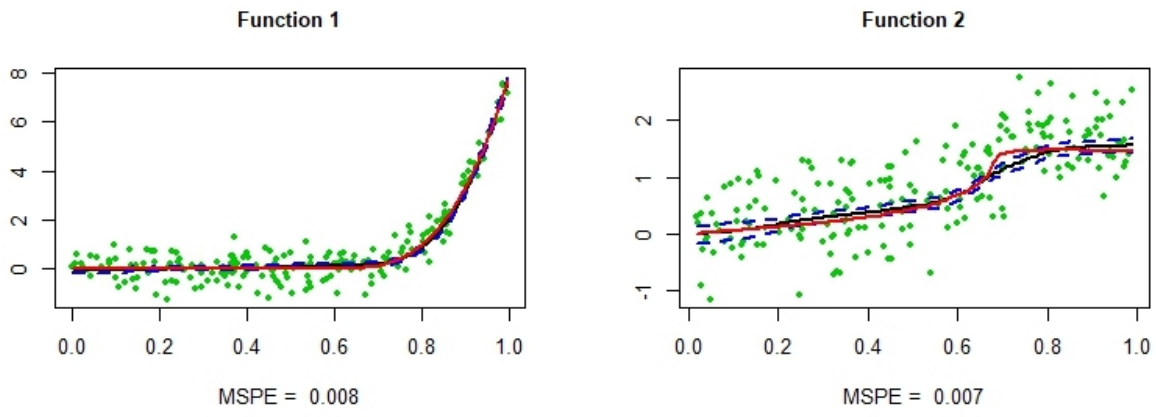

Figure 10: *Figure portraying out-of-sample prediction accuracy using bsar for f_1 and f_2. Red solid curve corresponds to the true function, black solid curve is the mean prediction, the region within two dotted blue curves represent 95% pointwise prediction Interval and the green dots are* 200 *test data points.* MSPE *values corresponding to each of the method are also shown in the plots.*

E Concentration result of the posterior mean μ

We start by introducing some new notations and assumptions. For two variables X, Y, we denote the conditional probability measure, conditional expectation and conditional variance of Y given X as $\mathbb{P}_{Y|X}$ $\mathbb{E}_{Y|X}$, and $\text{var}_{Y|X}$, respectively. For two quantities a, b, we write $a \gtrsim b$ when a is bounded below by a multiple of b. For matrices A, B of the same size, we say $A \leq B$ if $B - A$ is positive semi-definite. In the following, we state the assumptions on the basis choice and prior preferences.

Assumption 3. *We assume the number of basis N and sample size n satisfy $N/n \to 0$ as $N, n \to \infty$. Also, we assume the basis matrix $\Phi_{n\times N}$ satisfies*

$$c_1(n/N)\, I_N \leq \Phi^{\mathrm{T}}\Phi \leq c_2(n/N)\, I_N,$$

for some constants $0 < c_1 < c_2 < \infty$.

For an example of basis that satisfies Assumption 3, we take a qth ($q \geq 2$) order B-Spline basis function associated with $N - q$ knots. Moreover, under mild conditions, it can be shown that the optimal order of the number of basis $N \asymp n^c$ for some $c \in (0, 1)$ in the regression setting [Yoo et al., 2016].

Assumption 4. *For the prior distribution $\theta \sim \mathcal{N}(\mathbf{0}, \Omega_N)$ with N satisfying Assumption 3, we assume the covariance matrix Ω_N satisfies $\lambda_{\min}(\Omega_N) \gtrsim (N/n)$.*

Now we are ready to state the concentration result of the posterior center μ_N.

Proposition 4. *Under Assumption 3 and Assumption 4, for the truncated normal posterior $\mathcal{N}_{\mathcal{C}}(\mu_N, \Sigma_N)$ in §3 of the main manuscript, with at least probability $1 - 2\,N^{-2}$ with respect to $\mathbb{P}_{Y|X}$, we have*

$$\|\mu_N\|_\infty \leq \epsilon_N,$$

for $\epsilon_N \geq 2\,(c_2/c_1^2)^{1/2}\,(N \log N/n)^{1/2}$ with c_1, c_2 defined in Assumption 3.

Proof. Under model (3.2) with true coefficient vector $\theta_0 = \mathbf{0}$ we have $Y \mid \theta_0, X \sim \mathcal{N}(\mathbf{0}_n, I_n)$. We henceforth write the posterior center $\mu(Y) = \Sigma_N \Phi^{\mathrm{T}} Y$, also we have $\mathbb{E}_{Y|X}\{\mu(Y)\} = \mathbf{0}$ and $\mathrm{var}_{Y|X}\{\mu(Y)\} = \Sigma_N \Phi^{\mathrm{T}} \Phi \Sigma_N$. Further, we denote $\sigma_j^2 = \mathrm{var}_{Y|X}\{\mu_j(Y)\}$ for $j = 1, \ldots, N$. For basis matrix Φ and Ω_N satisfying Assumption 3 and Assumption 4 separately, we have

$$c_1(n/N) \leq \lambda_{\min}(\Omega_N^{-1} + \Phi^T \Phi) \leq \lambda_{\max}(\Omega_N^{-1} + \Phi^T \Phi) \leq (c_2 + D)(n/N).$$

Since under Assumption 4, the prior covariance matrix Ω_N satisfies $\|\Omega_N^{-1}\| \leq D\,(n/N)$ for some constant $D > 0$ and $\lambda_{\min}(\Omega_N^{-1}) \geq 0$. Further, we have

$$\frac{c_1}{(c_1 + D)^2}(N/n)\, I_N \leq \Sigma_N \Phi^{\mathrm{T}} \Phi \Sigma_N \leq \frac{c_2}{c_1^2}(N/n)\, I_N. \tag{E.1}$$

We define $\sigma_{\max}^2 = \max_{j \leq N}\{\sigma_j^2\}$, then (E.1) implies $\sigma_{\max}^2 \leq (c_2/c_1^2)(N/n)$. It is well known that $\max_{j \leq N} |\mu_j|$ is a Lipschitz function of μ_j's with the Lipschitz constant $\sigma_{\max}$. We can also upper bound the expectation as

$$\mathbb{E}_{Y|X}\Big(\max_{j \leq N} |\mu_i|\Big) \leq \{2 \log(2N)\}^{1/2} \max_{j \leq N}\{\sigma_j\} \leq M_0 (N \log N/n)^{1/2},$$

where $M_0 = 2(c_2/c_1^2)^{1/2}$.

Thus we take $\epsilon_N \geq 2M_0\,(N\log N/n)^{1/2}$, we have

$$\begin{aligned}\mathbb{P}_{Y|X}(\|\mu_N\|_\infty > \epsilon_N) &\leq \mathbb{P}_{Y|X}\Big\{\Big|\max_{j\leq N}|\mu_j| - \mathbb{E}_{Y|X}\Big(\max_{i\leq N}|\mu_i|\Big)\Big| > \epsilon_N - \mathbb{E}_{Y|X}\Big(\max_{i\leq N}|\mu_i|\Big)\Big\}\\ &\leq \mathbb{P}_{Y|X}\Big\{\Big|\max_{j\leq N}|\mu_j| - \mathbb{E}_{Y|X}\Big(\max_{i\leq N}|\mu_i|\Big)\Big| > \epsilon_N/2\Big\}\\ &\leq 2\exp\{-\epsilon_N^2/(8\,\sigma_{\max}^2)\} \leq 2\,N^{-2}.\end{aligned}$$

Then we have established the result. □

F Auxiliary results

Lemma 3. *(Slepian's lemma) Let X, Y be centered Gaussian vectors on $\mathbb{R}^d$. Suppose $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ for all i, and $\mathbb{E}(X_iX_j) \leq \mathbb{E}(Z_iZ_j)$ for all $i \neq j$. Then, for any $x \in \mathbb{R}$,*

$$\mathbb{P}\Big(\max_{1\leq i\leq d} X_i \leq x\Big) \leq \mathbb{P}\Big(\max_{1\leq i\leq d} Y_i \leq x\Big).$$

We use the Slepian's lemma in the following way in the main document. We have,

$$\mathbb{P}(X_1 \geq 0, \ldots, X_d \geq 0) = \mathbb{P}\Big(\min_{1\leq i\leq d} X_i \geq 0\Big) = \mathbb{P}\Big(\max_{1\leq i\leq d} X_i \leq 0\Big),$$

where the second equality uses $X \stackrel{d}{=} -X$. We use Slepian's inequality to arrive at equation (2.6) in the main document.

Lemma 4. *Let $Z_1, \ldots, Z_N$ be iid $\mathcal{N}(0,1)$ random variables. Then we have*

$$C_1\sqrt{2\log N} \leq \mathbb{E}\max_{i=1,\ldots,N} Z_i \leq \mathbb{E}\max_{i=1,\ldots,N}|Z_i| \leq \sqrt{2\log N}. \tag{F.1}$$

for some constant $0 < C_1 < 1$.

Lemma 5. *(Mill's ratio bound) Let $X \sim \mathcal{N}(0,1)$. We have, for $x > 0$, that*

$$\frac{x}{x^2+1}\,e^{-x^2/2} \leq 1 - \Phi(x) \leq \frac{1}{x}\,e^{-x^2/2},$$

where $\Phi(\cdot)$ is cumulative distribution function of X.

Lemma 6. *(Lemma 2.1 in Bickel and Lindner [2012]) Let matrix A be k-banded, symmetric, and positive definite. We denote $M = \|A\|$ and $m = 1/\|A^{-1}\|$, and for $n \in \mathbb{N}_0$, we define*

$$B_n = \gamma \sum_{j=0}^{n} (I - \gamma A)^j, \tag{F.2}$$

where $\gamma = 2/(M+m)$. Then B_n is a symmetric positive definite (nk)-banded matrix, also, $\|A^{-1} - B_n\| \leq \kappa^{n+1}/m$, $\kappa = (M-m)/(M+m) < 1$.

References

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726. 6, 7

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(12):3828 – 3842. In Memoriam: Evarist Giné. 22

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):1–20. 40

Bickel, P. J., Levina, E., et al. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604. 16

Bickel, P. J., Levina, E., et al. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227. 16

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics*, 65(1):198–205. 2

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148. 6

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press. 22

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(1):90–104. 2

Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: Applications to carcinogenicity studies. *Journal of the American Statistical Association*, 102(480):1158–1171. 2

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics-Theory and Methods*, 19(1):197–203. 6, 9, 10

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480. 4, 17, 18

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298. 22

Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics*, 38(5):961–976. 2, 4, 14, 17

Duan, L. L., Young, A. L., Nishimura, A., and Dunson, D. B. (2020). Bayesian constraint relaxation. *Biometrika*, 107(1):191–204. 22

Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470):618–627. 17

Gasull, A. and Utzet, F. (2014). Approximating Mill's ratio. *Journal of Mathematical Analysis and Applications*, 420(2):1832–1853. 10

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149. 9, 10

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, pages 400–400. 9

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media. 9, 10

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531. 3

Hashorva, E. and Hüsler, J. (2003). On multivariate Gaussian tails. *Annals of the Institute of Statistical Mathematics*, 55(3):507–522. 10

Lenk, P. J. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using Gaussian process priors. *Statistica Sinica*, 27:43–69. 20, 36

Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19:533–597. 10

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317. 22

Lu, D. (2016). A note on the estimates of multivariate Gaussian probability. *Communications in Statistics-Theory and Methods*, 45(5):1459–1465. 10

Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582. 2, 3, 4, 5, 14, 15, 23, 24

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(4):867–884. 2

Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406. 2, 4, 17

Ray, P., Pati, D., and Bhattacharya, A. (2019). Efficient Bayesian shape-restricted function estimation with constrained Gaussian process priors. *arXiv preprint arXiv:1902.04701*. 18, 36

Ruben, H. (1964). An asymptotic expansion for the multivariate normal distribution and Mill's ratio. *Journal of Research of the National Bureau of Standards, Mathematics and Mathematical Physics B*, 68:1. 10

Savage, I. R. (1962). Mill's ratio for multivariate normal distributions. *J. Res. Nat. Bur. Standards Sect. B*, 66:93–96. 10

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(2):166–181. 2

Sidák, Z. (1968). On multivariate normal probabilities of rectangles: their dependence on correlations. *The Annals of Mathematical Statistics*, 39(5):1425–1434. 10

Steck, G. P. (1979). Lower bounds for the multivariate normal Mill's ratio. *The Annals of Probability*, 7(3):547–551. 10

Sun, D. and Berger, J. O. (1998). Reference priors with partial information. *Biometrika*, 85(1):55–71. 22

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205. 22

Vershynin, R. (2018a). *High–Dimensional Probability : An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, United Kingdom New York, NY. 10

Vershynin, R. (2018b). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press. 25, 26

Yoo, W. W., Ghosal, S., et al. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, 44(3):1069–1102. 16, 37