Prior distributions for structured semi-orthogonal matrices

Michael Jauch[1], Marie-Christine Düker[2], and Peter Hoff[3]

[1]Department of Statistics, Florida State University
[2]Department of Data Science, FAU Erlangen-Nürnberg
[3]Department of Statistical Science, Duke University

Abstract

Statistical models for multivariate data often include a semi-orthogonal matrix parameter. In many applications, there is reason to expect that the semi-orthogonal matrix parameter satisfies a structural assumption such as sparsity or smoothness. From a Bayesian perspective, these structural assumptions should be incorporated into an analysis through the prior distribution. In this work, we introduce a general approach to constructing prior distributions for structured semi-orthogonal matrices that leads to tractable posterior inference via parameter-expanded Markov chain Monte Carlo. We draw upon recent results from random matrix theory to establish a theoretical basis for the proposed approach. We then introduce specific prior distributions for incorporating sparsity or smoothness and illustrate their use through applications to biological and oceanographic data.

1 Introduction

Many statistical models for multivariate data include a parameter belonging to the set of semi-orthogonal matrices $\mathcal{V}(k, p) = \{Q \in \mathbb{R}^{p \times k} \mid Q^\top Q = I_k\}$ with $k < p$. The set $\mathcal{V}(k, p)$ is referred to as the Stiefel manifold [17]. In these models, the columns of the semi-orthogonal matrix parameter can often be understood as spanning a low-dimensional subspace in which much of the interesting variation in high-dimensional data occurs. The rows often represent the coordinates of statistical units embedded in a low-dimensional latent space. Examples appear throughout the literature on covariance estimation, principal component analysis (PCA), and factor models [10, 24, 23, 8, 47, 61]; low-rank matrix or tensor approximation [30, 32, 18, 26, 64]; network data [31, 42, 41]; spatial statistics [46]; time series analysis [43]; and other topics. These models have been applied in diverse fields including climate science [46], ecology [55, 42], economics [42], neuroscience [41, 43, 47], and seismic imaging [64].

Example 1 (Model-based singular value decomposition)**.** Suppose Y is an $n \times p$ data matrix. It is common to assume that $Y = M + E$ where M has rank $k \ll \min\{n, p\}$ and E is a matrix of errors [30, 18, 26, 64]. The low-rank matrix M can be represented in terms of its singular value

decomposition (SVD) $M = UDV^\top$ where $U \in \mathcal{V}(k,n)$, $V \in \mathcal{V}(k,p)$, and D is a $k \times k$ diagonal matrix with positive entries on the diagonal. Let y_i, u_i, and e_i be the ith rows of Y, U, and E. The model-based SVD can be viewed as a factor model in which the p measurements in y_i are related to k latent factors in u_i through the equation $y_i = VDu_i + e_i$. The model-based SVD is also closely related to PCA. In fact, if Y is centered and $e_{ij} \overset{\text{ind.}}{\sim} \text{Normal}(0, \sigma^2)$, the MLEs of UD and V coincide with the principal component scores and loadings from classical PCA.

Example 2 (Network eigenmodel). Now suppose $Y = (y_{ij})$ is a $p \times p$ adjacency matrix with $y_{ij} \in \{0, 1\}$ indicating the presence or absence of a relationship between object i and object j. The network eigenmodel of Hoff [31] supposes that

$$y_{ij} \mid \pi_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\,(\pi_{ij})$$
$$\pi_{ij} = \Phi\left[c + \left(Q\Lambda Q^\top\right)_{ij}\right]$$

where Φ is the cumulative distribution function of a standard normal random variable and (c, Q, Λ) are unknown parameters. The parameter Q is a $p \times k$ semi-orthogonal matrix, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ is a $k \times k$ diagonal matrix, and c is a real number.

In applications, there is often reason to assume that a semi-orthogonal matrix parameter satisfies a structural assumption such as sparsity or smoothness. If each y_i from Example 1 contains measurements of some variable over a uniform, one-dimensional grid of points in time or space, then it may be reasonable to assume that the columns of V will resemble discretizations of smooth curves. If the scientific context suggests that y_{ij} should only relate to a subset of the latent factors in u_i, then it may be reasonable to assume that V is sparse. Such an assumption typically leads to more interpretable results. In some high-dimensional settings, making structural assumptions regarding a semi-orthogonal matrix parameter allows us to make meaningful inferences that would otherwise be impossible.

In Bayesian statistics, structural assumptions are incorporated into an analysis through the prior distribution. The vast majority of articles on Bayesian methodology for models with semi-orthogonal matrix parameters assign the semi-orthogonal matrix parameter a uniform prior. There are a handful of exceptions that propose non-uniform priors for structured semi-orthogonal matrices. Yoshida and West [63] introduced a prior for a sparse semi-orthogonal factor loadings matrix and a computational method combining variational inference with simulated annealing to find the posterior mode. Gao and Zhou [24] proposed a prior for a sparse semi-orthogonal matrix and analyzed its theoretical properties in the context of sparse PCA. The authors described a method for computing the posterior mean, but the discrete parameter space rendered posterior simulation intractable. Pourzanjani et al. [48] combined the Givens angle representation of a semi-orthogonal matrix [52] and the horseshoe prior [13] to construct a prior for sparse PCA. Jauch et al. [35] leveraged the matrix angular central Gaussian (MACG) distribution [16] to define a prior for a semi-orthogonal matrix parameter with smooth structure and conducted posterior simulation via parameter-expanded Markov chain Monte Carlo (MCMC). North et al. [46] developed a prior for a semi-orthogonal matrix parameter with column-specific smooth structure and demonstrated its value in applications to spatial data. In principle, the matrix Bingham–von Mises–Fisher family of distributions [38, 31], which is well established in the directional statistics literature, could provide non-uniform priors for structured semi-orthogonal matrices. In practice, the normalizing constants

can be prohibitively difficult to evaluate, limiting their use in hierarchical models. The limited extent of the literature discussing priors for structured semi-orthogonal matrices is due, in part, to the difficulty of defining probability distributions on the Stiefel manifold which reflect such prior information and lead to tractable posterior inference.

This article introduces a family of prior distributions for structured semi-orthogonal matrices that leads to tractable posterior inference via the parameter-expanded MCMC methodology of Jauch et al. [35]. The family of priors, which includes the uniform distribution and a class of MACG distributions as particular cases, is defined by projecting an unconstrained real-valued random matrix onto the Stiefel manifold. We draw upon results from random matrix theory to establish that aspects of the distribution of the real-valued matrix are inherited by the distribution of the semi-orthogonal matrix. Thus, if we want the prior distribution for a semi-orthogonal matrix parameter to reflect structural assumptions, we can build those assumptions into the distribution of the real-valued matrix. We introduce specific prior distributions for incorporating sparsity or smoothness and illustrate their use through applications to biological and oceanographic data.

We now outline the remainder of the article. Section 2 reviews important preliminaries, including the polar decomposition, the MACG distribution, and the parameter-expanded MCMC methodology of Jauch et al. [35]. Section 3 introduces the proposed family of prior distributions and presents theoretical results demonstrating that aspects of the distribution of the real-valued matrix are inherited by the distribution of the semi-orthogonal matrix. Section 4 discusses the details of incorporating sparse or smooth structure with the proposed family of prior distributions. As part of this discussion, we describe a novel continuous shrinkage prior for the real random matrix introduced in the construction of the prior for the semi-orthogonal matrix. Section 5 considers applications to biological and oceanographic data. We conclude in Section 6 with a discussion.

Notation: For the reader's convenience, notations used throughout the paper are collected here. The maximum and minimum eigenvalues of a symmetric matrix A are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. Analogously, we write $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ for the largest and smallest singular values. The spectral norm of a matrix A is given by $\|A\| = \sqrt{\lambda_{\max}(A^\top A)}$ and the Frobenius norm as $\|A\|_F = \sqrt{\mathrm{Tr}(A^\top A)}$. The maximum absolute column sum or 1-norm $\|A\|_1 = \max_{j=1,\ldots,p} \sum_{i=1}^{p} |A_{ij}|$. We say that a random variable X is bounded almost surely when there is a $M \geq 0$ such that $\mathrm{P}(|X| > M) = 0$, i.e., the set $\{\omega \mid |X(\omega)| > M\}$ has probability zero. For a sequence of random variables $(\xi_n)_{n \geq 1}$ and a deterministic sequence $(c_n)_{n \geq 1}$, we write $\xi_n = O_{a.s.}(c_n)$ to indicate that ξ_n/c_n is almost surely bounded for large enough n. We write $X \sim \mu$ to indicate that the random variable X was drawn from a distribution μ. Finally, $\overset{d}{=}$ denotes equality in distribution.

2 Preliminaries: the MACG distribution and polar expansion

The proposed prior distribution extends the MACG distribution of Chikuse [16] and leads to tractable posterior simulation via polar expansion, the parameter-expanded MCMC method of Jauch et al. [35]. Both the MACG distribution and polar expansion are built upon the polar decomposition. In this section, we briefly review each of these topics. The discussion draws heavily upon that of Jauch et al. [35].

The polar decomposition. The polar decomposition is the unique representation of a full rank matrix $X \in \mathbb{R}^{p \times k}$ as the product $X = Q_X S_X^{1/2}$ where $Q_X \in \mathcal{V}(k, p)$, S_X is a $k \times k$ symmetric positive definite (SPD) matrix, and $S_X^{1/2}$ is the symmetric square root of S_X. The components of the polar decomposition can be computed from X as $Q_X = X(X^\top X)^{-1/2}$ and $S_X = X^\top X$. In terms of the singular value decomposition $X = UDV^\top$, we have $Q_X = UV^\top$ and $S_X^{1/2} = VDV^\top$. The semi-orthogonal component Q_X can be understood as the projection of X onto $\mathcal{V}(k, p)$. More precisely, Q_X is the closest matrix in $\mathcal{V}(k, p)$ to X in the Frobenius norm, i.e. $Q_X = \mathrm{argmin}_{Q \in \mathcal{V}(k,p)} \|X - Q\|_F$ [29].

The MACG distribution. The random semi-orthogonal matrix $Q \in \mathcal{V}(k, p)$ is said to have a matrix angular central Gaussian MACG(Σ) distribution if $Q \overset{d}{=} Q_X$ where $X \sim N_{p,k}(0, \Sigma, I)$ [17]. The notation $N_{p,k}(0, \Sigma, I)$ indicates a centered matrix normal distribution with Σ as its $p \times p$ row covariance matrix and the identity as its $k \times k$ column covariance matrix [53, 19]. We can think of the MACG distribution as resulting from the projection of a centered normal random matrix onto the Stiefel manifold. The MACG(Σ) distribution has density $f(Q \mid \Sigma) = |\Sigma|^{-k/2} |Q^\top \Sigma^{-1} Q|^{-p/2}$ and is uniform on the Stiefel manifold when $\Sigma = I$ or $k = p$.

Posterior inference via polar expansion. Among other problems, Jauch et al. [35] considered MCMC simulation from the Stiefel manifold when the target distribution is a posterior distribution arising from a MACG(Σ) prior. Suppose we have data y whose distribution given the unknown parameter $Q \in \mathcal{V}(k, p)$ has density $p(y \mid Q)$. The prior density is $p(Q) = |\Sigma|^{-k/2} |Q^\top \Sigma^{-1} Q|^{-p/2}$. The posterior density satisfies $p(Q \mid y) \propto p(y \mid Q) \, p(Q)$.

To avoid the complications of direct simulation from the posterior density $p(Q \mid y)$ defined on the Stiefel manifold, we can reparametrize in terms of the matrix X from the definition of the MACG distribution. The prior distribution for X is the centered matrix normal distribution $N_{p,k}(0, \Sigma, I)$. The posterior density satisfies

$$p(X \mid y) \propto p(y \mid Q_X) \, N_{p,k}(X \mid 0, \Sigma, I). \tag{1}$$

To obtain an MCMC approximation to the posterior distribution of Q, we can simulate a Markov chain $X_1, \ldots, X_T$ on $\mathbb{R}^{p \times k}$ whose stationary distribution has density (1), calculate $Q_t = X_t(X_t^\top X_t)^{-1/2}$ for each iteration t, and take the empirical distribution of $Q_1, \ldots, Q_T$ as our approximation. The posterior distribution of X is non-standard, but knowing its density (up to a constant factor) allows us to apply standard MCMC methods, e.g. adaptive Hamiltonian Monte Carlo (HMC) [45, 33] as implemented in Stan [12] .

The same approach to posterior inference is available for other priors defined by projecting an unconstrained real-valued random matrix X onto the Stiefel manifold, including the family of prior distributions introduced in this article. We simply replace the centered matrix normal density in (1) with an alternative prior density for X. This computational tractability is an important advantage of the proposed family of prior distributions.

3 The family of prior distributions and its properties

We now introduce the proposed family of prior distributions. Let $Z = (Z_{ij})$ be a $p \times k$ matrix with i.i.d. real entries having mean zero and unit variance; let Ω be a $p \times p$ correlation matrix; and set $X = \Omega^{1/2} Z$. The proposed prior is the distribution of Q_X, the projection of X onto the Stiefel manifold $\mathcal{V}(k, p)$ obtained via the polar decomposition. There are two notable special cases. If the entries of Z are independent standard normal random variables, then $Q_X \sim \mathrm{MACG}(\Omega)$. If, additionally, we have that $\Omega = I_p$, then Q_X is uniform on $\mathcal{V}(k, p)$.

In the remainder of this section, we present results demonstrating that properties of the distribution of X are inherited by the distribution of Q_X. Theorem 1 establishes that invariance properties of X are shared by Q_X. Theorem 2 implies that the Wasserstein distance between the distribution of m entries of $\sqrt{p}Q_X$ and the distribution of the corresponding entries of X converges to zero as $m, p, k \to \infty$ provided that $m\sqrt{k/p} \to 0$. These results have important implications for statistical modeling: If we want the prior distribution for a semi-orthogonal matrix parameter to reflect structural assumptions, we can build these features into the distribution of X. Theorem 2 is particularly relevant to prior specification in the common scenario in which a semi-orthogonal matrix parameter has far more rows than columns. Beyond its relevance to statistical modeling, Theorem 2 may be of independent interest to those working in high-dimensional probability and random matrix theory.

3.1 Invariance properties

Let $\mathcal{L} \subseteq \mathcal{O}(p)$ and $\mathcal{R} \subseteq \mathcal{O}(k)$ where $\mathcal{O}(p)$ and $\mathcal{O}(k)$ are the orthogonal groups in dimensions p and k. A $p \times k$ random matrix Y is invariant to left multiplication by elements of $\mathcal{L}$ if $LY \overset{d}{=} Y$ for all $L \in \mathcal{L}$. The random matrix Y is invariant to right multiplication by elements of $\mathcal{R}$ if $YR \overset{d}{=} Y$ for all $R \in \mathcal{R}$. Finally, the random matrix Y is invariant to left multiplication by elements of $\mathcal{L}$ and right multiplication by elements of $\mathcal{R}$ if $LYR \overset{d}{=} Y$ for all $L \in \mathcal{L}$ and $R \in \mathcal{R}$.

Theorem 1 (Invariance). *If the random matrix X is invariant to left multiplication by elements of $\mathcal{L}$ and right multiplication by elements of $\mathcal{R}$, then so is Q_X.*

Theorem 1 characterizes how invariance properties of the distribution of X influence those of the distribution of Q_X and thus has important implications for choosing an appropriate prior from the proposed family. The uniform distribution on $\mathcal{V}(k, p)$ is the unique distribution which is invariant to left multiplication by elements of $\mathcal{O}(p)$ and right multiplication by elements of $\mathcal{O}(k)$. Theorem 1 implies that our prior distribution is uniform when X is invariant to left multiplication by elements of $\mathcal{O}(p)$ and right multiplication by elements of $\mathcal{O}(k)$. When we expect a semi-orthogonal matrix parameter to have structure such as sparsity or smoothness, a uniform prior is inappropriate, but less stringent invariance properties may still be desirable. For example, if row indices provide no substantive information, our prior distribution should be invariant to left multiplication by permutation matrices. If, additionally, there is no prior information regarding the signs of its entries, our prior distribution should be invariant to left multiplication by signed permutation matrices. Theorem 1 tells us how to achieve this.

3.2 Limit theorem

The next result sheds additional light on the similarity between the distribution of a random matrix X and that of its projection Q_X onto the Stiefel manifold $\mathcal{V}(k, p)$. Because the result is asymptotic in nature, we emphasize the dependence on p by writing X_p and Q_{X_p} throughout this section.

We quantify the similarity between the two distributions in terms of the Wasserstein distance.

Definition 3.1. Let $(\mathcal{X}, d)$ be a Polish metric space. For $\ell \in [1, \infty)$, the Wasserstein distance of order ℓ between two probability measures μ, ν on $\mathcal{X}$ with finite ℓth moments, is defined as

$$W_\ell(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} (d(x, y))^\ell \, d\gamma(x, y) \right)^{\frac{1}{\ell}}, \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions (or couplings) γ with marginals μ and ν.

The Wasserstein distance metrizes weak convergence [56, Theorem 6.9]. In our results, we consider the Wasserstein distance of order $\ell = 2$ with respect to either the Euclidean norm or Frobenius norm (depending on whether we are comparing the distributions of random vectors or random matrices).

Recall that $X_p = \Omega_p^{1/2} Z_p$. In order for the result to hold, we need to impose the following assumptions on the matrices Z_p and Ω_p :

Assumption 1. *The entries of $Z_p \in \mathbb{R}^{p \times k}$ are i.i.d., have mean zero, unit variance, and finite fourth moments.*

Assumption 2. *The matrix $\Omega_p \in \mathbb{R}^{p \times p}$ is a correlation matrix for which*

- *the quantity $c_\Omega(p) := p^{-1} \operatorname{Tr}\left(\Omega_p^2\right)$ converges to a constant $c_\Omega > 0$; and*

- *the spectral norm $\|\Omega_p\|$ is bounded above.*

Moment conditions and the conditions appearing in Assumption 2 are common in the random matrix theory literature. Proposition 2 in Section 4 verifies that Assumption 1 holds for the novel shrinkage prior proposed in Section 4. Proposition 3 in Section 4 verifies that Assumption 2 holds for a variety of different sequences of correlation matrices.

We can now state Theorem 2, which characterizes the asymptotic behavior of the Wasserstein distance between the distribution of m entries of $\sqrt{p} Q_X$ and the distribution of the corresponding entries of X.

Theorem 2 (Wasserstein distance)**.** *Suppose that $X_p = \Omega_p^{1/2} Z_p$ with Z_p as in Assumption 1 and Ω_p as in Assumption 2. The Wasserstein distance between m entries of $\sqrt{p} Q_{X_p}$, denoted by $Q_{X_p}(m)$, and the corresponding entries of X_p, denoted by $X_p(m)$, satisfies*

$$W_2 \left[law \left(\sqrt{p} \, Q_{X_p}(m) \right), law \left(X_p(m) \right) \right] = \mathcal{O}\left(c_\Omega(p) m \sqrt{\frac{k}{p}} \right). \tag{3}$$

An immediate consequence of Theorem 2 is that the Wasserstein distance converges to zero as $m, p, k \to \infty$ provided that $m\sqrt{k/p} \to 0$.

There is significant interest in normal approximations to the entries of random orthogonal or semi-orthogonal matrices in the random matrix theory community. For a detailed review, see Section 6 of Jauch et al. [34] and the references therein. We highlight an especially close connection between Theorem 2 and this literature. A result of Stam [54] showed that the distribution of m entries of a uniformly distributed unit vector converges to the distribution of m independent standard normal random variables as its length p grows. Watson [59] extended this result to a uniformly distributed semi-orthogonal matrix with a growing number of rows but a fixed number of columns. Theorem 2 offers a new perspective on this topic. It shows that the results of Stam [54] and Watson [59] describe special cases of a more general phenomenon. That is, the distribution of m entries of the projection Q_{X_p} of a real random matrix X_p onto the Stiefel manifold $\mathcal{V}(k, p)$ can be approximated by the distribution of the corresponding entries of X_p provided that m and k do not grow too fast compared to p and X_p satisfies the assumptions stated above.

Next, we highlight some connections and differences between the squared Wasserstein distance between the distributions of $\sqrt{p}Q_{X_p}$ and X_p and the so-called linear spectral statistic (LSS) of the re-normalized sample covariance matrix. Recall that $X_p^\top X_p = Z_p^\top \Omega_p Z_p$. The re-normalized sample covariance matrix is defined as

$$A_k := \frac{1}{\sqrt{kp\boldsymbol{c}_\Omega(p)}} \left(Z_p^\top \Omega_p Z_p - pI_k \right) \tag{4}$$

with $\boldsymbol{c}_\Omega(p) := p^{-1} \operatorname{Tr}\left(\Omega_p^2\right)$. For a smooth function $f(\cdot)$ of interest, the corresponding LSS is

$$\frac{1}{k} \sum_{j=1}^{k} f(\lambda_j(A_k)). \tag{5}$$

The empirical eigenvalue distribution of A_k is $F^{A_k}(x) = \frac{1}{k}\sum_{j=1}^{k} 1_{\{\lambda_j(A_k) \leq x\}}$.

The re-normalized sample covariance matrix (4) has been considered extensively in the random matrix theory literature. We briefly review this literature for the ultra-high dimensional setting where $p, k \to \infty$ with $k/p \to 0$. Bai and Yin [4] were the first to study the spectral distribution of A_k in this setting in the case where $\Omega_p = I_p$ in (4). They proved that the eigenvalue distribution of A_k converges to the semi-circle law. Chen and Pan [14] studied the behavior of the largest eigenvalue of A_k with $\Omega_p = I_p$. For a general correlation matrix Ω_p, Wang and Paul [58] derived conditions under which the empirical eigenvalue distribution converges to the semicircle law. Chen and Pan [15] established a central limit theorem for LSS of A_k. More recently, Qiu et al. [49] studied the LSS of A_k in the ultra-high dimensional regime.

The next lemma reveals the relationship between the squared Wasserstein distance between the distributions of $\sqrt{p}Q_{X_p}$ and X_p and the eigenvalues of the sample covariance matrix of X_p.

Lemma 1. *The squared Wasserstein distance between X_p and $\sqrt{p}Q_{X_p}$ can be written as*

$$W_2\left[law\left(\sqrt{p}\,Q_{X_p}\right), law\left(X_p\right)\right]^2 = 2p\operatorname{E}\left[\sum_{j=1}^{k}(1 - \lambda_j^{1/2}(S_p))\right] \quad with \quad S_p = p^{-1}Z_p^\top \Omega_p Z_p. \tag{6}$$

7

Based on Lemma 1, we can relate the Wasserstein distance to the LSS of A_k. There are a couple of observations to make. First, $(1 - \lambda_j^{1/2}(S_p)) \neq f(\lambda_j(A_k))$, i.e., we can not directly apply results on the spectral statistic (5). However, we do have that $(1 - \lambda_j^{1/2}(S_p)) = (1 - \lambda_j^{1/2}(\sqrt{\frac{k\boldsymbol{c}_\Omega(p)}{p}} A_k + \sqrt{\frac{p}{k\boldsymbol{c}_\Omega(p)}} I_k))$. Second, Lemma 1 considers the Wasserstein distance between the distributions of the full matrices Q_{X_p} and X_p rather than a subset of entries as in Theorem 2. Considering the full matrices allows us to write the Wasserstein distance as the expectation of the squared Frobenius norm, which can be calculated explicitly.

4   Incorporating Structure

We now discuss specific prior distributions from the proposed family that are appropriate for semi-orthogonal matrix parameters with sparse or smooth structure. Section 4.1 introduces a novel continuous shrinkage prior for the entries of Z that leads to an approximately sparse Q_X when $k \ll p$. Section 4.2 considers different correlation matrices Ω for incorporating row dependence, including some that lead to a Q_X with smooth structure, and verifies that Assumption 2 holds for sequences of these correlation matrices.

4.1   Sparsity

To construct a prior distribution for a sparse semi-orthogonal matrix, we will set $\Omega = I_p$ and let the entries of Z be i.i.d. from a continuous shrinkage prior with mean zero, unit variance, and finite fourth moments. Then, by Theorem 2, Q_X will inherit the approximate sparsity of X in the high-dimensional setting where $k \ll p$. Assumption 1 rules out many choices for the continuous shrinkage prior. For example, global-local shrinkage priors such as the horseshoe [13], generalized double Pareto [2], or Dirichlet-Laplace [6] introduce dependence through the global scale parameter and, in many cases, have infinite variance and fourth moments. These observations motivate us to propose a novel continuous shrinkage prior for the entries of Z that satisfies the requirements above and has a number of other appealing properties. In particular, we propose to let the entries of Z be i.i.d. with density

$$f(z \mid \ell) = \frac{(\ell/2)^{\ell/2}}{\Gamma(\ell/2)} |z|^{\ell-1} \exp\left(-\ell z^2/2\right), \quad \ell \in (0, 1]. \tag{7}$$

Proposition 1 provides a representation of this distribution as a scale mixture of normal distributions, which is helpful for establishing its theoretical properties and leads to improved computational efficiency when we consider posterior inference in Section 5.1.

Proposition 1. *For $\ell \in (0, 1)$, let each entry of Z be generated as $Z_{ij} \mid \theta_{ij} \overset{ind.}{\sim} \text{Normal}\left(0, \theta_{ij}/\ell\right)$ where $\theta_{ij} \overset{i.i.d.}{\sim} \text{Beta}\left[\ell/2, (1-\ell)/2\right]$. The entries of Z are then i.i.d. with density* (7).

Proposition 2 presents some other important properties of the distribution.

Proposition 2. *The following properties hold:*

I. *When $\ell = 1$, $f(z \mid \ell)$ is the standard normal density. As $\ell \to 0$, the kernel $z \mapsto |z|^{\ell-1} \exp(-\ell z^2/2)$ of the density* (7) *converges pointwise to the function $z \mapsto |z|^{-1}$.*

II. *For $\ell \in (0, 1)$, $\lim_{z \to 0} f(z \mid \ell) = \infty$.*

III. *The entries of the random matrix Z are i.i.d. with mean zero, unit variance, and finite fourth moment.*

IV. *The squared entries of Z are distributed $Z_{ij}^2 \overset{i.i.d.}{\sim}$ Gamma(shape $= \ell/2$, scale $= 2/\ell$).*

Property I tells us that the proposed distribution includes the standard normal distribution (when $\ell = 1$) and the improper normal-Jeffreys prior (as $\ell \to 0$) as limiting cases. As discussed earlier, when the entries of Z are i.i.d. standard normal random variables, Q_X is uniform on $\mathcal{V}(k, p)$. The normal-Jeffreys prior was studied in Figueiredo [22] and Bae and Mallick [3] as a means of obtaining sparse estimates in regression problems. Property II indicates that, like the univariate horseshoe density, the density (7) is unbounded at zero. Property III confirms that the proposed continuous shrinkage prior satisfies the requirements we stated at the beginning of this subsection. Proposition IV establishes that the squared entries of Z are independent gamma random variables with a common rate parameter, which allows us to derive the marginal distribution of the entries of Q_X when $k = 1$ in Example 3.

Example 3. Suppose that Z and Q_X each have a single column so that $Z = (z_1, \ldots, z_p)^\top$ and $Q_X = (q_1, \ldots, q_p)^\top$. In that case, the squared entries of Q_{X_p} have joint distribution $(q_1^2, \ldots, q_p^2)^\top \sim$ Dirichlet$(\ell/2, \ldots, \ell/2)$ and marginal distribution $q_i^2 \sim$ Beta(α, β) where $\alpha = \ell/2$ and $\beta = (p-1)\ell/2$. The entries of Q_{X_p} themselves have marginal density

$$q \mapsto \frac{1}{B(\alpha, \beta)} |q|^{2\alpha-1}(1 - q^2)^{\beta-1}, \quad -1 < q < 1.$$

Figure 2 compares the densities of z_i with those of $\sqrt{p}q_i$ for $\ell = .1$ when $p = 5$ and $p = 100$. In the high-dimensional case when $p = 100$, these densities are indistinguishable, as we expect based on Theorem 2.
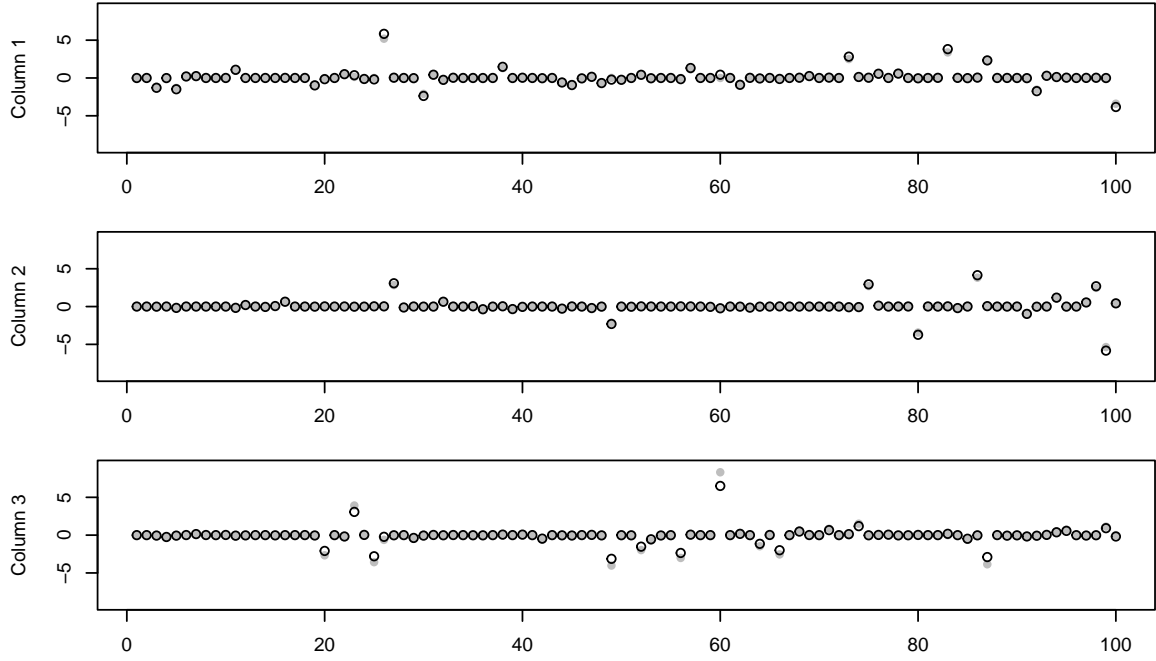
Figure 1: A comparison of the columns of a single realization X^* of X with those of $\sqrt{p}\, Q_{X^*}$ for the sparsity-inducing prior discussed in Section 4.1 with $\ell = .1$ and dimensions $p = 100, k = 3$. The entries of X^* appear as gray dots while the entries of $\sqrt{p}\, Q_{X^*}$ appear as black circles.
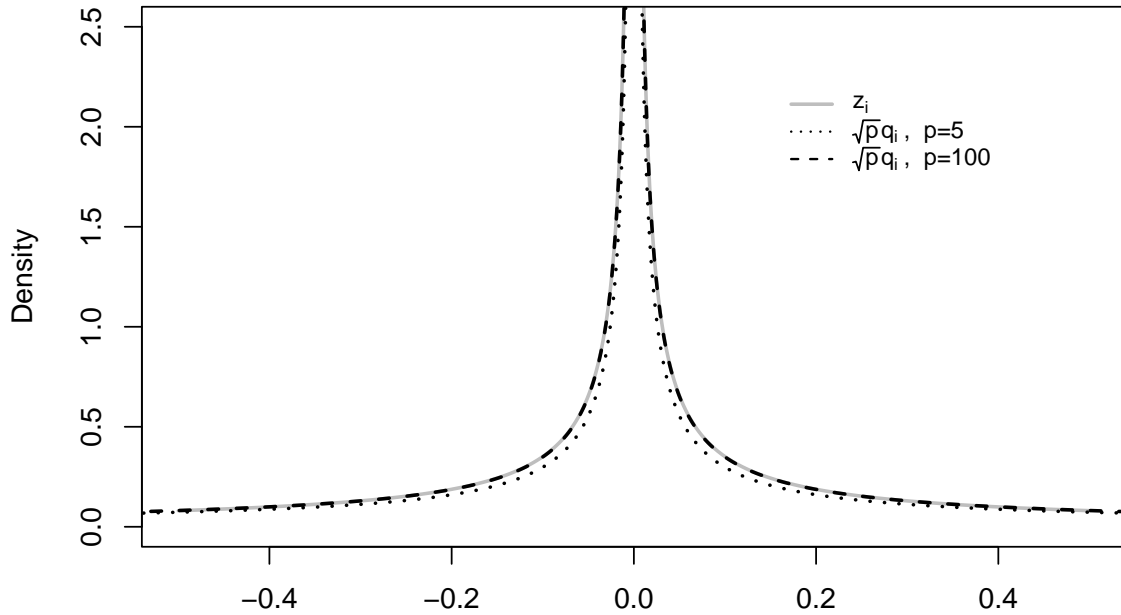

Figure 2: A comparison of the marginal densities of z_i and $\sqrt{p}q_i$ for $\ell = .1$ when $p = 5$ and $p = 100$, as discussed in Example 3.

4.2 Smoothness

To construct a prior distribution for a semi-orthogonal matrix with smooth structure, we introduce row dependence through the correlation matrix Ω. Example 4 considers the case where the entries of Z are i.i.d. standard normals and Ω is defined with the Matérn correlation function. Proposition 3 verifies that Assumption 2 of Theorem 2 holds for a variety of different sequences of correlation matrices.

Example 4. Let the entries of Z be i.i.d. standard normal random variables and let $\Omega = (\Omega_{ij})$ satisfy $\Omega_{ij} = C_{\rho,\nu}(|i-j|)$ where

$$C_{\rho,\nu}^{\text{Matérn}}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{d}{\rho}\right) \tag{8}$$

is the Matérn correlation function [50] with parameters $\rho, \nu > 0$. In this case, $Q_X \sim \text{MACG}(\Omega)$. Figure 3 compares the columns of a single realization X^* of X to those of $\sqrt{p}\,Q_{X^*}$ for parameters $\nu = 3, \rho = 12$ and dimensions $p = 100, k = 3$. The columns of the two matrices share a similar smooth structure, as we should expect based on Theorem 2. Proposition 3 confirms that Assumption 2 of Theorem 2 holds for a sequence of correlation matrices constructed from the Matérn correlation function (8).
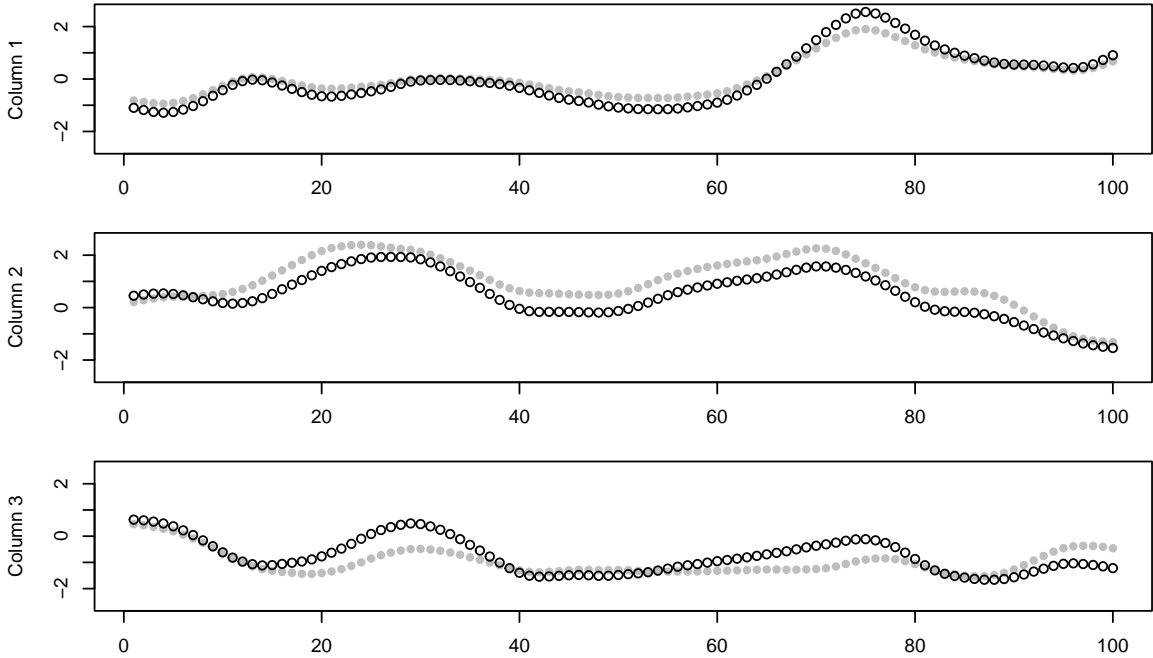

Figure 3: A comparison of the columns of a single realization X^* of X with those of $\sqrt{p}Q_{X^*}$ when the entries Z are i.i.d. standard normals and Ω is constructed from the Matérn correlation function, as discussed in Example 4. The entries of X^* appear as gray dots while the entries of $\sqrt{p}\,Q_{X^*}$ appear as black circles.

Proposition 3. *Assumption 2 of Theorem 2 is satisfied for* $\Omega_p = (\Omega_{ij})$ *constructed in any of the following ways:*

11

I. *(Power correlation)* $\Omega_{ij} = C_\rho^{Power}(|i-j|)$ *where* $C_\rho^{Power}(d) = \rho^d$ *for* $0 < \rho < 1$;

II. *(Squared exponential correlation)* $\Omega_{ij} = C_\rho^{SE}(|i-j|)$ *where* $C_\rho^{SE}(d) = \exp\left[-d^2/(2\rho^2)\right]$ *for* $\rho > 0$;

III. *(Matérn correlation)* $\Omega_{ij} = C_{\rho,\nu}^{Matérn}(|i-j|)$ *where* $C_{\rho,\nu}$ *is defined as in* (8) *for* $\rho, \nu > 0$;

IV. *(Banded correlation) Let* C *be any of the correlation functions above and set*

$$\Omega_{ij} = \begin{cases} C(|i-j|) & |i-j| \leq \tau \\ 0 & |i-j| > \tau \end{cases}$$

for a bandwidth $0 \leq \tau < p$.

5   Data Examples

5.1   Sparse network eigenmodel for protein interaction data

To demonstrate the proposed family of prior distributions in practice, we revisit a data example from Hoff [31] in which the author proposes a network eigenmodel for the protein interaction data of Butland et al. [9]. This data example has also appeared in Jauch et al. [35], Loyal and Chen [42], and other articles. In contrast to these previous works, all of which considered a uniform prior distribution for the eigenvectors of the network eigenmodel, we assign the eigenvectors the sparsity-inducing prior introduced in Section 4.1. Our experiments indicate that we can learn an appropriate level of sparsity from the data, leading to more interpretable results without sacrificing predictive performance.

The interaction data of $p = 270$ proteins of *Escherichia coli* are recorded in a binary, symmetric $p \times p$ matrix $Y = (y_{ij})$. We have $y_{ij} = 1$ if protein i and protein j interact while $y_{i,j} = 0$ otherwise. The network eigenmodel assumes that

$$y_{ij} \mid \pi_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\pi_{ij}\right)$$
$$\pi_{ij} = \Phi\left[c + \left(Q\Lambda Q^\top\right)_{ij}\right]$$

where Φ is the cumulative distribution function of a standard normal random variable and (c, Q, Λ) are unknown parameters. The parameter Q is a $p \times k$ semi-orthogonal matrix, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ is a $k \times k$ diagonal matrix, and c is a real number. As in Hoff [31] and Jauch et al. [35], the diagonal elements of Λ have independent $N(0, p)$ prior distributions while $c \sim N(0, 10^2)$. See Section 5.1 of Jauch et al. [35] for a discussion of the identifiability of the network eigenmodel.

The eigenvector matrix Q is assigned the sparsity-inducing prior introduced in Section 4.1. More precisely, we let $Z = (Z_{ij})$ be a $p \times k$ matrix with i.i.d. real entries distributed according to the density (7). The prior for Q is the distribution of $Q_Z = Z\left(Z^\top Z\right)^{-1/2}$, the projection of Z onto the Stiefel manifold obtained via the polar decomposition. The density (7) includes a parameter $\ell \in (0, 1)$ that determines the sparsity of the eigenvector matrix Q. Instead of choosing a fixed value, we assign $\ell \in (0, 1)$ a uniform prior and learn it from the data. In practice, we use the
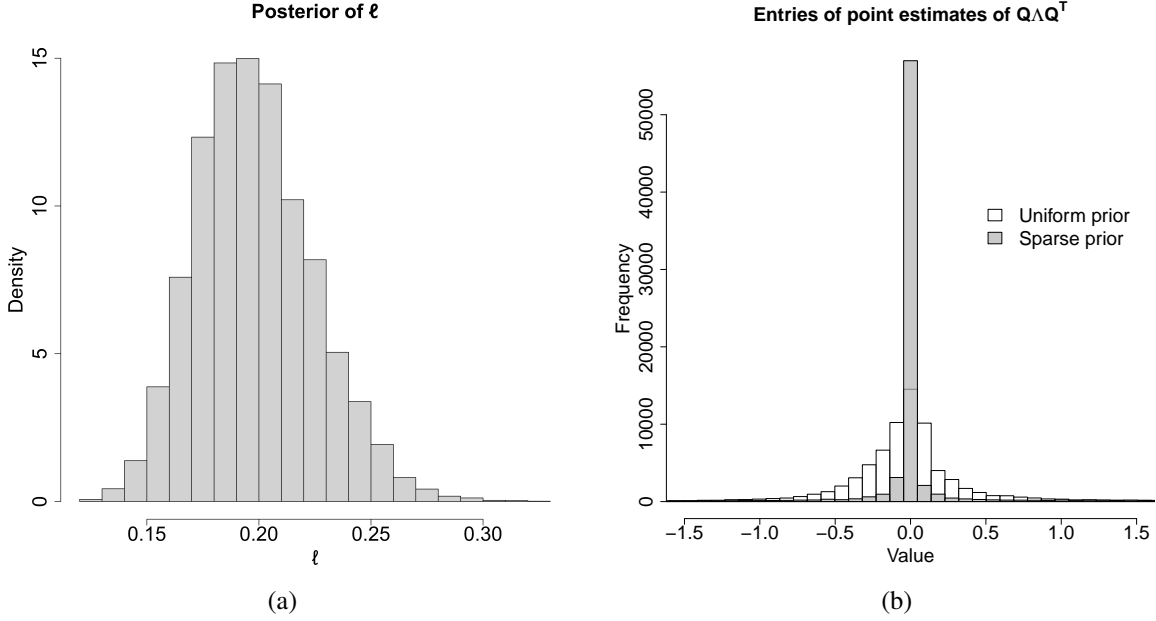
Figure 4: Panel (a) shows the marginal posterior distribution of the sparsity parameter ℓ. Panel (b) compares the entries of the point estimate of $Q\Lambda Q^T$ under the two priors for Q.

scale mixture representation given in Proposition 1 so that $Z_{ij} \mid \theta_{ij}, \ell \overset{\text{ind.}}{\sim} \text{Normal}\left(0, \theta_{ij}/\ell\right)$ where $\theta_{ij} \overset{\text{i.i.d.}}{\sim} \text{Beta}\left[\ell/2, (1-\ell)/2\right]$. Let Θ denote the $p \times k$ matrix whose ijth entry is θ_{ij}.

We obtain an MCMC approximation to the posterior distribution by applying polar expansion and adaptive HMC as implemented in Stan. After polar expansion, the posterior density is

$$
p\left(c, Z, \Theta, \Lambda, \ell \mid Y\right) \propto \prod_{1 \le i < j \le n} \Phi\left[c + \left(Q_Z \Lambda Q_Z^\top\right)_{ij}\right]^{y_{ij}} \left\{1 - \Phi\left[c + \left(Q_Z \Lambda Q_Z^\top\right)_{ij}\right]\right\}^{1-y_{ij}}
$$
$$
\times \prod_{i=1}^{p} \prod_{j=1}^{k} \text{Normal}\left(z_{ij} \mid 0, \theta_{ij}/\ell\right) \text{Beta}\left[\theta_{ij} \mid \ell/2, (1-\ell)/2\right]
$$
$$
\times \text{Normal}\left(c \mid 0, 10^2\right) \times \mathbb{1}_{(0,1)}(\ell) \times \prod_{j=1}^{k} \text{Normal}\left(\lambda_j \mid 0, p\right). \tag{9}
$$

Given a Markov chain $\{c_t, Z_t, \Theta_t, \Lambda_t, \ell_t\}_{t=1}^{T}$ whose stationary distribution has density (9), we approximate the marginal posterior distribution of Q by $\{Q_t\}_{t=1}^{T}$ where $Q_t = Z_t\left(Z_t^\top Z_t\right)^{-1/2}$ for each t.

Previous analyses of the protein interaction data based on the network eigenmodel have differed in their choice of dimension k. Hoff [31] and Jauch et al. [35] chose $k = 3$ without much discussion. The dimension selection procedure of Loyal and Chen [42] led to a choice of $k = 5$ which was further supported by a posterior predictive analysis. In our experiments, we set $k = 5$.

In the first experiment, we fit both the sparse network eigenmodel (with the sparsity-inducing prior for Q described above) and the standard network eigenmodel (with the uniform prior for Q) to the full protein network data set. Figure 4 provides posterior summary plots. Panel (a) shows the

marginal posterior distribution of the sparsity parameter ℓ. The posterior mean is around .20 while $(0.15, 0.25)$ is an approximate 95% credible interval. The small estimated value of ℓ indicates that the data favor a sparse Q. Panel (b) compares the entries of our point estimate of $Q\Lambda Q^T$ (the element-wise posterior median) under the two priors. As we might expect, the point estimate of $Q\Lambda Q^T$ under the sparsity-inducing prior has many nearly zero entries with a smaller number of very large entries. This leads to greater interpretability. Nearly zero entries of $Q\Lambda Q^T$ correspond to pairs of proteins whose probability of interaction is close to the baseline value of $\Phi(c)$, while large entries correspond to pairs of proteins whose probability of interaction deviates from that baseline value.

In the second experiment, we compared the out-of-sample predictive performance of the sparse network eigenmodel to that of the standard network eigenmodel. We repeatedly fit both models to randomly selected subsets of the dyads in the protein network data, estimated the probabilities that connections exist for the held out dyads, then calculated the area under the curve (AUC) for the two models. The procedure was repeated 16 times with 1/3 of the dyads held out each time. We took posterior means as our point estimates for the probabilities. We found that the predictive performance of the sparse network eigenmodel was comparable to that of the standard network eigenmodel. In particular, the AUC of the sparse network eigenmodel was greater in 9/16 of the repetitions with a median difference of .5% in favor of the sparse network eigenmodel. In light of its greater interpretability and comparable predictive performance, the sparse network eigenmodel is an appealing alternative to the standard network eigenmodel for this protein interaction data.

5.2 Smooth PCA of Hawaii ocean oxygen data

The Hawaii Ocean Time-series (HOT) program has measured physical and biogeochemical variables at a deep-water sampling site 100 km north of Oahu since October 1988. The measurements are made across a range of depths on an approximately monthly basis. Long-term time series like those collected by HOT are valuable in climate studies, because they help researchers understand variability and changes in physical and biogeochemical processes in the ocean.

In this section, we fit the model-based SVD from Example 1 to oxygen concentration data collected by the HOT program and analyzed in Zhang et al. [65]. To account for the functional relationship between oxygen concentration and depth, we assign V a prior appropriate for a semi-orthogonal matrix with smooth structure, as discussed in Section 4.2. The model and prior specification are similar to those from the PCA of the Canadian weather data in Section 5.2 of Jauch et al. [35]. That analysis can be viewed as another illustration of the proposed family of prior distributions. The oxygen concentration dataset can be obtained from the Hawaii Ocean Time-series Data Organization & Graphical System (HOT-DOGS) or the Github repository associated with Zhang et al. [65].

The data include oxygen concentrations in units of μmol/kg observed on $n = 133$ separate occasions from January 1, 2008 to December 31, 2021. On each occasion, oxygen concentrations were measured at $p = 100$ evenly spaced depths ranging from 2 meters to 200 meters below the ocean surface. The raw data matrix Y_{raw} has $n = 133$ rows and $p = 100$ columns with entry (i, j) recording the oxygen concentration at a depth of $2j$ meters on the ith occasion.

We subtract column means from Y_{raw} and model the resulting matrix as $Y = UDV^\top + \sigma E$. The semi-orthogonal matrices $U \in \mathcal{V}(k, n)$ and $V \in \mathcal{V}(k, p)$, the diagonal matrix $D = \text{diag}(d_1, \dots, d_k)$ with $d_1, \dots, d_k > 0$, and the scale $\sigma > 0$ are unknown parameters. The entries of E are indepen-

dent standard normal errors. As in Section 5.2 of Jauch et al. [35], we can view this model-based SVD as a PCA of the functional data. The rows of UD contain the principal component scores for each observation, while the columns of V form the corresponding basis of principal component curves. In this analysis, we set $k = 4$.

We assign the semi-orthogonal matrix V a hierarchical prior chosen to reflect the functional nature of the oxygen concentration data. In particular, we let

$$V \mid \rho \sim \text{MACG}\left(\Omega\right)$$
$$1/\rho \sim \text{Gamma}(\alpha, \beta)$$

where $\Omega = (\Omega_{ij})$ is constructed from the squared exponential correlation function from Proposition 3 with $\Omega_{ij} = C_\rho^{\text{SE}}(|2i - 2j|)$. The squared exponential correlation function can be obtained from the Matérn correlation function of Example 4 by letting $\nu \to \infty$. The length-scale parameter ρ, which controls the "wiggliness" of the principal component curves, is treated as unknown and assigned an inverse gamma prior. The inverse gamma prior places negligible probability mass near zero, effectively ruling out values of ρ that are too small to make scientific sense. We select the shape α and the rate β of the inverse gamma prior according to the heuristic of Section 5.2 of Jauch et al. [35], which we review here. For a fixed ρ, we can reason (based on Theorem 2, Proposition 3, and the fact that $p \gg k$) that *a priori* each column of V will resemble a centered Gaussian process (GP) with a squared exponential correlation function and length-scale ρ. The expected number of zero crossings by such a GP within the interval $[0, T]$ is $T/(\pi\rho)$ [1] [51, 40, 50]. We choose α and β so that ρ has a prior mean of $200/(2\pi)$ and a prior standard deviation (SD) of 10. If ρ were fixed at its prior mean, the expected number of zero crossings of the principal component curves would be approximately two.

We now specify the priors for the remaining parameters U, σ^2, and $d_1, \ldots, d_k$. We assign the semi-orthogonal matrix U a uniform prior. The priors for σ^2 and $d_1, \ldots, d_k$ are inverse gamma and truncated normal:

$$1/\sigma^2 \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right)$$

$$p(d_1, ..., d_k) \propto \mathbb{1}\left\{d_1, ..., d_k > 0\right\} \prod_{i=1}^{k} \text{Normal}(d_i \mid 0, \tau^2).$$

We select the hyperparameters ν, s^2, and τ using the empirical Bayes strategy from Section 5.2 of Jauch et al. [35], which is reviewed in the supplementary material.

We simulate from the posterior distribution via polar expansion combined with adaptive HMC, as described in Section 2. As our point estimate of V, we take the first $k = 4$ right singular vectors of the posterior mean of $UDV^\top$. The left side of Figure 5 compares our point estimate (in black) to the results of classical PCA (in gray). The principal component curves produced by our approach are slightly smoother than those obtained from classical PCA but are overall very similar. (The differences are more pronounced in the analysis of the Canadian weather data in Section 5.2 of Jauch et al. [35].) The estimated principal component curves resemble those from Zhang et al. [65] and lead to similar interpretations. The first PC relates to the overall oxygen concentration level across the entire range of depths, with a higher PC score corresponding to a higher concentration.

---

[1] The formula stated in Jauch et al. [35] is off by a factor of two.

The second PC relates to differences in oxygen concentrations at middle depths (approximately 50-125 meters) compared to deeper depths (approximately 150-200 meters), with a higher PC score corresponding to increased oxygen concentration at middle depths. The third and fourth PCs relate to more complex phenomena. The right side of Figure 5 compares a histogram estimate of the posterior density of ρ with its inverse gamma prior density. Compared to the prior (mean 31.83, SD 10), the posterior of ρ is much more concentrated with with a slightly lower mean (mean 25.13, SD 1.41), indicating that we can learn a suitable length-scale ρ from the data.

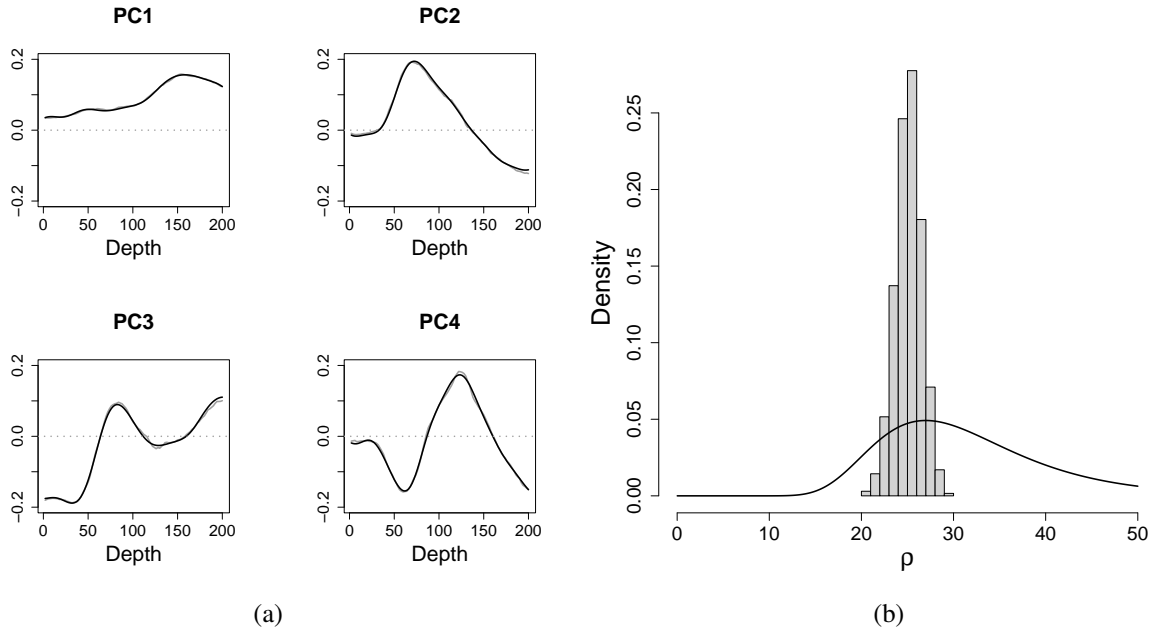(a)   (b)

Figure 5: The left side compares our point estimate (in black) to the results of classical PCA (in gray). The right side compares a histogram estimate of the posterior density of ρ with its inverse gamma prior density.

6   Discussion

We presented an approach to defining prior distributions for structured semi-orthogonal matrices that leads to tractable posterior inference via parameter expanded MCMC. The key idea was to introduce a suitably constructed latent random matrix X whose projection Q_X onto the Stiefel manifold is identified with the semi-orthogonal matrix parameter. We demonstrated that Q_X inherits the invariance properties of X (Theorem 1) and established that the Wasserstein distance between the distribution of elements of $\sqrt{p}Q_X$ and the distribution of the corresponding elements of X is small when $k \ll p$ (Theorem 2). These theoretical results, which may be of independent interest to those working in high-dimensional probability and random matrix theory, imply that we can incorporate structural assumptions into the prior distribution for a semi-orthogonal matrix by building these features into the distribution of X. To incorporate sparsity, we introduced a novel shrinkage prior with a number of appealing properties, including a representation as a scale mixture of normal distributions. To incorporate smoothness, we considered a variety of correlation

functions. We illustrated these priors in applications to protein interaction data and ocean oxygen data.

There are several interesting directions for future research:

- The theoretical contributions of this article were focused on properties of the proposed family of prior distributions. A natural next step would be to investigate properties of the resulting posterior distribution in specific models like the model-based SVD of Example 1, the network eigenmodel of Example 2, or the spiked covariance model [10, 24, 61] under the assumption that a semi-orthogonal matrix parameter is sparse or smooth.

- In some settings, it is natural to assume that a semi-orthogonal matrix parameter is *jointly sparse* with entire rows of zeros [57, 11, 62] or has heterogeneous structure with different levels of sparsity or smoothness across its columns [46]. The construction of Section 3, which requires that the columns of X are i.i.d., cannot easily accommodate these structural assumptions. Another natural next step would be to generalize that construction and develop analogous theory with an eye toward addressing this limitation.

- In applications involving structured semi-orthogonal matrix parameters, there is almost always uncertainty regarding the number of columns k. To account for this uncertainty, we could incorporate the model averaging and dimension selection ideas proposed in Hoff [30] or Loyal and Chen [42].

- An interesting and challenging mathematical problem would be to investigate the distance between the distributions of $\sqrt{p}Q_X$ and X (or transformations thereof) in terms of other probability metrics [27].

A   Proofs of results in Section 3

In this section, we present the proofs of our theoretical results in Section 3.

Proof of Theorem 1. Suppose that X is invariant to left multiplication by elements of $\mathcal{L}$ and right multiplication by elements of $\mathcal{R}$. More formally, let $L \in \mathcal{L}$ and $R \in \mathcal{R}$ be given and suppose that $X \overset{d}{=} LXR$. We write further $VDV^\top$ for the eigendecomposition of $X^\top X$. We aim to prove that the polar decomposition Q_{LXR} of LXR is equal to $LQ_X R$,

$$
\begin{aligned}
Q_{LXR} &\overset{d}{=} LXR(R^\top X^\top L^\top LXR)^{-1/2} \\
&= LXR(R^\top X^\top XR)^{-1/2} \\
&= LXR(R^\top VDV^\top R)^{-1/2} \\
&= LXRR^\top VD^{-1/2}V^\top R \\
&= LXVD^{-1/2}V^\top R \\
&= LX(X^\top X)^{-1/2}R \\
&= LQ_X R.
\end{aligned}
$$

$\square$

As in Section 3.2, for the subsequent proofs we use the notation X_p and Q_{X_p} to emphasize their dependence on p for our asymptotic results.

Proof of Theorem 2. We prove that the Wasserstein distance between finitely many entries m of $\sqrt{p}Q_{X_p}$ and the corresponding entries in X_p goes to zero as $k, p \to \infty$ under the assumptions stated in the proposition. For each p let

$$Q_{X_p}(m) = \left(q_1^{(p)}, \ldots, q_m^{(p)} \right)^\top \quad \text{and} \quad X_p(m) = \left(x_1^{(p)}, \ldots, x_m^{(p)} \right)^\top \tag{10}$$

be respectively vectors of m entries of Q_{X_p} and X_p corresponding to a fixed set of indices.

In order to find a more compact way of selecting m entries and to subsequently use that the columns of X_p are independent, we introduce some more notation. Let e_i denote the ith unit vector in $\mathbb{R}^k$. Define further matrices $D_m(m_i) \in \mathbb{R}^{m_i \times p}$, $i = 1, \ldots, k$. Each matrix $D_m(m_i)$ has one nonzero entry per row, such that, $D_m(m_i)$ has m_i nonzero entries. Applied to a column vector, the matrix $D_m(m_i)$ selects m_i entries. We assume that $\sum_{i=1}^k m_i = m$ such that m entries get selected by applying all matrices $D_m(m_i) \in \mathbb{R}^{m_i \times p}$, $i = 1, \ldots, k$.

Then, the Wasserstein distance of order two as defined in Definition 3.1 between m entries of $\sqrt{p}Q_{X_p}$ and the corresponding entries in X_p can be written as

$$W_2 \left[\text{law} \left(\sqrt{p}\, Q_{X_p}(m) \right), \text{law} \left(X_p(m) \right) \right]^2 \leq \mathrm{E} \left\| \sqrt{p}\, Q_{X_p}(m) - X_p(m) \right\|_F^2 \tag{11}$$

$$= \sum_{i=1}^k \mathrm{E} \left\| D_m(m_i)(\sqrt{p}\, Q_{X_p} - X_p)e_i \right\|_F^2 \tag{12}$$

$$= \frac{m}{k} \sum_{i=1}^k \mathrm{E} \left\| D_m(m_i)(\sqrt{p}\, Q_{X_p} - X_p) \right\|_F^2, \tag{13}$$

where in (11) we approximated the Wasserstein distance of order two by the Frobenius norm. The definition of $D_m(m_i)$ allows us to rewrite the Frobenius norm of m entries in terms of the matrices Q_{X_p} and X_p in (12). Finally, we use that the columns of X_p are independent and get (13).

With explanations given below, for k, p large, the Frobenius norm in (13) can be further bounded as

$$\| D_m(m_i)(\sqrt{p}Q_{X_p} - X_p) \|_F = \| D_m(m_i)(\sqrt{p}X_p(X_p^\top X_p)^{-\frac{1}{2}} - X_p) \|_F$$

$$= \left\| D_m(m_i)X_p \left(\left(\frac{1}{p}X_p^\top X_p \right)^{-\frac{1}{2}} - I_k \right) \right\|_F$$

$$\leq \| D_m(m_i)X_p \|_F \left\| I_k - \left(\frac{1}{p}X_p^\top X_p \right)^{-\frac{1}{2}} \right\| \tag{14}$$

$$= \| D_m(m_i)X_p \|_F \left\| \sum_{i=1}^\infty \binom{-1/2}{i} \left(\frac{1}{p}X_p^\top X_p - I_k \right)^i \right\| \tag{15}$$

$$\leq \| D_m(m_i)X_p \|_F \sum_{i=1}^\infty \left| \binom{-1/2}{i} \right| \left\| \frac{1}{p}X_p^\top X_p - I_k \right\|^i \tag{16}$$

$$\leq \| D_m(m_i)X_p \|_F \left\| \frac{1}{p}X_p^\top X_p - I_k \right\| \sum_{i=1}^\infty \left| \binom{-1/2}{i} \right| \tag{17}$$

18

$$= 2\|D_m(m_i)X_p\|_F \left\|\frac{1}{p}X_p^\top X_p - I_k\right\|, \tag{18}$$

where (14) follows since $\|AB\|_F \le \|A\|\|B\|_F$ for matrices A, B; see Theorem 1 in [21]. For Taylor expansions of matrix functions and the Taylor expansion of the square root of a matrix in particular, we refer to Section 6.8.1. in [28] and (6.38) therein. Theorem 4.7 in [28] states that the spectrum of $\frac{1}{p}X_p^\top X_p - I_k$ needs to be smaller than one to ensure convergence of the infinite series expansion. The Taylor expansion of $x \mapsto x^{-\frac{1}{2}}$ in (15) converges due to Lemma 2, which ensures that $\|\frac{1}{p}X_p^\top X_p - I_k\| < 1$ almost surely as long as $\frac{k}{p} \to 0$. For (16), we use triangle inequality which applies to the infinite series due to the subsequently proved convergence of the series. Submultiplicativity of the spectral norm allows us to exchange norm and exponent. The fact that $\|\frac{1}{p}X_p^\top X_p - I_k\| < 1$ almost surely by Lemma 2, is used to conclude that $\left\|\frac{1}{p}X_p^\top X_p - I_k\right\|^i \le \left\|\frac{1}{p}X_p^\top X_p - I_k\right\|$ almost surely and to infer (17). For (18), the series over the binomial coefficient can be calculated explicitly as

$$\sum_{i=1}^{\infty} \left|\binom{-1/2}{i}\right| = 2\sum_{i=1}^{\infty} \left|\binom{1/2}{i}\right| = 2\sum_{i=0}^{\infty} \left|\binom{1/2}{i}\right| - 2 = 2; \tag{19}$$

see p. 1206 in [60].

Finally, combining (13) and (18), we can bound the Wasserstein distance of order two by

$$W_2 \left[\text{law}\left(\sqrt{p}\,Q_{X_p}(m)\right), \text{law}\left(X_p(m)\right)\right]^2 \le \frac{m}{k}\sum_{i=1}^{k} \text{E}\left\|D_m(m_i)(\sqrt{p}\,Q_{X_p} - X_p)\right\|_F^2$$

$$\le 4\frac{m}{k}\sum_{i=1}^{k} \text{E}\left[\|D_m(m_i)X_p\|_F^2 \left\|\frac{1}{p}X_p^\top X_p - I_k\right\|^2\right] \tag{20}$$

$$\le 4c\frac{m}{k}\sum_{i=1}^{k} \text{E}\|D_m X_p\|_F^2 \boldsymbol{c}_\Omega(p)\frac{k}{p} \tag{21}$$

$$\le c\,\boldsymbol{c}_\Omega(p)m^2\frac{k}{p}, \tag{22}$$

where (20) is an application of (18). In (21), we used the convergence rate developed in Lemma 2. Let z_j denote the jth column of Z_p. Then, the last relation (22) is an explicit calculation of the expected value

$$\sum_{i=1}^{k} \text{E}\|D_m(m_i)X_p\|_F^2 = \sum_{i=1}^{k} \text{E}\,\text{Tr}\left[X_p^\top D_m^\top(m_i)D_m(m_i)X_p\right]$$

$$= \sum_{i=1}^{k} \text{E}\,\text{Tr}\left[Z_p^\top \Omega_p^{\frac{1}{2}} D_m^\top(m_i)D_m(m_i)\Omega_p^{\frac{1}{2}} Z_p\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} \text{E}\,\text{Tr}\left[z_j^\top \Omega_p^{\frac{1}{2}} D_m^\top(m_i)D_m(m_i)\Omega_p^{\frac{1}{2}} z_j\right]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} \mathrm{Tr} \left[\Omega_p^{\frac{1}{2}} D_m^\top(m_i) D_m(m_i) \Omega_p^{\frac{1}{2}} \right] = k \sum_{i=1}^{k} m_i = km.$$

$\square$

Proof of Lemma 1. By Lemma 3 below, we have

$$W_2 \left[\mathrm{law} \left(\sqrt{p} \, Q_{X_p} \right), \mathrm{law} \left(X_p \right) \right]^2 = \mathrm{E} \left\| \sqrt{p} \, Q_{X_p} - X_p \right\|_F^2. \tag{23}$$

The squared Frobenius norm can be written as

$$\begin{aligned}
\left\| \sqrt{p} \, Q_{X_p} - X_p \right\|_F^2 &= \mathrm{Tr} \left[\left(\sqrt{p} \, Q_{X_p} - X_p \right)^\top \left(\sqrt{p} \, Q_{X_p} - X_p \right) \right] \\
&= \mathrm{Tr} \left[p \, Q_{X_p}^\top Q_{X_p} \right] - 2 \, \mathrm{Tr} \left[\sqrt{p} \, Q_{X_p}^\top X_p \right] + \mathrm{Tr} \left[X_p^\top X_p \right].
\end{aligned}$$

We evaluate the three terms and their expectations separately and then combine the results. In terms of Z_p and Ω_p, the matrix Q_{X_p} is given as

$$\begin{aligned}
Q_{X_p} &= X_p (X_p^\top X_p)^{-1/2} \\
&= \Omega_p^{1/2} Z_p (Z_p^\top \Omega_p Z_p)^{-1/2}.
\end{aligned}$$

Thus, the first term is

$$\begin{aligned}
\mathrm{Tr} \left[p \, Q_{X_p}^\top Q_{X_p} \right] &= p \, \mathrm{Tr} \left[(Z_p^\top \Omega_p Z_p)^{-1/2} Z_p^\top \Omega_p^{1/2} \Omega_p^{1/2} Z_p (Z_p^\top \Omega_p Z_p)^{-1/2} \right] \\
&= p \, \mathrm{Tr} \left[(Z_p^\top \Omega_p Z_p)^{-1} Z_p^\top \Omega_p Z_p \right] \\
&= p \, \mathrm{Tr} \left[I_k \right] \\
&= pk.
\end{aligned}$$

Then, of course, we have $\mathrm{E} \, \mathrm{Tr} \left[p \, Q_{X_p}^\top Q_{X_p} \right] = pk$. The third term is $\mathrm{Tr} \left[X_p^\top X_p \right] = \mathrm{Tr} \left[Z_p^\top \Omega_p Z_p \right]$. Let Z_j denote the jth column of Z_p. Then

$$\begin{aligned}
\mathrm{E} \, \mathrm{Tr} \left[Z_p^\top \Omega_p Z_p \right] &= \mathrm{E} \left[\sum_{j=1}^{k} Z_j^\top \Omega_p Z_j \right] \\
&= k \, \mathrm{E} \left[\sum_{j=1}^{k} Z_1^\top \Omega_p Z_1 \right] \\
&= k \, \mathrm{Tr} \, \Omega_p \\
&= pk.
\end{aligned}$$

The second term is

$$\begin{aligned}
-2 \, \mathrm{Tr} \left[\sqrt{p} \, Q_{X_p}^\top X_p \right] &= -2 \, \mathrm{Tr} \left[\sqrt{p} \left(Z_p^\top \Omega_p Z_p \right)^{-1/2} Z_p^\top \Omega_p^{1/2} \Omega_p^{1/2} Z_p \right] \\
&= -2 \, \mathrm{Tr} \left[\left(p^{-1} Z_p^\top \Omega_p Z_p \right)^{-1/2} Z_p^\top \Omega_p Z_p \right]
\end{aligned}$$

$$
\begin{aligned}
&= -2p \operatorname{Tr}\left[\left(p^{-1} Z_p^\top \Omega_p Z_p\right)^{-1/2} p^{-1} Z_p^\top \Omega_p Z_p\right] \\
&= -2p \operatorname{Tr}\left[\left(p^{-1} Z_p^\top \Omega_p Z_p\right)^{1/2}\right] \\
&= -2p \operatorname{Tr}\left[S_p^{1/2}\right]
\end{aligned}
$$

where $S_p = p^{-1} Z_p^\top \Omega_p Z_p$. Let $\lambda_1 \geq \cdots \geq \lambda_k \geq 0$ be the eigenvalues of S_p. Putting this all together, we have that

$$
\begin{aligned}
\mathrm{E}\left\|\sqrt{p}\, Q_{X_p} - X_p\right\|_F^2 &= 2pk - 2p\, \mathrm{E}\operatorname{Tr}\left[S_p^{1/2}\right] \\
&= 2p\, \mathrm{E}\operatorname{Tr}\left[I_k - S_p^{1/2}\right] \\
&= 2p\, \mathrm{E}\left[\sum_{j=1}^{k}(1 - \lambda_j^{1/2})\right].
\end{aligned}
$$

$\square$

B Technical lemmas and their proofs

The following Lemma 2 is based on Lemma 6.1 in [49]. Theorem 4 in [14] gives a more general statement for the case $\Omega_p = I_p$. In our case, it is sufficient to require finite fourth moments for Z_{ij}. In particular, it is not necessary to assume $\mathrm{E}\,|Z_{ij}|^{6+\varepsilon_0} < \infty$. All assumptions on Ω_p in Theorem 2 are naturally satisfied for $\Omega_p = I_p$.

Lemma 2. *Suppose the assumptions in Theorem 2 are satisfied. Then,*

$$
\left\|\frac{1}{p} X_p^\top X_p - I_k\right\| = O_{a.s.}\left(\sqrt{\frac{\mathbf{c}_\Omega(p)k}{p}}\right) \tag{24}
$$

for k, p growing according to the assumptions in Theorem 2 and $\mathbf{c}_\Omega(p) := p^{-1}\operatorname{Tr}\left(\Omega_p^2\right)$.

Proof. We first prove a high probability bound for $\sqrt{\frac{p}{\mathbf{c}_\Omega(p)k}}\left\|\frac{1}{p} X_p^\top X_p - I_k\right\|$ with subsequent application of the Borel-Cantelli lemma to infer almost sure convergence.

Recall that $X_p = \Omega_p^{\frac{1}{2}} Z_p$ with Z_p having i.i.d. entries. With $\mathbf{c}_\Omega(p) := p^{-1}\operatorname{Tr}\left(\Omega_p^2\right)$, we set

$$
\eta = 2 \limsup_{p \to \infty} \|\Omega_p\| \frac{1}{\sqrt{\mathbf{c}_\Omega(p)}} \tag{25}
$$

and

$$
\begin{aligned}
O_k &= \sqrt{\frac{p}{\mathbf{c}_\Omega(p)k}}\left(\frac{1}{p} X_p^\top X_p - I_k - \operatorname{diag}\left(\frac{1}{p} X_p^\top X_p - I_k\right)\right) \\
&= \sqrt{\frac{p}{\mathbf{c}_\Omega(p)k}}\left(\frac{1}{p} Z_p^\top \Omega_p Z_p - \operatorname{diag}\left(\frac{1}{p} Z_p^\top \Omega_p Z_p\right)\right),
\end{aligned} \tag{26}
$$

where $\operatorname{diag}(A)$ for a quadratic matrix A refers to the matrix A with all off-diagonal entries set to zero. With (26), we can write

$$
\sqrt{\frac{p}{\mathbf{c}_\Omega(p)k}}\left(\frac{1}{p} X_p^\top X_p - I_k\right) = O_k + \sqrt{\frac{p}{\mathbf{c}_\Omega(p)k}}\operatorname{diag}\left(\frac{1}{p} X_p^\top X_p - I_k\right). \tag{27}
$$

21

We follow the proof of Lemma 6.1 in [49]. Note that their results suffice for us to prove that for any $\varepsilon > 0$,

$$\mathrm{P}\left(\sqrt{\frac{p}{c_\Omega(p)k}}\left\|\frac{1}{p}X_p^\top X_p - I_k\right\| > \varepsilon\right) = o(k^{-1}). \tag{28}$$

In contrast, [14] prove that, for any $\varepsilon > 0, l > 0$

$$\mathrm{P}\left(\sqrt{\frac{p}{k}}\left\|\frac{1}{p}Z_p^\top Z_p - I_k\right\| > \sqrt{\frac{k}{p}}\right) = o(k^{-l}). \tag{29}$$

We aim to combine the arguments in [14] and [49] to get

$$\mathrm{P}\left(\sqrt{\frac{p}{c_\Omega(p)k}}\left\|\frac{1}{p}X_p^\top X_p - I_k\right\| > \varepsilon\right) = o(k^{-\ell}). \tag{30}$$

As in the proof of Lemma 6.1 in equations (S2.42), (S2.43), and (S2.44) in [49], it suffices to show that for any $\varepsilon > 0, \ell > 1$,

$$\mathrm{P}\left(\|O_k\| > \eta + \varepsilon\right) = o(k^{-\ell}), \tag{31}$$

$$\mathrm{P}\left(\frac{1}{\sqrt{kpc_\Omega(p)}}\max_{1 \le r \le k}\left|\sum_{s=1}^p (Z_{sr}^2 - 1)\right| > \varepsilon\right) = o(k^{-\ell}), \tag{32}$$

$$\mathrm{P}\left(\frac{1}{\sqrt{kpc_\Omega(p)}}\max_{1 \le r \le k}\left|\sum_{s \ne t}^p (\Omega_p)_{st} Z_{sr} Z_{ts}\right| > \varepsilon\right) = o(k^{-\ell}), \tag{33}$$

where $\Omega_p = ((\Omega_p)_{st})_{s,t=1,\dots,p}$ and noting that $(\Omega_p)_{ss} = 1$. We consider the three relations separately. First, (31) is inferred from equation (8) in [14]. Furthermore, (32) follows by equation (9) in [14]. Finally, (33) can be inferred by following (S2.47) in [49]. To be more precise, we rephrase (S2.47) here which gives, for $\delta > 0$,

$$\mathrm{P}\left(\frac{1}{\sqrt{kpc_\Omega(p)}}\max_{1 \le r \le k}\left|\sum_{s \ne t}^p (\Omega_p)_{st} Z_{sr} Z_{ts}\right| > \varepsilon\right) \le k(\varepsilon\sqrt{kpc_\Omega(p)})^{-(4+\delta)}K(\mathrm{Tr}(\Omega_p^2))^{2+\delta/2} \tag{34}$$

$$\le kp^{2+\delta/2}(\varepsilon kpc_\Omega(p))^{-(2+\delta/2)}K(c_\Omega(p))^{2+\delta/2} \tag{35}$$

$$= \varepsilon^{-(2+\delta/2)}k^{-(1+\delta/2)}K \tag{36}$$

for some constant K coming from Lemma S1.2 in [49].

Finally, we can infer, for some constant c,

$$\sum_{k=1}^\infty \mathrm{P}\left(\left\|\frac{1}{p}X_p^\top X_p - I_k\right\| > \sqrt{\frac{kc_\Omega(p)}{p}}\varepsilon\right) \le c\sum_{k=1}^\infty k^{-\ell} < \infty \tag{37}$$

and therefore, by Borel-Cantelli, almost sure boundedness. $\qquad\square$

As observed in [29], the polar decomposition is the nearest unitary matrix in the Frobenius norm. For completeness, we give a formal argument here, proving that the squared Wasserstein distance between X_p and $\sqrt{p}\,Q_{X_p}$ equals the corresponding L_2-distance.

Lemma 3. *Recall $X_p \in \mathbb{R}^{p \times k}$ and $Q_{X_p} = X_p(X_p^\top X_p)^{-1/2}$. Then, the Wasserstein distance in (3.1) satisfies*

$$W_2 \left[law \left(\sqrt{p}\, Q_{X_p} \right), law \left(X_p \right) \right]^2 = \mathrm{E} \left\| \sqrt{p}\, Q_{X_p} - X_p \right\|_F^2. \tag{38}$$

Proof. In order to show that (38) is indeed satisfied, we use Brenier's theorem (see [7]) to derive that the optimal transport map between X_p and $\sqrt{p}X_p(X_p^\top X_p)^{-1/2}$ is given by

$$T(X_p) = \sqrt{p}X_p(X_p^\top X_p)^{-1/2}.$$

By Brenier's theorem, for the two random matrices X_p (source) and $Y = \sqrt{p}X_p(X_p^\top X_p)^{-1/2}$ (target), the squared Wasserstein distance is equal to

$$W_2^2(\mu, \nu) = \inf_{T: T_\# \mu = \nu} \int_{\mathbb{R}^{p \times k}} \|X_p - T(X_p)\|_F^2 \, d\mu(X_p),$$

where $T_\# \mu = \nu$ means that the map T pushes the distribution μ (of X_p) to the distribution ν (of Y). Therefore, (38) is satisfied.

While not necessary for the proof, we note as an additional remark that Brenier's Theorem states further that if μ and ν are absolutely continuous measures with finite second moments, then the optimal transport map T exists and is unique, given by the partial derivative of a convex function ϕ through

$$T(X_p) = \nabla\phi(X_p).$$

For the transport map $T(X_p) = \sqrt{p}X_p(X_p^\top X_p)^{-1/2}$ we have

$$\phi(X_p) = \frac{\sqrt{p}}{2} \mathrm{Tr}(X_p(X_p^\top X_p)^{-1/2} X_p^\top).$$

In particular, the gradient of ϕ with respect to X_p satisfies,

$$\nabla\phi(X_p) = \sqrt{p}X_p(X_p^\top X_p)^{-1/2} = T(X_p).$$

Since the function $(X_p^\top X_p)^{-1/2}$ involves an inverse matrix square root, which is smooth and convex over positive semidefinite matrices and since the trace operation and scaling preserve convexity, ϕ is a valid convex potential function, and $T(X_p) = \nabla\phi(X_p)$ satisfies Brenier's conditions. $\qquad \square$

C Proofs of results in Section 4

Proof of Proposition 1. Proposition 1 presents a Gaussian scale mixture representation of the density (7) for $\ell \in (0, 1)$. To arrive at such a representation, we follow Andrews and Mallows [1].

The first theorem of Andrews and Mallows [1] provides a necessary and sufficient condition for the symmetric density (7) to admit a Gaussian scale mixture representation, i.e. for there to exist independent random variables U_{ij} and V_{ij} with U_{ij} standard normal such that $Z_{ij} = U_{ij}/V_{ij}$ has density (7). We must have

$$\left(-\frac{d}{dy} \right)^k f \left(y^{1/2} \mid \ell \right) \geq 0$$

for $y > 0$. Substituting in the explicit definition of $f(\cdot \mid \ell)$ from (7), the condition becomes

$$\left(-\frac{d}{dy}\right)^k \frac{(\ell/2)^{\ell/2}}{\Gamma(\ell/2)} y^{(\ell-1)/2} \exp\left(-\ell y/2\right) \geq 0$$

for $y > 0$. In other words, we need to show that the function $y \mapsto y^{(\ell-1)/2} \exp\left(-\ell y/2\right)$ is completely monotonic [44]. From the examples in (1.2) of [44], we can conclude that the functions $y \mapsto y^{(\ell-1)/2}$ and $y \mapsto \exp\left(-\ell y/2\right)$ are completely monotonic for $\ell \in (0,1)$. Theorem 1 of [44] allows us to conclude that the function $y \mapsto y^{(\ell-1)/2} \exp\left(-\ell y/2\right)$ is completely monotonic when $\ell \in (0,1)$ because it is a product of completely monotonic functions.

Now that we know a Gaussian scale mixture representation exists, we turn to finding the density of V_{ij}. The existence of the Gaussian scale mixture representation implies, for all non-zero z, that

$$f(z \mid \ell) = (2\pi)^{-1/2} \int_0^\infty v \exp\left(-v^2 z^2/2\right) \, dG(v) \tag{39}$$

where G is the distribution function of V_{ij}. This representation and the relation (2.2) in Andrews and Mallows [1] imply that the Laplace transform of G is

$$\int_0^\infty \exp\left(-sv\right) \, dG(v) = 2 \int_0^\infty f(z \mid \ell) \exp\left(-s^2 z^{-2}/2\right) \, dz$$

$$= 2\frac{(\ell/2)^{\ell/2}}{\Gamma(\ell/2)} \int_0^\infty z^{\ell-1} \exp\left[-\left(\ell z^2 + s^2/z^2\right)/2\right] \, dz$$

$$= \frac{(\ell/2)^{\ell/2}}{\Gamma(\ell/2)} \int_0^\infty y^{\ell/2-1} \exp\left[-\left(\ell y + s^2/y\right)/2\right] \, dy \tag{40}$$

$$= \frac{(\ell/2)^{\ell/2}}{\Gamma(\ell/2)} \frac{2K_{\ell/2}\left(\sqrt{\ell s^2}\right)}{(\ell/s^2)^{\ell/4}} \tag{41}$$

$$= \frac{2^{1-\ell/2}\ell^{\ell/4}}{\Gamma\left(\ell/2\right)} s^{\ell/2} K_{\ell/2}\left(\sqrt{\ell}s\right)$$

$$= \frac{2^{1-\ell/2}\ell^{\ell/4}}{\Gamma\left(\ell/2\right)} s^{\ell/2} K_{-\ell/2}\left(\sqrt{\ell}s\right) \tag{42}$$

$$= \int_1^\infty \exp\left(-\sqrt{\ell}st\right) \cdot \frac{2\left(t^2-1\right)^{-\frac{1+\ell}{2}}}{B\left(\frac{1-\ell}{2},\frac{\ell}{2}\right)} \, dt \tag{43}$$

$$= \int_{\sqrt{\ell}}^\infty \exp\left(-s\tilde{v}\right) \cdot \frac{2\ell^{-1/2}\left(\tilde{v}^2/\ell - 1\right)^{-\frac{1+\ell}{2}}}{B\left(\frac{1-\ell}{2},\frac{\ell}{2}\right)} \, d\tilde{v}. \tag{44}$$

The equality (40) comes from changing variables to $y = z^2$, while (41) comes from recognizing that the integrand in (40) is proportional to a generalized inverse Gaussian density [37]. The equality (42) is an application of 10.27.3 from DLMF [20], while (43) is an application of 10.32.8 from DLMF [20]. Finally, (44) is the result of changing variables to $\tilde{v} = \sqrt{\ell}t$. We can conclude from this derivation that the random variable V_{ij} has density

$$v \mapsto \frac{2\ell^{-1/2}\left(v^2/\ell - 1\right)^{-\frac{1+\ell}{2}}}{B\left(\frac{1-\ell}{2},\frac{\ell}{2}\right)} \mathbb{1}_{\left(\sqrt{\ell},\infty\right)}(v). \tag{45}$$

24

We now define a random variable $W_{ij} = V_{ij}^2/\ell - 1$. One use can use the density (45) and a change of variables to show that $W_{ij} \sim \text{BetaPrime}\left(\frac{1-\ell}{2}, \frac{\ell}{2}\right)$ [36]. By well-known properties of the beta prime and beta distributions, we have that that $W_{ij} \overset{d}{=} (1-\theta_{ij})/\theta_{ij}$ where $\theta_{ij} \sim \text{Beta}\left(\frac{\ell}{2}, \frac{1-\ell}{2}\right)$. Proposition 1 follows from observing that

$$Z_{ij} = \frac{U_{ij}}{V_{ij}} = \frac{U_{ij}}{\sqrt{\ell\,(W_{ij}+1)}} \overset{d}{=} \frac{U_{ij}}{\sqrt{\ell\left(\frac{1-\theta_{ij}}{\theta_{ij}}+1\right)}} = U_{ij}\sqrt{\frac{\theta_{ij}}{\ell}}.$$

$\square$

Proof of Proposition 2. **Part I.** Both of the claims in Part I are immediate.

Part II. The claim follows from the definition of (7) and basic properties of limits.

Part III. That the entries of Z are i.i.d. follows by assumption. We need to verify that a random variable Z_{ij} with density (7) has a mean of zero, a variance of one, and a finite fourth moment. There are two cases: $\ell = 1$ and $\ell \in (0, 1)$. In the former case, Z_{ij} is a standard normal random variable and we can immediately conclude that Z_{ij} has a mean of zero, a variance of one, and a finite fourth moment. In the latter case, in which $\ell \in (0, 1)$, Proposition 1 establishes that Z_{ij} can be generated as $Z_{ij} \mid \theta_{ij} \overset{\text{ind.}}{\sim} \text{Normal}(0, \theta_{ij}/\ell)$ where $\theta_{ij} \overset{\text{i.i.d.}}{\sim} \text{Beta}\left[\ell/2, (1-\ell)/2\right]$. By the law of total expectation, we have that $\text{E}(Z_{ij}) = \text{E}\left[\text{E}(Z_{ij} \mid \theta_{ij})\right] = 0$. Again by the law of total expectation, we can calculate

$$\text{E}\left(|Z_{ij}|^r\right) = \text{E}\left[\text{E}\left(|Z_{ij}|^r \mid \theta_{ij}\right)\right]$$
$$= \int \left[\int \frac{z^r}{\sqrt{2\pi\theta/\ell}}\exp\left(-\frac{z^2}{2\theta/\ell}\right)dz\right]\frac{\theta^{\ell/2-1}(1-\theta)^{(1-\ell)/2-1}}{B\left[\ell/2, (1-\ell)/2\right]}d\theta$$
$$= \int (\theta/\ell)^{r/2}\,(r-1)!!\,\frac{\theta^{\ell/2-1}(1-\theta)^{(1-\ell)/2-1}}{B\left[\ell/2, (1-\ell)/2\right]}d\theta \tag{46}$$
$$= \frac{(r-1)!!}{\ell^{r/2}}\int \theta^{r/2}\,\frac{\theta^{\ell/2-1}(1-\theta)^{(1-\ell)/2-1}}{B\left[\ell/2, (1-\ell)/2\right]}d\theta$$
$$= \frac{(r-1)!!}{\ell^{r/2}}\prod_{s=0}^{r/2-1}\frac{\ell/2+s}{1/2+s} \tag{47}$$

for $r = 2, 4$. The equality (46) is the result of substituting in the expression for the rth moment of the Gaussian distribution. The double factorial $n!!$ is defined as the product of all positive integers from 1 to n having the same parity as n. The equality (47) is a consequence of substituting in the expression for the $r/2$-th raw moment of the beta distribution. Plugging $r = 2$ into (47) yields $\text{Var}(Z_{ij}) = \text{E}(|Z_{ij}|^2) = 1$. Setting $r = 4$ yields $\text{E}(|Z_{ij}|^4) = 1 + 2/\ell < \infty$.

Part IV. The claim follows from a change of variables. $\square$

D Discussion on Assumption 2

In this section, we verify Assumption 2 for different for the examples of correlation matrices given in Proposition 3.

Proof of Proposition 3. **Part I** For the first property in Assumption 2, note that

$$\frac{1}{p} \operatorname{Tr}(\Omega_p^2) = \frac{1}{p} \sum_{i,j=1}^{p} \Omega_{ij}^2 = \frac{1}{p} \sum_{i,j=1}^{p} \rho^{2|i-j|}$$

$$= \sum_{|k|<p} \left(1 - \frac{k}{p}\right) \rho^{2|k|} \to \sum_{k=-\infty}^{\infty} \rho^{2|k|} \tag{48}$$

as $p \to \infty$. For the second property in Assumption 2, we get, with explanations given below,

$$\|\Omega\| \le \|\Omega\|_1 = \max_{j=1,\ldots,p} \sum_{i=1}^{p} |\rho^{|i-j|}|$$

$$= \max_{j=1,\ldots,p} \sum_{k=j-p}^{j-1} |\rho^{|k|}|$$

$$\le \max_{j=1,\ldots,p} \int_{j-p-1}^{j-1} \rho^{|x|} dx$$

$$= 2b \max_{i=1,\ldots,p} \int_{j-p-1}^{j-1} \frac{1}{2b} \exp\left[-\frac{|x|}{b}\right] dx \tag{49}$$

$$\le 2b = 2\frac{1}{\log(1/\rho)}, \tag{50}$$

where (49) follows since $\rho^{|x|} = \exp(|x| \log(\rho)) = \exp(-|x| \log(1/\rho))$ and setting $b = 1/\log(1/\rho)$. Finally, (50) is due to integrating the density of the Laplace distribution over the real line.

    Part II For the first property in Assumption 2, note that

$$\frac{1}{p} \operatorname{Tr}(\Omega_p^2) = \frac{1}{p} \sum_{i,j=1}^{p} \Omega_{ij}^2 = \frac{1}{p} \sum_{i,j=1}^{p} \exp\left[-(i-j)^2/\rho^2\right]$$

$$= \frac{1}{p} \sum_{|k|<p} \left(1 - |k|\right) \exp\left[-k^2/\rho^2\right]$$

$$= \sum_{|k|<p} \left(1 - \frac{|k|}{p}\right) \exp\left[-k^2/\rho^2\right] \to \sum_{k=-\infty}^{\infty} \exp\left[-k^2/\rho^2\right], \tag{51}$$

as $p \to \infty$. The infinite series in (51) is also known as the theta function and is finite.

    For the second property in Assumption 2, we get

$$\|\Omega_p\| \le \|\Omega_p\|_1 = \max_{j=1,\ldots,p} \sum_{i=1}^{p} |\exp\left[-\frac{1}{2}(i-j)^2/\rho^2\right]| \tag{52}$$

$$= \max_{j=1,\ldots,p} \sum_{i=j-p}^{j-1} |\exp\left[-\frac{1}{2}k^2/\rho^2\right]|$$

$$\le \max_{j=1,\ldots,p} \int_{j-p-1}^{j-1} \exp\left[-\frac{1}{2}x^2/\rho^2\right] dx$$

$$= \sqrt{2\pi\rho^2} \max_{j=1,\ldots,p} \int_{j-p-1}^{j-1} \frac{1}{\sqrt{2\pi\rho^2}} \exp\left[-\frac{1}{2}x^2/\rho^2\right] dx \leq \sqrt{2\pi\rho^2}, \qquad (53)$$

where (52) follows since the operator norm can be bounded by the maximum absolute column sum. In (53), we bounded the integral by the density of a Gaussian random variable with zero mean and variance ρ^2 over the real line.

Part III For the first property in Assumption 2, note that

$$\frac{1}{p}\operatorname{Tr}(\Omega_p^2) = \frac{1}{p}\sum_{i,j=1}^{p}\Omega_{ij}^2 = \frac{1}{p}\sum_{i,j=1}^{p}C_\nu^2(|i-j|)$$

$$= 2\sum_{k=1}^{p-1}\left(1-\frac{k}{p}\right)C_\nu^2(k) + 1$$

$$= \frac{2^{2-2\nu}}{\Gamma^2(\nu)}2\sum_{k=1}^{p-1}\left(1-\frac{k}{p}\right)\left(\sqrt{2\nu}\frac{|k|}{\rho}\right)^{2\nu}K_\nu^2\left(\sqrt{2\nu}\frac{|k|}{\rho}\right) + 1$$

$$\to \frac{2^{2-2\nu}}{\Gamma^2(\nu)}2\sum_{k=1}^{\infty}\left(\sqrt{2\nu}\frac{k}{\rho}\right)^{2\nu}K_\nu^2\left(\sqrt{2\nu}\frac{k}{\rho}\right) + 1 < \infty,$$

where the infinite series is finite since $K_\nu(x) \sim \sqrt{\frac{\pi}{2}}x^{-\frac{1}{2}}e^{-x}$ for $\nu > 0$ and as $x \to \infty$ by (A.5) in [25]; and since

$$\frac{\pi}{2}\frac{2^{2-2\nu}}{\Gamma^2(\nu)}\sum_{k=1}^{\infty}\left(\sqrt{2\nu}\frac{k}{\rho}\right)\exp\left(-2\sqrt{2\nu}\frac{k}{\rho}\right) = -\frac{\pi}{2}\frac{2^{2-2\nu}}{\Gamma^2(\nu)}e^2/(-1+e^2)^2 < \infty.$$

For the second property in Assumption 2, we get, with explanations given below,

$$\|\Omega\| \leq \|\Omega\|_1 = \max_{j=1,\ldots,p}\sum_{i=1}^{p}C_\nu(|i-j|)$$

$$= \max_{j=1,\ldots,p}\sum_{k=j-p}^{j-1}C_\nu(|k|)$$

$$\leq \max_{j=1,\ldots,p}\int_{j-p-1}^{j-1}C_\nu(|x|)dx \qquad (54)$$

$$\leq \int_{-\infty}^{\infty}\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{|x|}{\rho}\right)^{\nu}K_\nu\left(\sqrt{2\nu}\frac{|x|}{\rho}\right)dx$$

$$\leq \frac{2^{1-\nu}}{\Gamma(\nu)}\frac{1}{\sqrt{2\nu}}\int_{-\infty}^{\infty}\left(\frac{|x|}{\rho}\right)^{\nu}K_\nu\left(\frac{|x|}{\rho}\right)dx$$

$$= \sqrt{2}\rho\sqrt{\frac{\pi}{2}}2^{1-\nu}\frac{\Gamma(\nu+\frac{1}{2})}{\Gamma(\nu)}\frac{1}{\sqrt{2\nu}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2}\rho}\frac{1}{\Gamma(\nu+\frac{1}{2})}\sqrt{\frac{2}{\pi}}\left(\frac{|x|}{\rho}\right)^{\nu}K_\nu\left(\frac{|x|}{\rho}\right)dx \qquad (55)$$

$$= \sqrt{2}\rho\sqrt{\frac{\pi}{2}}2^{1-\nu}\frac{\Gamma(\nu+\frac{1}{2})}{\Gamma(\nu)}\frac{1}{\sqrt{2\nu}}. \qquad (56)$$

27

The inequality (54) is due to the function $x \mapsto C_\nu(|x|)$ being monotonically decreasing. The monotonicity follows since $\frac{d}{dx}x^\nu K_\nu(x) = -x^\nu K_{\nu-1}(x)$; see for example [5]. Furthermore, $K_\nu(x) > 0$ for all $x > 0$ and $\nu \in \mathbb{R}$ as stated in Appendix A.1. in [25]. Due to equation (4.1.32) in [39] with the parameters therein chosen as $\sigma = \sqrt{2}\rho$, $\tau = \nu + \frac{1}{2}$, the expression in the integral in (55) is a density. We therefore infer (56).

Part IV The banded correlation matrix trivially satisfies Assumption 2 since the respective correlation functions satisfy Assumption 2. □

References

[1] Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 36:99–102, 1974.

[2] Armagan, A., Dunson, D. B., and Lee, J. Generalized double pareto shrinkage. *Statistica Sinica*, 23 (1):119, 2013.

[3] Bae, K. and Mallick, B. K. Gene selection using a two-level hierarchical bayesian model. *Bioinformatics*, 20(18):3423–3430, 07 2004.

[4] Bai, Z. D. and Yin, Y. Q. Convergence to the Semicircle Law. *The Annals of Probability*, 16(2): 863–875, 1988.

[5] Baricz, Á. Bounds for modified Bessel functions of the first and second kinds. *Proceedings of the Edinburgh Mathematical Society*, 53(3):575–599, 2010.

[6] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[7] Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[8] Bryan, J. G., Niles-Weed, J., and Hoff, P. D. The multirank likelihood for semiparametric canonical correlation analysis. *arXiv:2112.07465*, 2021.

[9] Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, page 531–537, 2005.

[10] Cai, T. T., Ma, Z., and Wu, Y. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074 – 3110, 2013.

[11] Cai, T., Ma, Z., and Wu, Y. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3):781–815, 2015.

[12] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: a probabilistic programming language. *Journal of Statistical Software*, 76:1–32, 2017.

[13] Carvalho, C. M., Polson, N. G., and Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[14] Chen, B. and Pan, G. Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, 18(4):1405–1420, 2012.

[15] Chen, B. and Pan, G. CLT for linear spectral statistics of normalized sample covariance matrices with the dimension much larger than the sample size. *Bernoulli*, 21(2):1089–1133, 2015.

[16] Chikuse, Y. The matrix angular central gaussian distribution. *Journal of Multivariate Analysis*, 33: 265–274, 1990.

[17] Chikuse, Y. *Statistics on Special Manifolds*. Springer New York, 2003.

[18] Choi, Y., Taylor, J., and Tibshirani, R. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, pages 2590–2617, 2017.

[19] Dawid, A. P. Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, 68:265–274, 1981.

[20] DLMF. *NIST Digital Library of Mathematical Functions*. `https://dlmf.nist.gov/`, Release 1.2.3 of 2024-12-15. URL `https://dlmf.nist.gov/`. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[21] Fang, Y., Loparo, K., and Feng, X. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

[22] Figueiredo, M. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159, 2003.

[23] Franks, A. M. and Hoff, P. Shared subspace models for multi-group covariance estimation. *Journal of Machine Learning Research*, 20(171):1–37, 2019.

[24] Gao, C. and Zhou, H. H. Rate-optimal posterior contraction for sparse pca. *Ann. Statist.*, 43:785–818, 2015.

[25] Gaunt, R. E. Inequalities for modified Bessel functions and their integrals. *Journal of Mathematical Analysis and Applications*, 420(1):373–386, 2014.

[26] Gavish, M. and Donoho, D. L. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

[27] Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[28] Higham, N. J. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[29] Higham, N. J. and Schreiber, R. S. Fast polar decomposition of an arbitrary matrix. *SIAM Journal on Scientific and Statistical Computing*, 11(4):648–655, 1990.

[30] Hoff, P. D. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102:674–685, 2007.

[31] Hoff, P. D. Simulation of the matrix bingham-von mises-fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18:438–456, 2009.

[32] Hoff, P. D. Equivariant and Scale-Free Tucker Decomposition Models. *Bayesian Analysis*, 11(3):627 – 648, 2016.

[33] Hoffman, M. D. and Gelman, A. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

[34] Jauch, M., Hoff, P. D., and Dunson, D. B. Random orthogonal matrices and the cayley transform. *Bernoulli*, 26:1560–1586, 2020.

[35] Jauch, M., Hoff, P. D., and Dunson, D. B. Monte carlo simulation on the stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, 30:622–631, 2021.

[36] Johnson, N. L., Kotz, S., and Balakrishnan, N. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995.

[37] Jørgensen, B. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media, 1982.

[38] Khatri, C. G. and Mardia, K. V. The von mises-fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):95–106, 1977.

[39] Kotz, S., Kozubowski, T., and Podgórski, K. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media, 2001.

[40] Kratz, M. F. Level crossings and other level functionals of stationary Gaussian processes. *Probability Surveys*, 3(none):230 – 288, 2006.

[41] Li, Y., Choi, K. S., Dunlop, B. W., Craighead, W. E., Mayberg, H. S., Garmire, L., Guo, Y., and Kang, J. Bsnmani: Bayesian scalar-on-network regression with manifold learning. *arXiv:2410.02965*, 2024.

[42] Loyal, J. D. and Chen, Y. A spike-and-slab prior for dimension selection in generalized linear network eigenmodels. *arXiv:2309.11654*, 2023.

[43] Meng, R. and Bouchard, K. E. Bayesian inference of structured latent spaces from neural population activity with the orthogonal stochastic linear mixing model. *PLOS Computational Biology*, 20(4): e1011975, 2024.

[44] Miller, K. S. and Samko, S. G. Completely monotonic functions. *Integral Transforms and Special Functions*, 12(4):389–402, 2001.

[45] Neal, R. M. *MCMC using Hamiltonian dynamics*, page 113–162. CRC Press, 2011.

[46] North, J. S., Risser, M. D., and Breidt, F. J. A flexible class of priors for orthonormal matrices with basis function-specific structure. *Spatial Statistics*, 64:100866, 2024.

[47] Park, H. G. Bayesian estimation of covariate assisted principal regression for brain functional connectivity. *Biostatistics*, page kxae023, 07 2024.

[48] Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. Bayesian inference over the stiefel manifold via the givens representation. *Bayesian Analysis*, 16:39–666, 2021.

[49] Qiu, J., Li, Z., and Yao, J. Asymptotic normality for eigenvalue statistics of a general sample covariance matrix when and applications. *The Annals of Statistics*, 51(3):1427 – 1451, 2023.

[50] Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for Machine learning*. MIT Press, 2006.

[51] Rice, S. O. Mathematical analysis of random noise. *The Bell System Technical Journal*, 24(1):46–156, 1945.

[52] Shepard, R., Brozell, S. R., and Gidofalvi, G. The representation and parametrization of orthogonal matrices. *The Journal of Physical Chemistry A*, 119:7924–7939, 2015.

[53] Srivastava, M. S. and Khatri, C. G. *An Introduction to Multivariate Statistics*. North-Holland/New York, 1979.

[54] Stam, A. J. Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces. *Journal of Applied Probability*, 19:221–228, 1982.

[55] Stouffer, D. B., Godoy, O., Dalla Riva, G. V., and Mayfield, M. M. The dimensionality of plant–plant competition. *bioRxiv*, 2021. doi: 10.1101/2021.11.10.467010.

[56] Villani, C. *Optimal transport: old and new*. Springer-Verlag, 2009.

[57] Vu, V. Q. and Lei, J. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905 – 2947, 2013.

[58] Wang, L. and Paul, D. Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$. *Journal of Multivariate Analysis*, 126:25–52, 2014.

[59] Watson, G. S. Limit theorems on high dimensional spheres and Stiefel manifolds. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 559–570. Academic Press, 1983.

[60] Wegkamp, M. and Zhao, Y. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 2016.

[61] Xie, F. Spectral norm posterior contraction in Bayesian sparse spiked covariance matrix model. *Electronic Journal of Statistics*, 18(2):5198 – 5257, 2024.

[62] Xie, F., Cape, J., Priebe, C. E., and Xu, Y. Bayesian Sparse Spiked Covariance Model with a Continuous Matrix Shrinkage Prior. *Bayesian Analysis*, 17(4):1193 – 1217, 2022.

[63] Yoshida, R. and West, M. Bayesian learning in sparse graphical factor models via annealed entropy. *Journal of Machine Learning Research*, 11:1771–1798, 2010.

[64] Yuchi, H. S., Mak, S., and Xie, Y. Bayesian Uncertainty Quantification for Low-Rank Matrix Completion. *Bayesian Analysis*, 18(2):491 – 518, 2023.

[65] Zhang, J., Cao, J., and Wang, L. Robust bayesian functional principal component analysis, 2023. URL https://arxiv.org/abs/2307.09731.