JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2023.2220170

Bayesian Inference Using the Proximal Mapping: Uncertainty Quantification Under Varying Dimensionality

Maoran Xu[a], Hua Zhou[b], Yujie Hu[c], and Leo L. Duan[d]

[a]Department of Statistical Science, Duke University, Durham, NC; [b]Departments of Biostatistics and Computational Medicine, University of California, Los Angeles, CA; [c]Department of Geography, University of Florida, Gainesville, FL; [d]Department of Statistics, University of Florida, Gainesville, FL

ABSTRACT

In statistical applications, it is common to encounter parameters supported on a varying or unknown dimensional space. Examples include the fused lasso regression, the matrix recovery under an unknown low rank, etc. Despite the ease of obtaining a point estimate via optimization, it is much more challenging to quantify their uncertainty. In the Bayesian framework, a major difficulty is that if assigning the prior associated with a p-dimensional measure, then there is zero posterior probability on any lower-dimensional subset with dimension $d < p$. To avoid this caveat, one needs to choose another dimension-selection prior on d, which often involves a highly combinatorial problem. To significantly reduce the modeling burden, we propose a new generative process for the prior: starting from a continuous random variable such as multivariate Gaussian, we transform it into a varying-dimensional space using the proximal mapping. This leads to a large class of new Bayesian models that can directly exploit the popular frequentist regularizations and their algorithms, such as the nuclear norm penalty and the alternating direction method of multipliers, while providing a principled and probabilistic uncertainty estimation. We show that this framework is well justified in the geometric measure theory, and enjoys a convenient posterior computation via the standard Hamiltonian Monte Carlo. We demonstrate its use in the analysis of the dynamic flow network data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2021
Accepted May 2023

KEYWORDS
Concentration of Lipschitz functions; Generalized density; Generalized projection; Hausdorff dimension; Nonexpansiveness

1. Introduction

Modern statistical applications often involve data that are high dimensional. To allow signal recovery under a relatively low sample size, one often needs to assume that the parameter $\theta \in \mathbb{R}^p$ in fact lies in/near some lower dimensional space. Commonly used assumptions include sparsity (Meinshausen and Bühlmann 2006; Zhao, Rocha, and Yu 2006; Meier, Van De Geer, and Bühlmann 2008; Bickel, Ritov, and Tsybakov 2009), low rank (Shen and Huang 2008; Ji et al. 2010), geometric constraints (Goodall and Mardia 1999; Saarela and Arjas 2011), etc. In most cases, the dimensionality d is unknown. For example, we usually do not know the exact rank in the low-rank matrix factorization.

Bayesian framework provides a principled way to quantify the uncertainty on those models. A potential caveat is that if the assigned prior is associated with a p-dimensional continuous measure, then there is zero posterior probability allocated on any of the lower-dimensional subsets with dimension $d < p$. Instead, from a generative perspective, one should first choose a discrete prior to select d, then generate θ within the chosen space. For example, the spike-and-slab prior (Mitchell and Beauchamp 1988) assigns a binomial distribution on d as the number of nonzero coefficients in the variable selection problem; the Bayesian adaptive regression spline uses a Poisson prior on the number of knots d, which determines the rank of the spline matrix (DiMatteo, Genovese, and Kass 2001). On the other hand, the discrete prior creates a highly combinatorial problem, and existing estimation methods such as the Reversible-jump Markov chain Monte Carlo (Green and Hastie 2009) are not very efficient to explore the high posterior probability region.

An appealing alternative is to avoid specifying any low-dimensional prior, but to induce a prior for θ with the measure in $\mathbb{R}^p$ yet having the mass concentrated near some low-dimensional sets. Specifically, the key is to reparameterize the parameter θ as some transformation of a sparse vector β, and then assign simple continuous shrinkage prior on β to favor near-zero values. For example, in spline regression, one uses β as the sparse weights in the linear combination of some basis functions. In this category, there is a rich literature covering tasks of variable selection (Park and Casella 2008; Carvalho, Polson, and Scott 2009; Rockova and George 2018), matrix decomposition (Bhattacharya and Dunson 2011; Legramanti, Durante, and Dunson 2020), functional data analysis (Shin, Bhattacharya, and Johnson 2020), covariance estimation (Li, Craig, and Bhadra 2019; Kastner 2019), among others.

Clearly, this strategy has its limitations—when we cannot reparameterize the low-dimensional sets of θ, the prior specification becomes awkward. This is not uncommon. For example, the fused lasso (Tibshirani et al. 2005) is a frequentist regularization very popular in the image/signal processing, which assumes sparsity not only in the parameter $\theta \in \mathbb{R}^p$, but also in the

CONTACT Maoran Xu maoranxu@ufl.edu Department of Statistical Science, Duke University, Durham, NC.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

$(p-1)$ increments between the neighboring elements $(\theta_{j+1}-\theta_j)$. Although we could imagine assigning some shrinkage prior on $\beta = D\theta \in \mathbb{R}^{2p-1}$ with D the corresponding matrix, such a prior is ill-defined: as D is not invertible, we cannot compute θ from β; as β resides in the column span of D, it has a dimension p, which is less than $(2p-1)$—therefore, the shrinkage prior one blindly assigned would be in fact an incomplete density for a degenerate measure, making it difficult to calibrate the hyper-parameters within and assess the effects of the prior regularization.

Motivated to generalize the Bayesian approaches for handling most of the low-dimensional regularizations (including potentially complicated ones), while avoiding the caveats of having to explicitly specify a discrete prior, we consider a "projection"-style approach. Starting from a continuous prior for β with measure in $\mathbb{R}^p$, we transform it into θ using a special mapping, so that θ has an induced prior on several low-dimensional sets. The projection idea was previously considered in several cases, such as the mixture of components with different dimensions (Petris and Tardella 2003), the isotonic regression (Dunson and Neelon 2003), monotone curve fitting (Lin and Dunson 2014) and more generally, constrained space modeling (Sen, Patra, and Dunson 2018). Nevertheless, in this article, we explore a much more general transformation known as the proximal mapping—it not only includes common Euclidean projection to a constrained set, but also useful non-projection transformation such as soft-thresholding, nuclear norm control, set expansion, etc. This mapping has been well studied in the optimization literature, with appealing properties that are convenient for canonical Bayesian inference, such as in the concentration of measure and convenient computation via the Hamiltonian Monte Carlo. We will carefully justify this prior via the geometric measure theory and demonstrate the strengths via several examples.

2. Method

2.1. Background on the Proximal Mapping

We first provide a brief review on the proximal mapping and motivate its use as a transformation tool. Let θ be the parameter of interest in a certain space Θ, with $\Theta \subseteq \mathbb{R}^p$. With another parameter $\beta \in \boldsymbol{\beta} \subseteq \mathbb{R}^p$, the *proximal mapping* is a transform of β to θ:

$$\theta = \text{prox}_{\lambda g}(\beta) := \underset{z\in\Theta}{\arg\min} \left\{ \lambda g(z) + \frac{1}{2}\|z-\beta\|_2^2 \right\}, \quad (1)$$

where g is a lower semicontinuous and convex function, and $\lambda > 0$ is a scalar as a hyper parameter. This effectively induces a parameter space:

$$\Theta_{\lambda g} = \{\text{prox}_{\lambda g}(\beta) : \beta \in \mathbb{R}^p\}.$$

For an intuitive understanding, the proximal mapping could be viewed as a generalized projection. Given a constrained set C, we can choose $g(z) = \mathcal{X}_C(z)$, the characteristic function of a constrained set taking value 0 if $z \in C$, or ∞ if $z \notin C$. The mapping becomes $\theta = P_C(\beta) = \arg\min_{z\in C}\|z-\beta\|_2^2$, the Euclidean projection of β into the set C. Furthermore, we can replace $\mathcal{X}_C$ with other function for g, leading to a wider class of transformation.

Example 1 (Soft thresholding). Perhaps the most famous example is $g(z) = \|z\|_1$ from lasso (Tibshirani 1996). It has a closed-form proximal mapping known as the soft-thresholding operator $\text{prox}_{\lambda g}(\beta) = \text{sign}(\beta)\max(|\beta|-\lambda, 0)$, with all operations carried out element-wise. The induced parameter space $\Theta_{\lambda g}$ is in fact the union of multiple sets with varying dimensions: $\{\theta \in \mathbb{R}^p : \theta_j = 0 \text{ for } j \in S, |S| = d\}$, where S is some index set and $d \in (0, 1, \ldots, p)$, each is a Euclidean subspace of dimension $(p-d)$ — conveniently, we do not need to explicitly specify the dimension d, since it is automatically induced through the transformation.

This suggests that the proximal mapping can be used as a convenient tool to develop priors on lower-dimension subsets. We now list a few useful proximal mappings in Table 1. In addition, the proximal mapping allows us to easily consider multiple constraints or g functions, since the intersection of convex sets and summation of convex functions are still convex. The general form can be computed using the alternative direction of method of multipliers algorithm Bertsekas (2014), and we will demonstrate one case in the data application. For example, consider θ being sparse while constrained in some convex set; this would be challenging to model for conventional approaches due to the lack of reparameterization.

2.2. Proximal Prior

We now use the above in a Bayesian modeling framework. Suppose we have data generated from a likelihood $L(y;\theta)$, where we want to assign a prior on θ in some space with dimensionality smaller or equal to p. We use the following generative

Table 1. Some useful proximal mappings.

Space of β	$g(z)$	$\text{prox}_{\lambda g}(\beta)$	Usage
$\mathbb{R}^p$	$\mathcal{X}_C$, C convex set	$P_C(\beta)$	Projection to a set [See Table 6.1 of Beck (2017) for an expanded list]
$\mathbb{R}^p$	$\|z\|_1$	$\text{sign}(\beta)\max(\|\beta\|-\lambda, 0)$, computed element-wise	Sparsity
$\{\beta \in \mathbb{R}^{k\times k}$, positive semidefinite$\}$	$\|Z\|_*$, nuclear norm	$U\Lambda_0 V^{\mathrm{T}}$, with $\beta = U\Lambda V^{\mathrm{T}}$ the singular value decomposition, $(\Lambda_0)_{ii} = \max(0, \Lambda_{ii}-\lambda)$	Low rank
$\mathbb{R}^{m\times n}$	$\|Z\|_{2,1} = \sum_i \sqrt{\sum_j Z_{ij}^2}$	$[\beta_i \max(1-\frac{\lambda}{\|\beta_i\|_2}, 0)]_{i=1}^m$ with β_i as the ith row	Row / group sparsity
$\mathbb{R}^p$	$\|Dz\|_1$ with $D \in \mathbb{R}^{k\times p}$	Solvable via the alternating direction method of multipliers	Fused lasso, convex clustering
$\mathbb{R}^p$	$\text{dist}_C(z) = \inf_{x\in C}\|z-x\|_2$, distance to a set	$aP_C(\beta) + (1-a)\beta$ with $a = \min\{\lambda/\text{dist}_C(\beta), 1\}$	Set expansion to C

process for θ:

$$\begin{aligned}\beta &\sim \Pi^0_{\boldsymbol{\beta}},\\ \lambda &\sim \Pi^0_{\boldsymbol{\lambda}},\\ \theta &= \text{prox}_{\lambda g}(\beta),\end{aligned} \tag{2}$$

where $\Pi^0_{\boldsymbol{\beta}}$ is a continuous distribution in $\mathbb{R}^p$, such as the non-degenerate Gaussian $\beta \sim \text{N}(\mu, \Sigma)$ and we use $\Pi^0_{\boldsymbol{\lambda}}$ to denote a generative distribution for $\lambda > 0$.

Here g is a convex and lower-semicontinuous function such as those in Table 1. Potentially, g could be known up to some other hyper-parameter γ; in that case, we denote it by g_γ and use $\Pi^0_{\boldsymbol{\gamma}}$ as the prior for γ. For a clear notation, we use bold subscript such as in $\Pi^0_{\boldsymbol{\theta}}$ as a book-keeping index to refer to the variable whose prior is being defined.

It is not hard to see that $\lambda g(z) + 2^{-1}\|z - \beta\|_2^2$, as the combination of the convex g and a quadratic term, is strictly convex with a unique minimizer. Therefore, each β maps to a unique θ, hence, we have a measurable mapping, which means we have a valid prior distribution for θ using (2). We denote the conditional prior distribution for θ as $\Pi^0_{\boldsymbol{\theta}}(\theta \mid \lambda, \gamma)$, and its marginal distribution as $\Pi^0_{\boldsymbol{\theta}}(\theta)$ after integrating out γ and λ. For convenience, we will refer to either form as a "proximal prior."

We first show that, a proximal prior can produce a convenient equivalence to a hierarchical prior of first selecting a low-dimensional set and then assigning a conditional density within this set. We denote the space induced by $\text{prox}_{\lambda g}(\boldsymbol{\beta})$ as Θ, and assume that it can be partitioned into $\Theta = \Theta^0 \cup \Theta^1 \cup \cdots \cup \Theta^p$, where Θ^k denotes a k-dimensional subset of Θ, and $\Theta^j \cap \Theta^k = \varnothing$ if $j \neq k$ (this can be achieved even if a higher dimensional set $\tilde{\Theta}^k$ overlaps/contains a lower-dimensional set Θ^j, we set $\Theta^k = \tilde{\Theta}^k \setminus \cup_{j=1}^{k-1}\Theta^j$). Then the prior kernel (a mix of density and mass functions) evaluated at $\theta = t$ can be written as

$$\Pi^0_{\boldsymbol{\theta}}(t) = \sum_{k=0}^{p} \Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)\mathbf{1}(t \in \Theta^k)\text{pr}(\theta \in \Theta^k), \tag{3}$$

where $\sum_{k=0}^p \text{pr}(\theta \in \Theta^k) = 1$ and $\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)$ is a conditional density that integrates to 1 over $t \in \Theta^k$ using an appropriate k-dimensional integral with respect to some proper measure λ^k, denoted by $\int_{\Theta^k} \Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)\lambda^k(\text{d}t) = 1$. The integral and measure will be formally defined in the theory section.

Therefore, from a generative view, the above can be understood as first picking a set Θ^k with probability $\text{pr}(\theta \in \Theta^k)$, then drawing a value t within the space of Θ^k. This includes those corner cases where $\text{prox}_{\lambda g}$ cannot map to some dimensional sets: that is, for some k's, we can have $\text{pr}(\theta \in \Theta^k) = 0$.

Accordingly, with $L(y;\theta)$ the likelihood, the posterior of θ can be derived as

$$\Pi(\theta = t \mid y) = \sum_{k=0}^{p} \underbrace{z_k^{-1} L(y;t)\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)\mathbf{1}(t \in \Theta^k)}_{\Pi(\theta = t \mid \theta \in \Theta^k, y)} \underbrace{\frac{z_k \text{pr}(\theta \in \Theta^k)}{\sum_{k=0}^d z_k \text{pr}(\theta \in \Theta^k)}}_{\text{pr}(\theta \in \Theta^k \mid y)}, \tag{4}$$

where $z_k = \int_{\Theta^k} L(y;\theta = t)\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)\lambda^k(\text{d}t)$. We assume posterior propriety almost everywhere, such that $z_k < \infty$ for all $k : \text{pr}(\theta \in \Theta^k) > 0$.

The proximal priors simplify these procedures. Using the transformation $\theta = \text{prox}_{\lambda g}(\beta)$, for any measurable set $\mathcal{A} \in \Theta$,

$$\begin{aligned}&\text{pr}(\theta \in \mathcal{A} \mid y)\\ =&\sum_{k=0}^{p} z_k^{-1} \left\{ \int_{\mathcal{A}\cap\Theta^k} L(y;t)\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)\text{d}t \right\} \frac{z_k \text{pr}(\theta \in \Theta^k)}{\sum_{k=0}^d z_k \text{pr}(\theta \in \Theta^k)}\\ \overset{(a)}{=}&\sum_{k=0}^{p} z_k^{-1} \left[\int_{\text{prox}_{\lambda g}^{-1}(\mathcal{A}\cap\Theta^k)} L\left\{y; \text{prox}_{\lambda g}(b)\right\} \frac{\Pi^0_{\boldsymbol{\beta}}(b)\mathbf{1}[\text{prox}_{\lambda g}(\beta) \in \Theta^k]}{\text{pr}[\text{prox}_{\lambda g}(\beta) \in \Theta^k]}\text{d}b \right]\\ &\frac{z_k \text{pr}(\theta \in \Theta^k)}{\sum_{k=0}^d z_k \text{pr}(\theta \in \Theta^k)}\\ =&\frac{1}{\sum_{k=0}^d z_k \text{pr}(\theta \in \Theta^k)} \sum_{k=0}^{p} \int_{\text{prox}_{\lambda g}^{-1}(\mathcal{A}\cap\Theta^k)} L\left\{y; \text{prox}_{\lambda g}(b)\right\} \Pi^0_{\boldsymbol{\beta}}(b)\text{d}b,\end{aligned}$$

where in (a) we mean that $\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^k)$ contains the Jacobian term in the change-of-variables $t = \text{prox}_{\lambda g}(b)$ (details of the Jacobian calculation provided in the theory section). At any given $\beta = b$, we can omit the integral and summation, and obtain a remarkably simple posterior density of β:

$$\Pi(\beta = b \mid y) \propto L\left\{y; \text{prox}_{\lambda g}(b)\right\} \Pi^0_{\boldsymbol{\beta}}(b). \tag{5}$$

Remark 1. To clarify, although the hierarchical form of $\Pi^0_{\boldsymbol{\theta}}$ provides a nice interpretation to our proximal prior, such an equivalence is not strictly necessary for the proximal modeling framework to work. To be rigorous, the above equivalence requires a few regularity conditions, to be formalized in the theory section.

Therefore, compared to (4), the posterior density (5) is much easier for Bayesian applications. This also suggests a new strategy of "data augmentation using optimization" [instead of marginalization as in Tanner and Wong (1987)]—if we can write the parameter θ as some proximal mapping from β, then we can sample β first as an augmented variable; after sampling, we compute $\theta = \text{prox}_{\lambda g}(\beta)$ and discard the information from β.

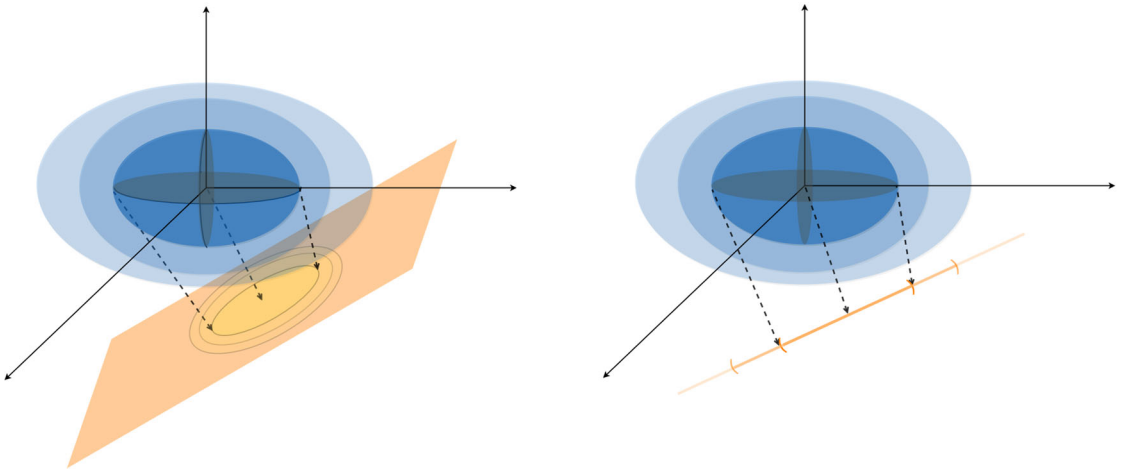
We now use one example to illustrate the equivalence.

Example 2 (Affinely constrained prior under varying rank). Suppose we want to assign a prior for θ in a set of affine constraints $C = \{\theta \in \mathbb{R}^p : A^{\text{T}}\theta = b\}$, where $A^{\text{T}} \in \mathbb{R}^{m\times p}$ is another parameter, with $m < p$ and $b \in \text{Col}(A^{\text{T}})$ the column space of A^{T} (so that C is not empty). Since A is not fixed, we do not know the rank of A, hence, not the dimensionality of C. Using the proximal prior with $g_A(z) = 0$ if $A^{\text{T}}\theta = b, g_A(z) = \infty$ otherwise (hence, $\text{prox}_{\lambda g_A}$ is invariant to any finite value of $\lambda > 0$), and $\beta \sim \text{N}(\mu, \Sigma)$, we have a closed-form proximal mapping

$$\theta = \text{prox}_{\lambda g}(\beta) = \beta - A(A^{\text{T}}A)^{-}(A^{\text{T}}\beta - b),$$

where $(\cdot)^-$ is the Moore-Penrose inverse. We illustration this mapping in Figure 1.

The θ-marginal proximal prior is a discrete mixture over different rank of A:



(a) Given $\text{rank}(A) = 1$, the mapping $\theta = \text{prox}_{\lambda g}(\beta)$ creates a two-dimensional prior on the hyperplane (orange).

(b) Given $\text{rank}(A) = 2$, the mapping $\theta = \text{prox}_{\lambda g}(\beta)$ creates a one-dimensional prior on the line (orange).

Figure 1. Illustrative example of constructing a prior on an affinely constrained set $C = \{\theta : A^{\mathrm{T}}\theta = b\}$. A challenge arises when the rank(A) is unknown, the dimensionality of the prior is unknown. The proximal prior bypasses this hurdle by transforming a continuous prior (blue) into the constrained space (orange), without the need to explicitly specify the dimensionality.

$$\begin{aligned}\Pi^0_{\boldsymbol{\theta}}(\theta) &= \int \Pi^0_{\boldsymbol{\theta}}(\theta \mid A)\, \Pi^0_{\boldsymbol{A}}(A)\mathrm{d}A \\ &= \sum_{d=0}^{p} \text{pr}^0(\text{rank}(A) = d) \underbrace{\frac{\int_{A:\text{rank}(A)=d} \Pi^0_{\boldsymbol{\theta}}(\theta \mid A)\, \Pi^0_{\boldsymbol{A}}(A)\mathrm{d}A}{\int_{A:\text{rank}(A)=d} \Pi^0_{\boldsymbol{A}}(A)\mathrm{d}A}}_{\Pi^0_{\boldsymbol{\theta}}[\theta|\text{rank}(A)=d]}\end{aligned}$$

and $\Pi^0_{\boldsymbol{\theta}}(\theta \mid A)$ is the degenerate Gaussian density with mean $A\left(A^{\mathrm{T}}A\right)^{-1} b + P_{A^\perp}\mu$ and covariance $P_{A^\perp}\Sigma P_{A^\perp}$, where $P_{A^\perp} = I - A\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}$. Although the summation may not have a closed-form, the weights and conditional density can be tractable in applications.

For illustration, we consider a Bayesian envelope linear regression for multivariate response $Y_i \in \mathbb{R}^p$:

$$\begin{aligned} &Y_i = \mu + \tilde{\Theta}X_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \text{N}\left(0_p,\ \Gamma\Omega\Gamma^{\mathrm{T}} + \Gamma_*\Omega_*\Gamma_*^{\mathrm{T}}\right) \\ &\theta = \Gamma\eta,\end{aligned}$$

for $i = 1, \ldots, n$, each covariate $X_i \in \mathbb{R}^m$ (with $p \le m$), and the noise $\epsilon_i \in \mathbb{R}^p$; $[\Gamma\ \Gamma_*]$ together from a $p \times p$ orthonormal matrix, with Γ a $p \times u$ submatrix, $\eta \in \mathbb{R}^{u\times m}$ full rank, and both Ω and Ω_* positive definite matrices. The regression coefficient matrix $\tilde{\Theta} \in \mathbb{R}^{p\times m}$ is of rank u, with u unknown. The motivation is that by making $\tilde{\Theta}X_i$ in the subspace spanned by the leading eigenvectors of the covariance, $\Gamma_*^{\mathrm{T}}Y_i$ is small in magnitude and independent from X_i, leading to a sufficient dimension reduction (Cook, Li, and Chiaromonte 2010). For Bayesian inference, Khare, Pal, and Su (2017) proposed to use a matrix-Bingham prior on Γ with a pre-specified u, so that it has conjugate forms in a Gibbs sampler for posterior computation. On the other hand, since Γ is in an orthogonal and low-rank space, it is difficult to generalize to other forms of prior such as letting u vary.

Using the affine constraint proximal mapping, we can bypass these challenges. We reparameterize $\Gamma_*\Omega_*\Gamma_*^{\mathrm{T}} = AA^{\mathrm{T}}$, $\Gamma\Omega\Gamma^{\mathrm{T}} = P_{A^\perp}WP_{A^\perp}^{\mathrm{T}}$ with $\text{rank}(A) = p-u$, W positive definite, and $\tilde{\Theta}$ by a linear constraint $A^{\mathrm{T}}\tilde{\Theta} = O$. The proximal mapping yields $\tilde{\Theta} = P_{A^\perp}B$, with $B \in \mathbb{R}^{p\times m}$ the matrix form for β. Using matrices $Y^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, $X^{\mathrm{T}} \in \mathbb{R}^{n\times m}$, we rewrite the envelope regression likelihood as

$$\begin{aligned} &L(Y; A, W, B, \mu) \\ &\propto \exp\left(-\frac{1}{2}\text{tr}\left\{[Y^{\mathrm{T}} - 1\mu^{\mathrm{T}} - X^{\mathrm{T}}(P_{A^\perp}B)^{\mathrm{T}}]\right.\right. \\ &\qquad \left.\left.(P_{A^\perp}WP_{A^\perp}^{\mathrm{T}})^{-}[Y - \mu 1^{\mathrm{T}} - (P_{A^\perp}B)X]\right\}\right) \\ &\times |A^{\mathrm{T}}A|^{-n/2}\exp\left(-\frac{1}{2}\text{tr}\left\{[Y^{\mathrm{T}} - 1\mu^{\mathrm{T}}](AA^{\mathrm{T}})^{-}[Y - \mu 1^{\mathrm{T}}]\right\}\right) \\ &\quad |P_{A^\perp}WP_{A^\perp}^{\mathrm{T}}|_+^{-n/2},\end{aligned}$$

where $|\cdot|_+$ is the pseudo-determinant. To complete the proximal prior specification, we use $B \in \mathbb{R}^{p\times m}$ and $R \in \mathbb{R}^{p\times p}$, with their elements iid from $\text{N}(0, 1)$ (hence, full rank almost surely); then we set $A = R\Lambda$, and $\Lambda = \text{diag}(\Lambda_{i,i})_{i=1}^p$ with $\Lambda_{i,i} = z_i 1(z_i > \rho_q)$ and $z_i \sim \text{Exp}(1)$, and ρ_q the q-quantile of $\text{Exp}(1)$. We provide a numerical simulation in S1.3 of the Supplementary Materials.

Focusing on the θ-marginal prior, we have: (a) the mixture weight $\text{pr}^0[\text{rank}(A) = p - u] = \binom{p}{u}q^u(1-q)^{p-u}$, where q can be specified as a priori; (b) the conditional density containing

$$\Pi^0_{\boldsymbol{\theta}}(\theta \mid A) = (2\pi)^{-u/2}|P_{A^\perp}|_+^{-1/2}\exp(-\theta^{\mathrm{T}}P_{A^\perp}^{-}\theta/2)$$

$$= (2\pi)^{-u/2} \exp(-\|\theta\|^2/2) 1(A^{\mathrm{T}}\theta = 0).$$

The second line is due to $P_{A^\perp}$ being idempotent $\theta = P_{A^\perp}\beta = P_{A^\perp}\theta$, $P_{A^\perp}P_{A^\perp}^{-}P_{A^\perp} = P_{A^\perp}$, $\theta^{\mathrm{T}}P_{A^\perp}\theta = \theta^{\mathrm{T}}\theta$ and $|P_{A^\perp}|_+ = 1$. This density is invariant to scaling of A.

Remark 2. To clarify, we use the above example (with a relatively simple $\Pi^0_{\boldsymbol{\theta}}$) to illustrate the equivalence between the hierarchical specification of $\Pi^0_{\boldsymbol{\theta}}$ and the continuous-$\Pi^0_{\boldsymbol{\beta}}$-and-mapping specification. In general cases, the former $\Pi^0_{\boldsymbol{\theta}}$ may be intractable due to the lack of closed-form, hence, motivating the proximal mapping strategy as proposed in this article.

In Example 2, if we intuitively compare the two distributions before and after the mapping, it reduces (or at least retains) the distance to the center $\|\theta - \mathrm{prox}_{\lambda g}(\mu)\|_2 \le \|\beta - \mu\|_2$. The property shown in this example is known as the "non-expansiveness," which in fact holds for all proximal mappings:

$$\|\mathrm{prox}_{\lambda g}(\beta_1) - \mathrm{prox}_{\lambda g}(\beta_2)\|_2 \le \|\beta_1 - \beta_2\|_2,$$

for any β_1, β_2 in the domain of $\mathrm{prox}_{\lambda g}$. This is in particular meaningful for Bayesian inference, as it conveniently controls the concentration of measure for θ.

Theorem 1. If the data y come from a distribution $\mathcal{F}_{\theta^*}$ with a fixed parameter θ^*, and any $\epsilon \in (0,1)$, the posterior distributions of θ and β satisfy

$$\begin{aligned}&\mathrm{pr}(\|\theta - \theta^*\| > \epsilon \mid y)\\ &\le \int \min_{\beta^*:\mathrm{prox}_{\lambda g_\gamma}(\beta^*)=\theta^*} \mathrm{pr}\left[\left\|\beta - \beta^*\right\| > \epsilon \mid y, \lambda, \gamma\right]\\ &\quad \Pi(\lambda, \gamma \mid y) d(\lambda, \gamma).\end{aligned}$$

In addition, if $\mathrm{tr}[\mathrm{cov}(\beta \mid y)] < \infty$, then

$$\mathrm{tr}[\mathrm{cov}(\theta \mid y)] \le \mathrm{tr}[\mathrm{cov}(\beta \mid y)].$$

Using the envelope regression example, we know that

$$\begin{aligned}\Pi(B \mid y, A, W) \propto \exp\big\{ &-(1/2)\mathrm{tr}[XX^{\mathrm{T}}B^{\mathrm{T}}P_{A^\perp}(P_{A^\perp}WP_{A^\perp}^{\mathrm{T}})^{-}P_{A^\perp}B]\\ &+ \mathrm{tr}(B^{\mathrm{T}}B) - 2\mathrm{tr}[X(Y^{\mathrm{T}} - 1\mu^{\mathrm{T}})(P_{A^\perp}WP_{A^\perp}^{\mathrm{T}})^{-}P_{A^\perp}B]\big\},\end{aligned}$$

which is a multivariate Gaussian for $\mathrm{vec}(B)$. On the other hand, since we know $\tilde{\Theta} = P_{A^\perp}B$, we know for any given A, $\|\tilde{\Theta} - \tilde{\Theta}^*\|_F^2 = \mathrm{tr}[(B - B^*)^{\mathrm{T}}P_{A^\perp}P_{A^\perp}(B - B^*)] \le \|B - B^*\|_F^2$ with $P_{A^\perp}B^* = \tilde{\Theta}^*$, due to $P_{A^\perp}$ being idempotent and having eigenvalues equal to either 1 or 0.

2.3. Prior Specification on λ

In the proximal mapping (1), the hyper-parameter λ plays an important role, hence, we need to carefully choose its prior. To first obtain some intuition, note when $\lambda \to 0$, we have $\mathrm{prox}_{\lambda g}(\beta) \to \beta$ if $g(z) < \infty$ for all z, the identity mapping; when $\lambda \to \infty$, we have $\mathrm{prox}_{\lambda g}(\beta) \to \arg\min_z g(z)$. Therefore, as λ increases, θ becomes farther away from β, hence, the distribution $\Pi^0_{\boldsymbol{\beta}}(\beta)$ gets more "deformed" at a larger λ. We now formalize this deformation intuition, while relaxing the finite-valuedness of g. For conciseness, we postpone all the proofs in the Appendix.

Theorem 2 (Monotonicity of deformation in λ). For any function g with range $\mathbb{R}\cup\{\infty\}$, if $0 < \lambda_1 < \lambda_2$, then $\|\beta - \mathrm{prox}_{\lambda_1 g}(\beta)\|_2 \le \|\beta - \mathrm{prox}_{\lambda_2 g}(\beta)\|_2$.

This result means that we can find a measurement between 0 and 1 to quantify the deformation:

$$\omega_\lambda := \frac{\mathbb{E}_{\boldsymbol{\beta}}\|\beta - \mathrm{prox}_{\lambda g}(\beta)\|_2}{\mathbb{E}_{\boldsymbol{\beta}}\|\beta - \lim_{\lambda^*\to\infty}\mathrm{prox}_{\lambda^* g}(\beta)\|_2}, \tag{6}$$

where the expectation is taken with respect to the prior of β.

When lacking prior knowledge on λ, we can use a Beta prior on $w \in (0,1)$ and solve for λ:

$$\omega \sim \mathrm{Beta}(a_\omega, b_\omega), \qquad \lambda = \min_{x>0}(x : \omega_x = \omega). \tag{7}$$

In this article, we use a non-informative $a_\omega = b_\omega = 1$. As a toy example, let β be univariate with a finite variance, using the proximal mapping with $g(z) = z^2/2$, we have $\mathrm{prox}_{\lambda g}(\beta) = \beta/(1+\lambda)$. Therefore, we have $\lambda = (1-\omega)/\omega$ with an induced prior $\Pi^0_{\boldsymbol{\lambda}}(\lambda) = 2/(1+\lambda)^2$ for $\lambda > 0$.

In more general cases, (7) often cannot be solved analytically. However, we can numerically compute a prior for λ, using a strategy similar to Berger et al. (2009)—for K chosen points $\lambda_1, \ldots, \lambda_K$ in $(0,\infty)$, we can use the empirical estimates of the expectation based on simulated $\beta \sim \Pi^0_{\boldsymbol{\beta}}(\beta)$, and solve for $\omega_1, \ldots, \omega_K$; afterwards, we can easily interpolate to obtain the λ associated with any ω.

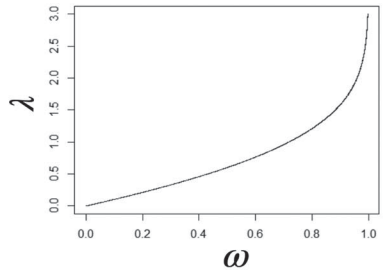
Example 1 ((Continued). Soft-thresholding prior). To illustrate, we compute the prior of λ for the soft-thresholding prior based on $\theta = \mathrm{sign}(\beta)\max(|\beta| - \lambda, 0)$. Based on $\beta \in \mathbb{R}^p$ and $\beta \sim \mathrm{N}(0, I_p)$, we compute the prior density of λ and plot it in Figure 2.

In this section, we discussed the choice of $\Pi^0_{\boldsymbol{\lambda}}$ with the generality of all possible g and $\mathrm{prox}_{\lambda g}$ in mind. On the other hand, for some specific case such as $g(z) = \|z\|_1$ and soft-thresholding $\mathrm{prox}_{\lambda g}$, there is a connection to some existing prior in the literature, such as the classic spike-and-slab prior. For example, if $\Pi^0_{\boldsymbol{\beta}}(\beta) \propto \exp(-\|\beta\|_1/\alpha)$ and λ to be the ω-quantile of $\mathrm{Exp}(\alpha^{-1})$, then we can obtain a spike-and-slab prior with Laplace slab $\Pi^0_{\boldsymbol{\theta}}(\theta \mid \lambda) = \prod_{j=1}^p [w_\lambda \delta_0(\theta_j) + (1 - w_\lambda)(2\alpha)^{-1}\exp(-\theta_j/\alpha)]$. A closely related discovery is the neuronized prior (Shin and Liu 2021) using truncated activation function, for which there is an equivalence to a spike-and-slab prior with two-normal-product slab. In these cases, there are often alternative choices for $\Pi^0_{\boldsymbol{\lambda}}$ that are justified via large sample theory. Due to the page constraint, we defer the detailed discussion and numerical experiments to S1.1 of the supplementary materials.

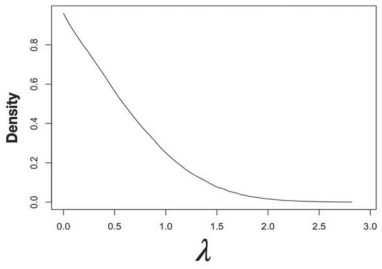
3. Geometric Measure Theory on the Varying Dimensional Sets

3.1. Hausdorff Dimension and Low Dimensional Density

We now give a more rigorous exposition on the distribution induced by the proximal mapping. Without loss of generality, we consider θ as a p-element vector. Since θ may correspond to a measure of a set in the lower dimensional space, the p-dimensional Lebesgue measure of any lower-dimensional set is



(a) Value of λ corresponding to a certain deformation ω. Larger λ leads to larger deformation from $\Pi^0_{\boldsymbol{\beta}}(\beta)$ to $\Pi^0_{\boldsymbol{\theta}}(\theta)$, which can be quantified by $\omega \in [0, 1]$.



(b) The prior for λ, corresponding to a uniform prior on the deformation measurement ω.

Figure 2. Illustration of a numerically computed prior for λ, which controls sparse level for the soft-thresholding mapping $\theta = \text{sign}(\beta)\max(|\beta| - \lambda, 0)$.

zero hence is not useful. We need some tools from the geometric measure theory to address this issue. To start, consider a set $\mathcal{A}$ and suppose we do not know its dimensionality. Instead, we can cover $\mathcal{A}$ with sets B_i's, each B_i has its diameter $\text{diam}(B_i) = \sup\{|x - y| : x, y \in B_i\} \le \delta$. We call any $\bigcup_i B_i \supset \mathcal{A}$, $\text{diam}(B_i) < \delta$ as a δ-covering of $\mathcal{A}$.

Then we take the infimum over all δ-coverings of $\mathcal{A}$, and letting the δ decrease, we obtain the s-dimensional Hausdorff measure of $\mathcal{A}$:

$$\mathcal{H}^s(\mathcal{A}) = \lim_{\delta \to 0} \inf \left\{ \sum_{i=1}^{\infty} \text{diam}(B_i)^s : \text{diam}(B_i) \le \delta, \mathcal{A} \subseteq \bigcup_i B_i \right\}. \tag{8}$$

Intuitively, the above can be taken as the minimum total "volume" of the covering—except s is a parameter that varies.

In fact, $\mathcal{H}^s(\mathcal{A})$ is a nonincreasing function in $s \ge 0$ (Edgar 2007). More importantly, for any Borel $\mathcal{A}$, and $0 < s_1 < s_2$, if $\mathcal{H}^{s_1}(\mathcal{A}) < \infty$ then $\mathcal{H}^{s_2}(\mathcal{A}) = 0$; and if $\mathcal{H}^{s_2}(\mathcal{A}) > 0$ then $\mathcal{H}^{s_1}(\mathcal{A}) = \infty$ [Theorem 6.1.6 (Edgar 2007)]. This means for any Borel set $\mathcal{A}$, there is a unique $s_0 \in [0, \infty) \cup \{\infty\}$ as a transition point, over which the dimensionality drops from ∞ to 0:

$$\mathcal{H}^s(\mathcal{A}) = \infty, \text{ for any } s < s_0;$$
$$\mathcal{H}^s(\mathcal{A}) = 0, \text{ for any } s > s_0.$$

Such an s_0 is referred to as the Hausdorff dimension of $\mathcal{A}$, equivalently:

$$\dim_{\mathcal{H}}(\mathcal{A}) = \inf\{s \ge 0 : \mathcal{H}^s(\mathcal{A}) = 0\}. \tag{9}$$

Note that $\dim_{\mathcal{H}}(\mathcal{A}) \ge 0$ does not have to be an integer; nevertheless, when it is, the Hausdorff measure is proportional to the commonly used s-dimensional Lebesgue measure

$$\lambda^s(\mathcal{A}) = \inf \left\{ \sum_{i=1}^{\infty} \text{vol}(B_i) : \mathcal{A} \in \bigcup B_i, B_i \text{ is an open cube} \right\},$$

via $\lambda^s(\mathcal{A}) = w_s \mathcal{H}^s(\mathcal{A})$, where $w_s = \pi^{s/2}[2^s \Gamma(s/2 + 1)]^{-1}$ due to the volume formula of an s-dimensional ball. In addition, when $s = 0$, $\mathcal{H}^0(\mathcal{A})$ is same as the counting measure.

Now recall that $\text{prox}_{\lambda g}$ is nonexpansive, which leads to the following theorem:

Theorem 3. For any Borel set $\mathcal{A}$ and proximal mapping $\text{prox}_{\lambda g}$, we have

1. $\mathcal{H}^s\{\text{prox}_{\lambda g}(\mathcal{A})\} \le \mathcal{H}^s(\mathcal{A})$ for any $s \ge 0$;
2. $\dim_{\mathcal{H}}\{\text{prox}_{\lambda g}(\mathcal{A})\} \le \dim_{\mathcal{H}}(\mathcal{A})$.

Remark 3. In the above, the statement 2 is particularly useful: it tells us that $\text{prox}_{\lambda g}$ only maps to lower or equal dimensional space.

Now, starting from a probability distribution defined by a certain Radon measure μ in $\mathbb{R}^p$ for some low-dimensional sets in Θ^s, one interesting question is how to differentiate this and obtain a "density," as $\Pi_{\boldsymbol{\theta}}(\theta = t \mid \theta \in \Theta^s)$ used in (3) and (4).

For a point $\theta \in \Theta^s$, the ball $B_r(\theta)$ centered at θ with radius $r > 0$ has the lower and upper s-dimensional derivatives:

$$f^s_{\mu,*}(\theta) = \liminf_{r \to 0} \frac{\mu\{B_r(\theta)\}}{w_s r^s}, \qquad f^{s,*}_{\mu}(\theta) = \limsup_{r \to 0} \frac{\mu\{B_r(\theta)\}}{w_s r^s}.$$

Therefore, if we have the two limits coincide, we would have a definition of an s-dimensional density: $f^s_{\mu}(\theta) = f^s_{\mu,*}(\theta) = f^{s,*}_{\mu}(\theta)$, commonly referred to as the s-density.

Remark 4. To understand the s-density as a generalized concept of "density," for those continuous distributions associated with a p-dimensional Lebesgue measure, such as the nondegenerate Gaussian distribution, the p-density is the probability density function; whereas for the discrete distributions, the 0-density is the same as the probability mass function.

Next, similar to the probability density function, s-density may not always exist. Therefore, it is important to state the two required conditions, s is an integer and θ is in a rectifiable set, as formalized in the following theorem.

Theorem 4 (Besicovitch-Marstrand-Preiss theorem). (Preiss 1987) Let μ be a locally finite Radon measure on $\mathbb{R}^p$, if there exists a real $s \ge 0$ such that $f^s_{\mu}(\theta)$ exists, and it is positive on a set of positive μ-measure, then s must be an integer. On the other hand, let $\mathcal{A} \subset \mathbb{R}^p$ be Borel with $\mathcal{H}^s(\mathcal{A}) \in (0, \infty)$ and s

an integer, then $f^s_\mu(\theta)$ exists for $\theta \in \mathcal{A}$ almost everywhere with respect to $\mathcal{H}^k$, if and only if the set $\mathcal{A}$ is rectifiable.

To explain "rectifiability," a Borel set $\mathcal{A} \subset \mathbb{R}^p$ is rectifiable if there is a countable family of Lipschitz maps $T_i : \mathbb{R}^s \to \mathbb{R}^p$ which cover almost all $\mathcal{A}$ except for sets with zero $\mathcal{H}^s$ measure. That is, intuitively speaking, almost every p-element vector $\theta \in \mathcal{A}$ can be represented as some transformation of $x \in \mathbb{R}^s$—note that this is not the same as a simple reparameterization, as we may obtain $\mathcal{A}$ via multiple f_i's (up to countably many).

3.2. Calculation of the s-Density

We now provide a way to calculate the s-density. Focusing on a subset Θ^k with $\dim(\Theta^k) = k$ and $\boldsymbol{\beta}^k = \text{prox}^{-1}_{\lambda g}(\Theta^k)$. We now transform $\Pi_0(\beta = b \mid \beta \in \boldsymbol{\beta}^k)$ into an s-density with $s = k$.

Theorem 5. If $\boldsymbol{\beta}^k$ is $(\mathcal{H}^p, p)$-rectifiable and $\dim_{\mathcal{H}}(\boldsymbol{\beta}^k) = p$, Θ^k is $(\mathcal{H}^k, k)$-rectifiable and $\dim_{\mathcal{H}}(\Theta^k) = k$, with $p \geq k$, and $J_k\text{prox}_{\lambda g}(\beta) > 0$ a.e.-μ_β. Then the s-density of θ induced by $\text{prox}_{\lambda g}$ is

$$\Pi(\theta = t \mid \theta \in \Theta^k) = \int_{\text{prox}^{-1}_{\lambda g}(t)} \frac{\Pi(\beta = b)/\text{pr}\{\text{prox}_{\lambda g}(\beta) \in \Theta^k\}}{J_k\text{prox}_{\lambda g}(b)} w_{(p-k)} d\mathcal{H}^{p-k}(b),$$

where $J_k\text{prox}_{\lambda g}(b)$ is the k-dimensional Jacobian of $\text{prox}_{\lambda g}$ at b.

Note that if the low-dimensional set Θ^k can be reparameterized as a transformation an k-element vector, then it is possible to change (10) to an integration with respect to an k-dimensional Lebesgue measure.

To explain the assumptions above, a set $\mathcal{A}$ is $(\mathcal{H}^s, s)$-rectifiable when $\mathcal{H}^s(\mathcal{A}) < \infty$, and there is a set as the countable union of Lipschitz images from bounded sets $\mathcal{B} = \bigcup_j \{T_j(\mathcal{C}_j) : \mathcal{C}_i \subset \mathbb{R}^s \text{ and bounded}, T_j \text{ Lipschitz}\}$ such that $\mathcal{H}^s(\mathcal{A} \setminus \mathcal{B}) = 0$. As the result, if s-density exists, we could use (10) when both θ and β are finite.

Morgan (2016) gives the k-dimensional Jacobian $J_kT(x)$ of function $T : \mathbb{R}^n \to \mathbb{R}^m$, differentiable at x. Let $D_T(x) \in \mathbb{R}^{n\times m}$ be the derivative matrix of $T(x)$ at x, with $\{D_T(x)\}_{ij} = \partial T(x)_j/\partial x_i$, $1 \leq i \leq n, 1 \leq j \leq m$, then the k-dimensional Jacobian can be computed as

$$J_kT(x) = \sqrt{\sum_{\substack{M \text{ is the } k\times k \\ \text{submatrix of } D_T(x)}} (\det M)^2}. \tag{10}$$

Note than when $m = n = k$, $J_kT(x) = |\det\{D_T(x)\}|$ as more commonly seen.

Importantly, by Rademacher's theorem (Federer 2014), a Lipschitz function is differentiable almost everywhere. Therefore, $D_{\text{prox}_{\lambda g}}(\beta)$ exists almost surely with respect to μ_β. In the following example, we illustrate the use of the above theorem to compute the s-density for the affinely constrained prior.

Example 2 (The s-density of the affinely constrained prior). Using (10), one can verify that the s-density of affinely constrained prior recovers the "degenerate Gaussian density." Starting from $\beta \sim \text{N}(\mu, \Sigma)$ and let us assume A is $p \times d$ and $\text{rank}(A) = d$, then it is not hard to compute that $J_{(p-d)}\text{prox}_{\lambda g}(\beta) = 1$. Using the proximal mapping, at $\theta = t$, we have $P_A\beta = t - A(A^{\text{T}}A)^{-1}b$ with $P_A = \{I - A(A^{\text{T}}A)^{-1}A^{\text{T}}\}$, hence, we can integrate over the region $\text{prox}^{-1}_{\lambda g}(t)$ by reparameterizing $\beta = P^-_A\{t - A(A^{\text{T}}A)^{-1}b\} + Ax$ where $x \in \mathbb{R}^d$. Integrating over x, we have the s-density with $s = p - d$:

$$\Pi^0_{\boldsymbol{\theta}}(t \mid \theta \in \Theta^{p-d}) \propto \exp\Big[-\frac{1}{2}\{t - A(A^{\text{T}}A)^{-1}b - P_A\mu\}^{\text{T}} P^-_A\Sigma^{-1}P^-_A\{t - A(A^{\text{T}}A)^{-1}b - P_A\mu\}\Big],$$

which is commonly referred to as the "degenerate density" for a degenerate Gaussian, with its covariance $P_A\Sigma P_A$ having a rank $(p - d)$.

We list a few more examples commonly considered in statistics, where for each we have a guaranteed existence of s-density: regression under linear equality constraints, matrix factorization under low-rank constraint, sparse regression, covariance modeling in positive-definite space, and directional modeling in orthonormal space.

Remark 5. To clarify, the existence of s-density for θ is not necessary in our modeling framework using proximal mapping, since we can always carry out computation using a valid p-dimensional density of β. On the other hand, the existence of s-density would be required if one wants to interpret the prior via an equivalent prior $\Pi^0_{\boldsymbol{\theta}}$ as in (3).

4. Posterior Computation

As shown in (5), when using β instead of θ, the posterior has a simple density on $\mathbb{R}^p$, and the proximal mapping is differentiable almost everywhere with respect to μ_β. Therefore, as long as $\Pi(\theta; y)$ is a continuous and differentiable function in θ almost everywhere, we can use the Hamiltonian Monte Carlo (HMC) for posterior computation. Now we first briefly review the HMC algorithm, then address the gradient calculation for the proximal mapping.

To sample from target distribution $\beta \sim \Pi_{\beta|y}(\cdot)$, the HMC uses an auxiliary momentum variable v and samples from a joint distribution $\Pi(\beta, v) = \Pi(\beta \mid y)\Pi(v)$, where a common choice of $\Pi(v)$ is the density of $\text{N}(0, M)$. Denote $U(\beta) = -\log\Pi(\beta \mid y)$ and $K(v) = -\log\Pi(v) = v^{\text{T}}M^{-1}v/2$, which are referred to as the potential energy and kinetic energy, respectively. The total Hamiltonian energy function is $H(\beta, v) = U(\beta) + K(v)$.

At each state (β, v), a new state is generated by simulating Hamiltonian dynamics, which satisfies the Hamilton's equations:

$$\frac{\partial\beta}{\partial t} = \frac{\partial H(\beta, v)}{\partial v} = M^{-1}v; \qquad \frac{\partial v}{\partial t} = -\frac{\partial H(\beta, v)}{\partial\beta} = \frac{\partial\log\Pi(\beta \mid y)}{\partial\beta}. \tag{11}$$

The exact solution for (11) is often intractable, while we can numerically approximate the evolution by algorithms such as the leapfrog scheme. The leapfrog is a reversible and volume-preserving integrator, which updates the evolution $(\beta^t, v^t) \to$

$(\beta^{t+\epsilon}, v^{t+\epsilon})$ via

$$\begin{aligned} v &\leftarrow v + \frac{\epsilon}{2}\frac{\partial \log \Pi(\beta \mid y)}{\partial \beta}, \\ \beta &\leftarrow \beta + \epsilon M^{-1} v, \\ v &\leftarrow v + \frac{\epsilon}{2}\frac{\partial \log \Pi(\beta \mid y)}{\partial \beta} \end{aligned} \tag{12}$$

for $t = 0, \epsilon, \ldots, L\epsilon$, and sets $(\beta^*, v^*) \leftarrow (\beta^{L\epsilon}, v^{L\epsilon})$. To correct the numeric error due to approximation, (β^*, v^*) is treated as a proposal and accepted with the Metropolis-Hastings (MH) probability

$$\min[1, \exp\{-H(\beta^*, v^*) + H(\beta, v)\}].$$

We now discuss the gradient computation:

$$\frac{\partial \log \Pi(\beta \mid y)}{\partial \beta} = \frac{\partial \text{prox}_{\lambda g}(\beta)}{\partial \beta}\left\{\frac{\partial \log L(y;\theta)}{\partial \theta}\Big|_{\theta=\text{prox}_{\lambda g}(\beta)}\right\} + \frac{\partial \log \Pi^0_{\boldsymbol{\beta}}(\beta)}{\partial \beta}.$$

When $\text{prox}_{\lambda g}(\beta)$ has a closed-form, we can use the automatic differentiation toolbox to calculate the gradient $\partial \text{prox}_{\lambda g}(\beta)/\partial\beta$; on the other hand, when the closed-form does not exist, some numeric approximation is needed.

Note that the partial gradient is $\partial \text{prox}_{\lambda g}(\beta)/\partial\beta_j = \lim_{\epsilon\to 0}\{\text{prox}_{\lambda g}(\beta + e_j\epsilon) - \text{prox}_{\lambda g}(\beta)\}/\epsilon$ with e_j is the standard basis with the jth element equal to one, and all others equal to zero; using a small ϵ gives us the finite difference approximation. Nevertheless, when β is high dimensional, this would involve $(p+1)$ times of calculating the proximal mapping, which can be computationally prohibitive. To solve this problem, we follow Spall (1992) and use the simultaneous perturbation stochastic approximation:

$$\frac{\partial \text{prox}_{\lambda g}(\beta)}{\partial \beta_j} \approx \frac{1}{m}\sum_{k=1}^{m}\frac{\{\text{prox}_{\lambda g}(\beta + \Delta^{(k)}\epsilon) - \text{prox}_{\lambda g}(\beta)\}}{\Delta_j^{(k)}\epsilon}, \tag{13}$$

for $j = 1, \ldots, p$, where $\Delta^{(k)} = \{\Delta_1^{(k)}, \ldots, \Delta_p^{(k)}\}$ has each $\Delta_j^{(k)} \in \{-1, 1\}$ independently generated using $\text{pr}\{\Delta_j^{(k)} = 1\} = \text{pr}\{\Delta_j^{(k)} = -1\} = 0.5$. The right hand side is based on the first order approximation to the finite difference form. The advantage is that we only need to evaluate the proximal mapping for m times. In this article, we use $\epsilon = 10^{-7}$ and $m = 20$ and find empirically good stability for the HMC algorithm.

For the HMC as a gradient-based algorithm, another potential concern is that $\text{prox}_{\lambda g}(\beta)$ may have zero gradient at certain value of β, for example, the soft-thresholding $\text{sign}(\beta)\max(|\beta| - \lambda, 0)$ will have zero gradient for those $\beta_j : |\beta_j| < \lambda$. Fortunately, two things prevent such a β_j from being stuck at a certain value. First, although the log-likelihood $\log L[y; \text{prox}_{\lambda g}(\beta)]$ may have a zero gradient for β_j, the log-prior $\log \Pi^0_{\boldsymbol{\beta}}$ does not (as it does not depend on $\text{prox}_{\lambda g}$)—in those cases, β_j will be updated through its prior distribution, until it enters the region where $L[y; \text{prox}_{\lambda g}(\beta)]$ is no longer invariant in β_j. This behavior is quite similar to the one with augmented "continuous particle" for sampling binary distribution via HMC (Pakman and Paninski 2013), where they demonstrated excellent mixing of Markov chains. Second, the HMC preserves the joint density of $\Pi(\beta \mid y)\Pi(v) = \Pi(\beta^* \mid y)\Pi(v^*)$ (with the MH correction), and as we sample a new v at the start of each iteration, the effective range of β^* to reach is $\{\beta^* : \Pi(\beta^* \mid y) = \Pi(\beta \mid y)\Pi(v)/\Pi(v^*), \Pi(v^*) > 0, \Pi(v^*) \le \Pi(\hat{v})\}$, with $\hat{v}$ the mode of $\Pi(v)$. Therefore, as long as $\Pi(v) < \Pi(\hat{v})$, we can have $\Pi(v^*) > \Pi(v)$, and $\Pi(\beta^* \mid y) < \Pi(\beta \mid y)$ allowing β^* to move away from a local-optimal state. In practice, we use the No-U-Turn algorithm (Hoffman and Gelman 2014), which ensures that we run the dynamics for long enough, so that the new proposal is away from the current state. We provide some diagnostic plots in S2 of the supplementary materials.

5. Simulation Studies

5.1. Set Expansion Prior for Hypothesis Testing

We now demonstrate the usefulness of the proximal prior in standard statistical inference, such as the hypothesis testing of whether θ is in a constrained set C. Consider two hypotheses $H_0 : \theta \in C$ and $H_1 : \theta \in \bar{C}$, where $C \bigcap \bar{C} = \varnothing$. For testing, one typically assumes a mixture prior

$$\Pi^0_{\boldsymbol{\theta}}(\theta) = p_0\phi(\theta)\mathbf{1}(\theta \in C) + (1 - p_0)\bar{\phi}(\theta)\mathbf{1}(\theta \in \bar{C}), \tag{14}$$

where ϕ and $\bar{\phi}$ are the prior kernel function of θ under H_0 and H_1, respectively; and p_0 is the prior probability assigned to C. The Bayes factor of H_0 relative to H_1 is defined as

$$\text{BF}_{01} = \frac{\int_C L(y;\theta)\phi(\theta)\text{d}\theta}{\int_{\bar{C}} L(y;\theta)\bar{\phi}(\theta)\text{d}\theta} = \frac{\text{pr}(\theta \in C \mid y)}{\text{pr}(\theta \in \bar{C} \mid y)}\frac{(1 - p_0)}{p_0},$$

for which, a smaller value of BF_{01} provides stronger evidence against H_0. Often, C is not of the same dimension with $\bar{C}$. For example, when testing a point null hypothesis $H_0 : \theta_1 = 0$, we have $\dim(C) < \dim(\bar{C})$. The standard practice has been assigning appropriate ϕ under C (and $\bar{\phi}$ under $\bar{C}$), with $\phi(\theta)$ and $\bar{\phi}(\theta)$ being 0 on $\bar{C}$ and C, respectively. However, when the null hypothesis is low-dimensional, such as testing linear equality, assigning density supported on the null set C can become quite challenging.

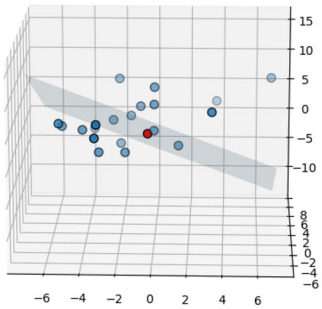
For a convex null set C, we could define a proximal prior based on the distance function, such that the prior density is positive on both C and $\bar{C}$. The distance function from point β to set C is defined as $\text{dist}_C(\beta) = \inf_{x\in C}\|x - \beta\|_2 = \|\beta - P_C(\beta)\|_2$. The proximal mapping of the distance function to set C is of the form

$$\text{prox}_{\lambda\text{dist}_C}(\beta) = \begin{cases} \beta + \frac{\lambda}{\text{dist}_C(\beta)}\{P_C(\beta) - \beta\}, & \text{if } \text{dist}_C(\beta) \ge \lambda \\ P_C(\beta), & \text{if } \text{dist}_C(\beta) < \lambda. \end{cases}$$

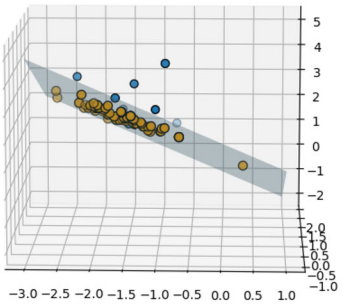
Clearly, this proximal mapping projects the points in the λ-neighborhood of C into C, and keeps the rest of the points out of C. Thus we get a prior that puts positive mass on both C and $\bar{C}$ and can also be expressed in the form of (14).

We can easily estimate the Bayes factor

$$\text{BF}_{01} = \frac{\text{pr}\{\text{dist}_C(\beta) < \lambda \mid y\}}{\text{pr}\{\text{dist}_C(\beta) \ge \lambda \mid y\}}\frac{\text{pr}\{\text{dist}_C(\beta) \ge \lambda\}}{\text{pr}\{\text{dist}_C(\beta) < \lambda\}}$$



(a) The data $y_1, \ldots, y_{20} \sim \mathrm{N}(\theta_0, 3^2 I_3)$ (the blue dots), with the true mean $\theta_0 = (-0.5, 0.3, 1.2)^{\mathrm{T}}$ (the red dot), shown along with the hyperplane $C = (\theta : \theta_1 + \theta_2 + \theta_3 = 1)$.



(b) The posterior samples from $\pi(\theta \mid y)$, which are mostly distributed on the hyperplane (the orange dots), while there are several outliers far from this hyperplane (the blue dots).

Figure 3. The set expansion prior for testing $\{\theta = (\theta_1, \theta_2, \theta_3) : \theta_1 + \theta_2 + \theta_3 = 1\}$.

via posterior sampling methods. If the prior ratio $\mathrm{pr}(\theta \in C)/\mathrm{pr}(\theta \in \bar{C})$ is not specified, in order to obtain adequate number of samples in both C and $\bar{C}$, we can choose a fixed λ (instead of assigning a prior on λ) such that $\mathrm{pr}\{\mathrm{dist}_C(\beta) \geq \lambda\} \approx \mathrm{pr}\{\mathrm{dist}_C(\beta) < \lambda\}$.

We conduct a simulated experiment: we have data $y_1, \ldots, y_{20}$ generated from $\mathrm{N}(\theta_0, 3^2)$ with $\theta_0 = (-0.5, 0.3, 1.2)^{\mathrm{T}}$. We want to test the linear equality hypothesis $H_0 : \theta_1 + \theta_2 + \theta_3 = 1$ against $H_1 : \theta_1 + \theta_2 + \theta_3 \neq 1$. The null set C is a hyperplane with Hausdorff dimension 2. We assign the set expansion prior to θ by assign $\mathrm{N}(0, 3^2)$ to β and set $\theta = \mathrm{prox}_{\lambda \mathrm{dist}_C}(\beta)$ with $\lambda = 2$, such that the prior ratio of C and $\bar{C}$ is around 0.48. Posterior sampling is implemented with the HMC with 5000 samples and 2000 burn-ins. We get an estimated Bayes factor $BF_{01} = 0.77$, and display 100 of the samples in Figure 3, panel (b).

5.2. Numerical Experiments on Variable Selection and Low Rank Matrix Model

In addition, we conduct numerical experiments for two models where solutions exist with conventional sparse priors: variable selection using a spike-and-slab prior, and low-rank matrix factorization with a discrete prior on the rank. We compare the computational performance in the combinatorial search-based MCMC algorithms for these models, with the HMC algorithm for our models using proximal priors. Further, we compare with other alternatives such as neuronized prior (Shin and Liu 2021) and multiplicative shrinkage prior (Bhattacharya and Dunson 2011; Legramanti, Durante, and Dunson 2020). We provide the details in S1.2 of the supplementary materials.

6. Data Application: Interpretable Factor Analysis of the Flow Network

We now demonstrate the practical usefulness of the proximal prior via analyzing the dynamic flow network data. The data (Zhu, Hu, and Collins 2020) include dynamic estimated traffic flow on major roads in Florida every 6 hr before Hurricane Irma made landfall until it covered the entire state (between 18:00 on September 06, 2017 and 18:00 on September 11, 2017). In total, the data contain 25 valid temporal records of flow networks, denoted by $Y^{(1)}, \ldots, Y^{(25)}$; each $Y^{(t)} \in \mathbb{R}^{n_V \times n_V}$ contains the traffic flows during a 6-hour period on the roads between $n_V = 382$ urban regions.

Each flow network is a weighted graph $\{V, E, Y^{(t)}\}$, with $V = (1, \ldots, n_V)$ the set of n_V nodes, $E = \{(i,j) : i,j \in V\}$ the edges, and the weight $Y^{(t)}_{i,j} \in \mathbb{R}$, representing the amount of flow between the two nodes, with $Y^{(t)}_{i,j} > 0$ a flow $i \to j$, and $Y^{(t)}_{i,j} < 0$ a flow $j \to i$. On the diagonal, $Y^{(t)}_{i,i} > 0$ indicates an external in-flow entering the network, while $Y^{(t)}_{i,i} < 0$ means an existing out-flow; $Y^{(t)}_{i,j} = 0$ if $(i,j) \notin E$.

To find useful patterns underneath the raw observation data, we use a low-dimensional latent factor model, with the factors $F^{(l)} \in \mathbb{R}^{n_V \times n_V}$ shared by all time points, while letting the loadings $\gamma_l^{(t)} \geq 0$ vary over time, subject to Gaussian measurement error $\mathcal{E}^{(t)}_{i,j} \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\mathcal{E}})$ for $i \leq j$.

$$Y^{(t)} = \sum_{l=1}^{d} \gamma_l^{(t)} F^{(l)} + \mathcal{E}^{(t)}. \tag{15}$$

Now, to make the factors useful in interpretation, we require each $F^{(l)}$ to be a *feasible flow*—an idealized flow network satisfying the following constraints, (i) skew-symmetry (except for the diagonal): $F^{(l)}_{i,j} = -F^{(l)}_{j,i}$ for $i < j$; (ii) flow-conservation, that the net sum of in-flows should equal to the out-flows for node j, $\sum_{i=1}^{n_V} F^{(l)}_{i,j} = 0$; (iii) to reduce noise, we assume that the elements of $F^{(l)}$ are sparse. Further, as we expect that most of the nodes do not have an external in-low/out-flow, we assume that (iv) most of the nodes having $F^{(l)}_{j,j} = \sum_{i \neq j} F^{(l)}_{i,j} = 0$. To

obtain the parameter in such a highly constrained space, we use the proximal mapping, with $\beta^{(l)} \in \mathbb{R}^{n_V \times n_V}$,

$$F^{(l)} = \text{prox}\{\beta^{(l)}\} = \underset{z \in \mathbb{R}^{n_V \times n_V}}{\arg\min}\ \lambda_1 \|z\|_1 + \lambda_1 \sum_{j=1}^{n_V} |z_{j,j}| + \frac{1}{2}\|\beta^{(l)} - z\|_2^2, \quad \text{subject to } z_{i,j} = -z_{j,i} \text{ for } i \neq j \text{ and } \sum_{i=1}^{n_V} z_{i,j} = 0 \text{ for } j = 1, \ldots, n_V. \tag{16}$$

The proximal mapping does not have a closed-form solution, however, can be efficiently computed using the alternating direction method of multipliers. Note that with constraint (i) and (ii), it is sufficient to use the lower-triangular entries to represent the rest. In the following, we use $(\beta_{i,j})_{i>j}$ to denote the $n_V(n_V-1)/2$ vector containing the lower-triangular entries. Therefore, (17) is equivalent to

$$\text{prox}\{(\beta_{i,j})_{i>j}\} = \underset{\{z_{i,j}\}_{i>j}}{\arg\min} \frac{1}{2}\sum_{i>j}(\beta_{i,j} - z_{i,j})^2 + \lambda_1 \sum_{i>j} |z_{i,j}| + \lambda_2 \sum_{j=1}^{n_V} |\sum_{i>j} z_{i,j} - \sum_{i<j} z_{j,i}|. \tag{17}$$

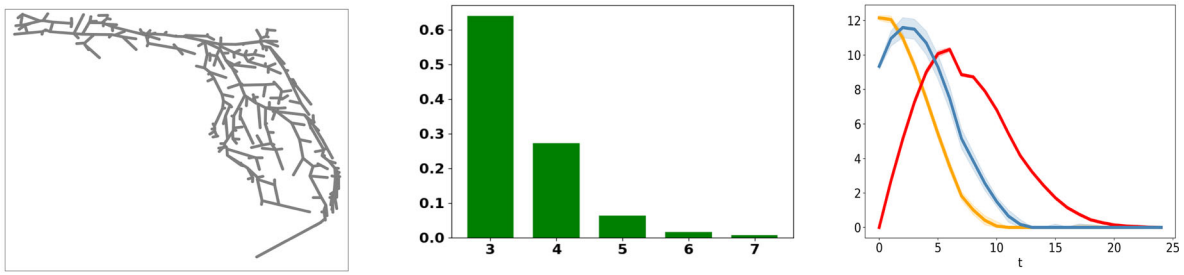
This proximal operator is evaluated via the alternating direction method of multipliers and solved iteratively. We provide the detailed algorithm in the Appendix.

On the prior of the loading, we assign a group shrinkage prior by using the (2, 1)-matrix norm in the proximal mapping. For the matrix $\gamma \in \mathbb{R}^{d \times T}$, we set:

$$\gamma = \text{prox}_{\lambda_2\|\cdot\|_{2,1}}(\rho) = \underset{z \in \mathbb{R}^{d \times T}}{\arg\min}\ \lambda_2 \|z\|_{2,1} + \frac{1}{2}\|\rho - z\|_F^2,$$

where $\|z\|_{2,1} = \sum_{l=1}^{d} \sqrt{\sum_{t=1}^{T} \{z_l^{(t)}\}^2}$. This prior has the advantage that $\{\gamma_l^{(1)}, \ldots, \gamma_l^{(T)}\}$ will be simultaneously zero for certain l—which allows us to use an overfitted model with a relatively large $d = 10$, with the posterior recovering only a small number of factors with nonzero loadings. We use independent standard normal as prior on the elements of $\beta^{(l)}$ and ρ.

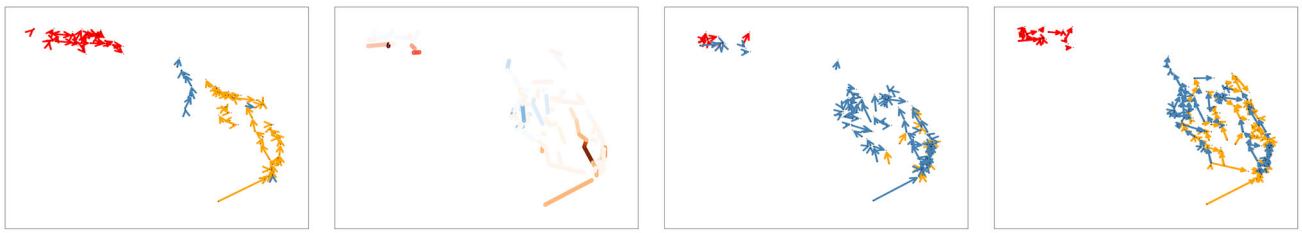
We run the HMC for 20,000 steps and discard the first 5000 as burn-ins, and we use thinning at every 10th iteration as the posterior sample. The posterior shows the highest probability at having three factors, and we visualize them in Figure 4(d)—clearly, by forcing the external in-flows and out-flows to be sparse, we have each factor roughly corresponding to a single connected sub-network.



(a) The traffic network showing the roads across the state of Florida.

(b) The posterior distribution of the number of non-zero factors.

(c) The estimated three loadings $\gamma_l^{(t)}$ changing over time.



(d) Estimated components by proximal prior.

(e) Edge-wise Variance of proximal prior.

(f) Estimated components by elementwise horseshoe.

(g) Estimated components by row-sparse horseshoe.

Figure 4. Data analysis on the dynamic flow network, observed during the Hurricane Irma evacuation. A latent factor model is fitted to the data, with the factors regularized by a proximal prior that force each network to be sparse in both the flows and the number of in-flows and out-flows. (d) shows three factor flows (posterior mode) estimated using the proximal prior. Each factor is close to a connected sub-flow network with (e) element-wise variance representing the uncertainty on the factor estimate. We compare the proximal prior with two realizations of shrinkage priors, where (f) shows the factors estimated using elementwise horseshoe prior, which are very fragmented and difficult to interpret. (g) shows the factors estimated using the row-sparse horseshoe, which contain many nodes with small or no flows.

Interestingly, examining the estimated loadings that change over the time points in Figure 4(c), we see that in the beginning of the evacuation, the factors 1 and 2 are dominant, but later there is a sudden decrease—this in fact corresponds to the time point when the hurricane makes the landfall, effectively forcing the traffic in those areas to shut down. After the 60th hr, the traffic moves up to the north part, and factor 3 represents the late stage of the evacuation.

To compare, we also test two continuous shrinkage priors on the factors. Specifically, we use (a) the elementwise horseshoe prior on each lower-diagonal $F^{(l)}_{i,j} \sim \mathrm{N}(0, \tau^2_{i,j}\sigma^2)$, $\tau_{i,j} \sim C^+(0,1)$, $\sigma^2 \sim$ Inverse-Gamma$(2, 0.01)$, and (b) the two-way group horseshoe prior, by letting $F^{(l)}_{i,j} \sim \mathrm{N}(0, \tau_i\tau_j\sigma^2)$, $\tau_i \sim C^+(0,1)$, $\sigma^2 \sim$ Inverse-Gamma$(2, 0.01)$. The purpose of (b) is to shrink each row of F simultaneously, while satisfying the skew-symmetry of $F^{(l)}$. Effectively these horseshoe priors accommodate the properties of (i)(ii)(iii) of a sparse feasible flow, nevertheless, they cannot accommodate (iv)—as each $F^{(l)}_{j,j}$ is completely determined given $F^{(l)}_{i,j}$ $(i \neq j)$, we could not further assign shrinkage prior on $F^{(l)}_{j,j} = \sum_{i\neq j} F^{(l)}_{j,j}$. As the result in Figure 4(f), the elementwise continuous shrinkage priors show a large number of external in-flows and out-flows, leading to fragmented small networks in each factor. As shown in Figure 4(g), the two-way group horseshoe finds many nodes with no flows at all, which is not very interpretable since we would like most of the nodes to have many in-flows and out-flows, as long as the total net-flow is zero.

7. Discussion

In this article, we exploit the proximal mapping to produce a new class of priors. As we have demonstrated, these priors and the associated probabilistic models can enable statistical inference (such as uncertainty quantification, hypothesis testing) on a wide range of problems, where in the past, one has been limited to point estimate only. The technique of "data augmentation using optimization" we have introduced could be generalized for other purposes, such as potentially new efficient algorithm for the posterior computation. Lastly, one could consider other type of optimization problems for a similar prior construction, such as the popular classes of semidefinite (Vandenberghe and Boyd 1996) and / or mixed integer programmings (Linderoth and Savelsbergh 1999), although how to provide a probabilistic treatment for these problems is still an open question.

Appendix

A.1. Proof of Theorem 1

Proof. For any value $\beta^* : \mathrm{prox}_{\lambda g_\gamma}(\beta^*) = \theta^*$ using g_γ at given γ and λ, $\epsilon < \|\mathrm{prox}_{\lambda g_\gamma}(\beta) - \mathrm{prox}_{\lambda g_\gamma}(\beta^*)\| \leq \|\beta - \beta^*\|$. Therefore, $\mathbf{1}(\theta, \|\theta - \theta^*\| > \epsilon \mid y, \lambda, \gamma) \leq \mathbf{1}\left(\beta, \left\|\beta - \beta^*\right\| > \epsilon \mid y, \lambda, \gamma\right)$. Taking the minimum over β^* on the right hand side and expectation on both sides, we obtain the first result.

Next, using the fact that for two independent copies β_1, β_2 from $\Pi(\beta \mid y)$, $2\mathrm{tr}[\mathrm{cov}(\beta \mid y) = \mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}(\|\beta_1 - \beta_2\|_2^2 \mid y) = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}(\|\beta_1 - \beta_2\|_2^2 \mid \lambda, \gamma, y)$, and the non-expansiveness of proximal mappings, we obtain the second result. □

A.2. Proof of Theorem 2

Proof. Let $0 < \lambda_1 < \lambda_2$, $v_1 = \mathrm{prox}_{\lambda_1 g}(x)$and $v_2 = \mathrm{prox}_{\lambda_2 g}(x)$, we prove: (i) $g(v_1) \geq g(v_2)$ and (ii) $\|v_1 - x\| \leq \|v_2 - x\|$. For (i),

$$\begin{aligned}
\frac{1}{2}\|v_2 - x\|^2 + \lambda_2 g(v_2) &= \frac{1}{2}\|v_2 - x\|^2 + \lambda_1 g(v_2) + (\lambda_2 - \lambda_1) g(v_2) \\
&\overset{(a)}{\geq} \frac{1}{2}\|v_1 - x\|^2 + \lambda_1 g(v_1) + (\lambda_2 - \lambda_1) g(v_2) \\
&= \frac{1}{2}\|v_1 - x\|^2 + \lambda_2 g(v_1) + (\lambda_2 - \lambda_1) \\
&\qquad \{g(v_2) - g(v_1)\} \\
&\overset{(b)}{\geq} \frac{1}{2}\|v_2 - x\|^2 + \lambda_2 g(v_2) + (\lambda_2 - \lambda_1) \\
&\qquad \{g(v_2) - g(v_1)\},
\end{aligned}$$

where (a) is due to v_1 is the minimizer of $\frac{1}{2}\|v - x\|^2 + \lambda_1 g(v)$, and similarly for v_2 in (b). Therefore, $(\lambda_2 - \lambda_1)\{g(v_2) - g(v_1)\} \leq 0$, which leads to $g(v_1) \geq g(v_2)$.

For (ii), slightly changing (a), we have $\frac{1}{2}\|v_1 - x\|^2 \leq \frac{1}{2}\|v_2 - x\|^2 - \lambda_1\{g(v_1) - g(v_2)\}$. Since $\lambda_1 \geq 0$, we have $\|v_1 - x\|^2 \leq \|v_2 - x\|^2$. □

A.3. Proof of Theorem 3

Proof. Since the proximal mapping satisfies 1-Lipschitz condition, $\mathrm{diam}\{\mathrm{prox}(B)\} \leq \mathrm{diam}(B)$ for all B in the domain of $\mathrm{prox}_{\lambda g}$. Using the definition of the Hausdorff measure, $\mathcal{H}^s\{\mathrm{prox}(A)\} \leq \mathcal{H}^s(A)$. To see the statement 2, we apply the result in statement 1 and see $\mathcal{H}^s\{\mathrm{prox}(\mathcal{A})\} = 0$ whenever $\mathcal{H}^s(\mathcal{A}) = 0$. □

A.4. Proof of Theorem 5

Proof. Using Theorem 3.2.22 of Federer (2014), for any $\mathcal{H}^p$ measurable function F on $\boldsymbol{\beta}^k$,

$$\int_{\boldsymbol{\beta}^k} F(\beta) J_k \mathrm{prox}_{\lambda g}(\beta) \mathcal{H}^p(\mathrm{d}\beta) = \int_{\Theta^k} \int_{\mathrm{prox}^{-1}_{\lambda g}(\theta)} F(b) \mathcal{H}^{p-k}(\mathrm{d}b) \mathcal{H}^k(\mathrm{d}\theta).$$

Using the assumption, we can exclude the zero-measure set where $J_{m_k}\mathrm{prox}_{\lambda g}(\beta) = 0$.

□

A.5. Algorithm to Compute the Proximal Mapping in the Flow Network Modeling

We formulate an equivalent problem to (17)

$$\begin{aligned}
\mathrm{prox}\{(\beta_{i,j})_{i>j}\} = \underset{\{z_{i,j}\}_{i>j}, x \in \mathbb{R}^{n_V}}{\arg\min} & \frac{1}{2}\sum_{i>j}(\beta_{i,j} - z_{i,j})^2 + \lambda_1 \sum_{i>j}|z_{i,j}| + \lambda_2 \sum_{j=1}^{n_V}|x_j| \\
& \text{subject to } C(z_{i,j})_{i>j} = x,
\end{aligned} \tag{18}$$

where $C \in \mathbb{R}^{n_V \times n_V(n_V-1)/2}$ is the matrix such that $\{C(z_{i,j})_{i>j}\}_k = \sum_{i>k} z_{i,k} - \sum_{i<k} z_{k,i}$. The scaled augmented Lagrangian for (18) is

$$\begin{aligned}
\mathcal{L} = & \frac{1}{2}\|(\beta_{i,j})_{i>j} - (z_{i,j})_{i>j}\|_2^2 + \lambda_1\|(\beta_{i,j})_{i>j}\|_1 + \lambda_2\|x\|_1 \\
& + \frac{\gamma}{2}\|C(z_{i,j})_{i>j} - x + u\|_2^2 - \frac{\gamma}{2}\|u\|_2^2.
\end{aligned}$$

In each iteration we update the x and z separately to minimize the Lagrangian:

$$(z_{i,j})_{i>j} \leftarrow \text{prox}_{\lambda_1/\gamma\|\cdot\|_1}[(I+\gamma C^{\mathrm{T}}C)^{-}\{\gamma C^{\mathrm{T}}(x-u)+\beta\}];$$
$$x \leftarrow \text{prox}_{\lambda_2/\gamma\|\cdot\|_1}\{C(z_{i,j})_{i>j}+u\},$$

and update u as in dual ascent:

$$u \leftarrow u + C(z_{i,j})_{i>j} - x$$

until convergence (i.e., $\|C(z_{i,j})_{i>j} - x\| \to 0$).

Supplementary Materials

Supplementary materials include (i) additional examples of proximal priors, along with additional experimental results and (ii) code to reproduce the results in this paper.

Funding

The authors would like to thank the support from the following funding. This research was partially funded by grants from the University of Florida Informatics Institute SEED Fund (UFII-SEED-2022: LD), the National Institute of General Medical Sciences (R35GM141798: HZ), the National Human Genome Research Institute (R01HG006139: HZ), and the National Science Foundation (DMS-2054253 and IIS-2205441: HZ).

ORCID

Hua Zhou https://orcid.org/0000-0003-1320-7118

References

Beck, A. (2017), *First-order Methods in Optimization*, Philadelphia, PA: SIAM. [2]

Berger, J. O., Bernardo, J. M., Sun, D., et al. (2009), "The Formal Definition of Reference Priors, *Annals of Statistics*, 37, 905–938. [5]

Bertsekas, D. P. (2014), *Constrained Optimization and Lagrange Multiplier Methods*, New York: Academic Press. [2]

Bhattacharya, A., and Dunson, D. B. (2011), "Sparse Bayesian Infinite Factor Models, *Biometrika*, 98, 291–306. [1,9]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), "Handling Sparsity via the Horseshoe," in *Artificial Intelligence and Statistics*, pp. 73–80. [1]

Cook, R. D., Li, B., and Chiaromonte, F. (2010), "Envelope Models for Parsimonious and Efficient Multivariate Linear Regression," *Statistica Sinica*, 20, 927–960. [4]

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve-Fitting With Free-Knot Splines, *Biometrika*, 88, 1055–1071. [1]

Dunson, D. B., and Neelon, B. (2003), "Bayesian Inference on Order-constrained Parameters in Generalized Linear Models," *Biometrics*, 59, 286–295. [2]

Edgar, G. (2007), *Measure, Topology, and Fractal Geometry*, New York: Springer. [6]

Federer, H. (2014), *Geometric Measure Theory*, Berlin: Springer. [7,11]

Goodall, C. R. and Mardia, K. V. (1999), "Projective Shape Analysis," *Journal of Computational and Graphical Statistics*, 8, 143–168. [1]

Green, P. J., and Hastie, D. I. (2009), "Reversible Jump MCMC," *Genetics*, 155, 1391–1403. [1]

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623. [8]

Ji, H., Liu, C., Shen, Z., and Xu, Y. (2010), "Robust Video Denoising using Low Rank Matrix Completion," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1791–1798, IEEE. [1]

Kastner, G. (2019), "Sparse Bayesian Time-Varying Covariance Estimation in Many Dimensions," *Journal of Econometrics*, 210, 98–115. [1]

Khare, K., Pal, S., and Su, Z. (2017), "A Bayesian Approach for Envelope Models," *The Annals of Statistics*, 45, 196–222. [4]

Legramanti, S., Durante, D., and Dunson, D. B. (2020), "Bayesian Cumulative Shrinkage for Infinite Factorizations," *Biometrika*, 107, 745–752. [1,9]

Li, Y., Craig, B. A., and Bhadra, A. (2019), "The Graphical Horseshoe Estimator for Inverse Covariance Matrices," *Journal of Computational and Graphical Statistics*, 28, 747–757. [1]

Lin, L., and Dunson, D. B. (2014), "Bayesian Monotone Regression Using Gaussian Process Projection," *Biometrika*, 101, 303–317. [2]

Linderoth, J. T., and Savelsbergh, M. W. (1999), "A Computational Study of Search Strategies for Mixed Integer Programming," *INFORMS Journal on Computing*, 11, 173–187. [11]

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, Series B, 70, 53–71. [1]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [1]

Morgan, F. (2016), *Geometric Measure Theory: A Beginner's Guide*, New York: Academic Press. [7]

Pakman, A., and Paninski, L. (2013), "Auxiliary-Variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions," in *Advances in Neural Information Processing Systems* (Vol. 26). [8]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1]

Petris, G., and Tardella, L. (2003), "A Geometric Approach to Transdimensional Markov chain Monte Carlo," *Canadian Journal of Statistics*, 31, 469–482. [2]

Preiss, D. (1987), "Geometry of Measures in Rn: Distribution, Rectifiability, and Densities," *Annals of Mathematics*, 125, 537–643. [6]

Rockova, V., and George, E. I. (2018), "The Spike-And-Slab Lasso," *Journal of the American Statistical Association*, 113, 431–444. [1]

Saarela, O., and Arjas, E. (2011), "A Method for Bayesian Monotonic Multiple Regression," *Scandinavian Journal of Statistics*, 38, 499–513. [1]

Sen, D., Patra, S., and Dunson, D. B. (2018), "Constrained Bayesian Inference through Posterior Projections," arXiv preprint arXiv:1812.05741. [2]

Shen, H., and Huang, J. Z. (2008), "Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation," *Journal of Multivariate Analysis*, 99, 1015–1034. [1]

Shin, M., Bhattacharya, A., and Johnson, V. E. (2020), "Functional Horseshoe Priors for Subspace Shrinkage," *Journal of the American Statistical Association*, 115, 1784–1797. [1]

Shin, M., and Liu, J. S. (2021), "Neuronized Priors for Bayesian Sparse Linear Regression," *Journal of the American Statistical Association*, 117, 1695–1710. [5,9]

Spall, J. C. (1992), "Multivariate Stochastic Approximation Using A Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control*, 37, 332–341. [8]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [3]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [2]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [1]

Vandenberghe, L., and Boyd, S. (1996), "Semidefinite Programming," *SIAM Review*, 38, 49–95. [11]

Zhao, P., Rocha, G., and Yu, B. (2006), "Grouped and Hierarchical Model Selection through Composite Absolute Penalties," Department of Statistics, UC Berkeley, Technical Report, 703. [1]

Zhu, Y.-J., Hu, Y., and Collins, J. M. (2020), "Estimating Road Network Accessibility During a Hurricane Evacuation: A Case Study of Hurricane Irma in Florida," *Transportation Research Part D: Transport and Environment*, 83, 102334. [9]