Bayesian Uncertainty Quantification for Low-Rank Matrix Completion

Henry Shaowu Yuchi*‡ Simon Mak†‡ Yao Xie*

February 2021

Abstract

We consider the problem of uncertainty quantification for an unknown low-rank matrix $\mathbf{X}$, given a partial and noisy observation of its entries. This quantification of uncertainty is essential for many real-world problems, including image processing, satellite imaging, and seismology, providing a principled framework for validating scientific conclusions and guiding decision-making. However, existing literature has mainly focused on the completion (i.e., point estimation) of the matrix $\mathbf{X}$, with little work on investigating its uncertainty. To this end, we propose in this work a new Bayesian modeling framework, called BayeSMG, which parametrizes the unknown $\mathbf{X}$ via its underlying row and column subspaces. This Bayesian subspace parametrization enables efficient posterior inference on matrix subspaces, which represents interpretable phenomena in many applications. This can then be leveraged for improved matrix recovery. We demonstrate the effectiveness of BayeSMG over existing Bayesian matrix recovery methods in numerical experiments, image inpainting, and a seismic sensor network application.

Keywords: hierarchical modeling, manifold sampling, matrix factorization, matrix completion, seismic imaging, uncertainty quantification

*H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology

†Department of Statistical Science, Duke University

‡Joint first authors.

1 Introduction

Low-rank matrices play a vital role in modeling many scientific and engineering problems, including (but not limited to) image processing, satellite imaging, and network analysis. In such applications, however, only a small portion of the desired matrix (which we denote as $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ in this article) can be observed. The reasons for this are two-fold. (i) The cost of observing all matrix entries can be high, requiring expensive computational, experimental, or communication expenditure; (ii) There can be missing observations at individual entries due to sensor malfunction, experimental failure, or unreliable data transmission. The *matrix completion* problem aims to complete the missing entries of $\mathbf{X}$ from a partial (and often-times noisy) observation. Matrix completion has attracted much attention since the seminal works of Candès and Tao (2010), Candès and Recht (2009), and Recht (2011). The theory and methodology behind point estimation are now well-understood for matrix completion, under the assumption that $\mathbf{X}$ is low-rank, with various convex and non-convex optimization algorithms developed for performing this recovery.

However, much of the literature (a detailed review is in Section 1.1) has focused on the completion, i.e., *point estimation*, of $\mathbf{X}$, with little work on exploring the uncertainty of such estimates. In many scientific and engineering applications, such estimates are much more useful when coupled with a measure of uncertainty. The principled characterization (and reduction) of this uncertainty is known as *uncertainty quantification* (UQ), see, e.g., Smith (2013). UQ is becoming increasingly important in various applications, providing a principled framework for validating scientific conclusions and guiding decision-making.

In this paper, we address the problem of UQ for the matrix completion problem from a Bayesian perspective. We propose a novel Bayesian modeling framework, called BayeSMG, which quantifies uncertainty in the desired matrix $\mathbf{X}$ via posterior sampling on its underlying subspaces. BayeSMG can be viewed as a hierarchical Bayesian extension of the singular matrix-variate Gaussian (SMG) distribution (see Gupta and Nagar, 1999; Mak and Xie, 2018), with hierarchical priors on matrix subspaces. In addition to providing point estimates on $\mathbf{X}$, the proposed model also yields UQ on $\mathbf{X}$ via an efficient Gibbs sampler on matrix subspaces. By integrating this subspace structure for posterior inference, we show that BayeSMG enjoys improved recovery performance and better interpretability compared with existing Bayesian models in extensive numerical experiments and a real-world seismic sensor network application.

1.1 Existing literature

Much of the existing literature on inferring $\mathbf{X}$ from partial observations falls under the topic of *matrix completion* - the completion (or point estimation) of $\mathbf{X}$ from observed entries. Early works in this area include the seminal works of Candès and Tao (2010), Candès and Recht (2009), and Recht (2011), which established conditions for exact completion via nuclear-norm minimization, under the assumption that observations are uniformly sampled without noise. This is then extended to the *noisy* matrix completion setting, where entries are observed with noise; important results include Candès and Plan (2010), Keshavan et al. (2010), Koltchinskii et al. (2011), and Negahban and Wainwright (2012), among others. There is now a rich body of work on matrix completion; recent overviews include Davenport and Romberg (2016) and Chi et al. (2019). However, completion focuses solely on the point estimation of matrix entries and does not provide uncertainty quantification on those unobserved. In scenarios where only a few entries are observed(see motivating applications), this uncertainty can be as valuable as point estimates in assessing the quality of the recovered matrix.

The current research literature has generally focused on point estimation of the unknown matrix $\mathbf{X}$. The problem of quantifying uncertainties in $\mathbf{X}$ has been relatively unexplored, but it is nonetheless an important one given the motivating applications. One recent pioneering work on this is Chen et al. (2019), which proposed entrywise confidence intervals for both convex and non-convex estimators on $\mathbf{X}$, via debiasing using low-rank factors of the matrix. The resulting debiased estimators admit nearly precise nonasymptotic distributional characterizations, which in turn enable optimal construction of confidence intervals for missing matrix entries and low-rank factors. Our approach has several distinctions from this work. First, the latter is a frequentist approach with appealing theoretical guarantees, whereas our approach is Bayesian and yields a richer quantification of uncertainty on $\mathbf{X}$ via a hierarchical Bayesian model. Second, to derive elegant theoretical results, the latter requires a sample size complexity condition on $\mathbf{X}$, similar to the minimum sample size condition in standard matrix completion analysis (see, e.g., Candès and Recht, 2009). Our UQ approach, in contrast, is applicable for any sample size n on $\mathbf{X}$, particularly for the "small-n" setting where observations are limited and uncertainty quantification is most needed.

Another approach for quantifying uncertainty is via Bayesian modeling. There is a growing literature on Bayesian matrix completion, of which the most popular approach is the Bayesian Probabilistic Matrix Factorization (BPMF) method

in Salakhutdinov and Mnih (2008). BPMF adopts the following probabilistic model on $\mathbf{X}$: $\mathbf{X} = \mathbf{M}\mathbf{N}^T$, $\mathbf{M} \in \mathbb{R}^{m_1 \times R}$, $\mathbf{N} \in \mathbb{R}^{m_2 \times R}$, where $R < m_1 \wedge m_2 := \min(m_1, m_2)$ is an upper bound on matrix rank. Each row of the factorized matrices $\mathbf{M}$ and $\mathbf{N}$ are then assigned i.i.d. Gaussian priors $\mathcal{N}(\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$ and $\mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, respectively. Conjugate normal hyperpriors are then assigned on the row and column means $\boldsymbol{\mu}_M \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_M/\beta)$, $\boldsymbol{\mu}_N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_N/\beta)$, with Inverse-Wishart hyperpriors on row and column covariance matrices $\boldsymbol{\Sigma}_M \sim \text{IW}(R, \mathbf{W}), \boldsymbol{\Sigma}_N \sim \text{IW}(R, \mathbf{W})$. The hyperparameters β and $\mathbf{W}$ are typically specified to provide weakly- or non-informative priors. This model allows for an efficient Gibbs sampler, which performs conjugate sampling on each *row* of $\mathbf{M}$ and each *row* of $\mathbf{N}$, along with conjugate updates on the mean vectors $(\boldsymbol{\mu}_M, \boldsymbol{\mu}_N)$ and covariance matrices $(\boldsymbol{\Sigma}_M, \boldsymbol{\Sigma}_N)$. With this, the BPMF can be shown to tackle problems as large as the Netflix dataset, with millions of user-movie ratings. Many existing Bayesian matrix completion methods (e.g., Lawrence and Urtasun, 2009; Zhou et al., 2010; Babacan et al., 2011; Alquier et al., 2014) can be viewed as variations or extensions of this BPMF framework.

From a modeling perspective, the key novelty in BayeSMG model is that it requires orthonormality in the factorized matrices, whereas the BPMF does not. Such a factorization can be viewed as parametrizing $\mathbf{X}$ via its singular value decomposition (SVD). This yields several advantages for our method, which we demonstrate later. First, by explicitly parametrizing row and column subspaces as model parameters, BayeSMG can incorporate prior knowledge on subspaces within the prior specification of such parameters. This prior information is often available in many signal processing and image processing problems, e.g., known signal structure or image features. Second, BayeSMG allows for *direct* inference on subspaces of $\mathbf{X}$ via posterior sampling, which is of direct interest in many problems, e.g., in sensor network localization (Zhang et al., 2020; an application we tackle later on) and topology identification problems (Eriksson et al., 2012). For subspace inference, our approach avoids performing an additional SVD step for every posterior sample (compared to the BPMF), which significantly speeds up inference for high-dimensional problems. Finally and perhaps most importantly, BayeSMG can leverage this posterior learning on subspaces to provide improved inference on $\mathbf{X}$. Compared to the BPMF, our approach can yield faster posterior contraction for unobserved entries when the underlying matrix has a low-rank structure, in both numerical simulations and applications. It enables a more accurate estimate and more precise uncertainty quantification of $\mathbf{X}$ over the BPMF.

The BayeSMG model also provides several novel theoretical insights. In Section 4, we show that the maximum a posteriori (MAP) estimator takes the form of a regularized matrix estimator, which provides a connection between the proposed method and existing matrix completion techniques. We also show that the BayeSMG model provides a probabilistic model on matrix coherence (Candès and Recht, 2009). Coherence has been widely used in the matrix completion literature as a theoretical condition for recovery, which measures the "recoverability" of a low-rank matrix. Through this, we then establish an error monotonicity result for BayeSMG, which provides a reassuring check on the UQ performance of the proposed model.

The paper is organized as follows. Section 2 introduces the BayeSMG model. Section 3 presents an efficient posterior sampling algorithm for $\mathbf{X}$ via manifold sampling on its subspaces. Section 4 reveals connections between the BayeSMG model and coherence, and its impact on error convergence. Section 5 investigates numerical experiments with synthetic and image data. Section 6 explores a real-world seismic sensor network application. Section 7 concludes with discussions.

2 The SMG model

We first describe the Singular Matrix-variate Gaussian (SMG) distribution, and how it can be utilized for modeling matrix subspaces.

2.1 Problem set-up

Let $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ be the matrix of interest, and assume $\mathbf{X}$ is low-rank, i.e., $R := \text{rank}(\mathbf{X}) \ll m_1 \wedge m_2$. Let $[m] := \{1, \cdots, m\}$. Suppose $\mathbf{X}$ is sampled with noise at an index set $\Omega \subseteq [m_1] \times [m_2]$ of size $|\Omega| = n$, yielding observations:

$$Y_{i,j} = X_{i,j} + \epsilon_{i,j}, \quad (i,j) \in \Omega. \tag{1}$$

Here, $Y_{i,j}$ is the observation at entry indexed by (i,j), corrupted by noise $\epsilon_{i,j}$. In this work, we assume $\epsilon_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \eta^2)$, i.e., the noise on each entry follows an i.i.d. Gaussian distribution with zero mean and variance η^2. Furthermore, let $\mathbf{Y}_\Omega := (Y_{i,j})_{(i,j)\in\Omega} \in \mathbb{R}^n$ denote the vector of noisy observations, and let $\mathbf{X}_{\Omega^c}$ be the vector of unobserved matrix entries, where $\Omega^c := ([m_1] \times [m_2]) \setminus \Omega$ is the set of unobserved indices.

With this framework, the desired goal of uncertainty quantification (UQ)

can be made more concrete. Given noisy observations $\mathbf{Y}_\Omega$, we wish to not only estimate the unobserved matrix entries $\mathbf{X}_{\Omega^c}$, but also quantify a notion of *uncertainty* on both observed or unobserved entries (since observation noise is present).

2.2 SMG model

We adopt the following SMG model for the low-rank matrix $\mathbf{X}$, which we assume to be normalized with a zero mean.

Definition 1 (SMG model, Definition 2.4.1 of Gupta and Nagar, 1999)**.** *Let $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix with entries $Z_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $(i,j) \in [m_1] \times [m_2]$. The random matrix $\mathbf{X}$ has a* singular matrix-variate Gaussian (SMG) *distribution if $\mathbf{X} \overset{d}{=} \mathcal{P}_\mathcal{U} \mathbf{Z} \mathcal{P}_\mathcal{V}$ for some choice of projection matrices $\mathcal{P}_\mathcal{U} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{P}_\mathcal{V} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m_1 \times R}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V} \in \mathbb{R}^{m_2 \times R}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $R < m_1 \wedge m_2$. We will denote this as $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, R)$.*

In other words, a realization from the SMG distribution can be obtained by first (i) simulating a matrix $\mathbf{Z}$ from a Gaussian ensemble with variance σ^2, i.e., a matrix with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then (ii) performing a left and right projection of $\mathbf{Z}$ using the projection matrices $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$. Recall that the projection operator $\mathcal{P}_\mathcal{U} = \mathbf{U}\mathbf{U}^T \in \mathbb{R}^{m_1 \times m_1}$ maps a vector in $\mathbb{R}^{m_1}$ to its orthogonal projection on the R-dimensional subspace $\mathcal{U}$ spanned by the columns of $\mathbf{U}$. By performing this projection, the resulting matrix $\mathbf{X} = \mathcal{P}_\mathcal{U}\mathbf{Z}\mathcal{P}_\mathcal{V}$ can be shown to be of rank $R < m_1 \wedge m_2$, with its row and column spaces $\mathcal{U}$ and $\mathcal{V}$ corresponding to the subspaces for $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$. The matrix $\mathbf{X}$ also lies in the space $\mathcal{T} := \bigcup_{u_k \in \mathcal{U}, v_k \in \mathcal{V}} \text{span}(\{\mathbf{u}_k \mathbf{v}_k^T\}_{k=1}^R)$. With a small choice of R, this provides a flexible probabilistic model for the low-rank matrix $\mathbf{X}$.

The SMG distribution provides several appealing properties for modeling low-rank matrices. First, it provides a prior modeling framework on the matrix $\mathbf{X}$ involving its row and column subspaces $\mathcal{U}$ and $\mathcal{V}$. It is known from Chikuse (2012) that, for each projection operator $\mathcal{P} \in \mathbb{R}^{m \times m}$ of rank R, there exists a unique R-dimensional hyperplane (or an R-plane) in $\mathbb{R}^m$ containing the origin which corresponds to the image of such a projection. It connects the space of rank R projection matrices and the *Grassmann manifold* $\mathcal{G}_{R,m-R}$, the space of R-planes in $\mathbb{R}^m$. Viewed this way, the projection matrices parametrizing $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, R)$ encode useful information on the row and column spaces of $\mathbf{X}$. Second, since the projection of a Gaussian random vector is still

Gaussian, the left-right projection of the Gaussian ensemble $\mathbf{Z}$ results in each entry of $\mathbf{X}$ being Gaussian-distributed as well. It is useful for deriving a UQ property of the BayeSMG model.

We now show several distributional properties of the SMG model:

Lemma 2 (Distributional properties of SMG)**.** *Let* $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, R)$, *with* $\mathcal{P}_\mathcal{U} \in \mathbb{R}^{m_1 \times m_1}$, $\mathcal{P}_\mathcal{V} \in \mathbb{R}^{m_2 \times m_2}$, $\sigma^2 > 0$ *and* $R < m_1 \wedge m_2$ *known. Then:*

(a) The density of $\mathbf{X}$ *is given by*

$$p(\mathbf{X}) = (2\pi\sigma^2)^{-R^2/2} \mathrm{etr}\left\{-\frac{1}{2\sigma^2}[(\mathbf{X}\mathcal{P}_\mathcal{V})^T(\mathcal{P}_\mathcal{U}\mathbf{X})]\right\}, \quad \mathbf{X} \in \mathcal{T}, \tag{2}$$

where $\mathrm{etr}(\cdot) := \exp\{\mathrm{tr}(\cdot)\}$.

(b) Consider the block decomposition of $\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U}$*:*

$$\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U} = \begin{pmatrix} (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_\Omega & (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega,\Omega^c} \\ (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})^T_{\Omega,\Omega^c} & (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega^c} \end{pmatrix}. \tag{3}$$

Conditional on the observed noisy entries $\mathbf{Y}_\Omega$*, the unobserved entries* $\mathbf{X}_{\Omega^c}$ *follow the distribution,* $[\mathbf{X}_{\Omega^c}|\mathbf{Y}_\Omega] \sim \mathcal{N}(\mathbf{X}^P_{\Omega^c}, \mathbf{\Sigma}^P_{\Omega^c})$*. Here,* $\gamma^2 = \eta^2/\sigma^2$*, and*

$$\begin{aligned} \mathbf{R}_N(\Omega) &:= (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_\Omega \in \mathbb{R}^{N \times N}, \\ \mathbf{X}^P_{\Omega^c} &:= (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})^T_{\Omega,\Omega^c}[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}]^{-1}\mathbf{Y}_\Omega, \\ \mathbf{\Sigma}^P_{\Omega^c} &:= \sigma^2\{(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega^c} - (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})^T_{\Omega,\Omega^c}[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}]^{-1}(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})^T_{\Omega,\Omega^c}\}. \end{aligned} \tag{4}$$

(c) Conditional on the observed noisy entries $\mathbf{Y}_\Omega$*, the observed entries* $\mathbf{X}_\Omega$ *follow the distribution* $[\mathbf{X}_\Omega|\mathbf{Y}_\Omega] \sim \mathcal{N}(\mathbf{X}^P_\Omega, \mathbf{\Sigma}^P_\Omega)$*, where* $\otimes$ *is the Kronecker product, and*

$$\begin{aligned} \mathbf{X}^P_\Omega &:= (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_\Omega[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}]^{-1}\mathbf{Y}_\Omega, \\ \mathbf{\Sigma}^P_\Omega &:= \sigma^2\{(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_\Omega - (\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})^T_\Omega[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}]^{-1}(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_\Omega\}. \end{aligned} \tag{5}$$

Remark: Lemma 2 reveals two key properties of the SMG model. First, *prior* to observing data, part (a) shows that the low-rank matrix $\mathbf{X}$ lies on the space $\mathcal{T}$, and follows a degenerate multivariate Gaussian distribution with mean zero and

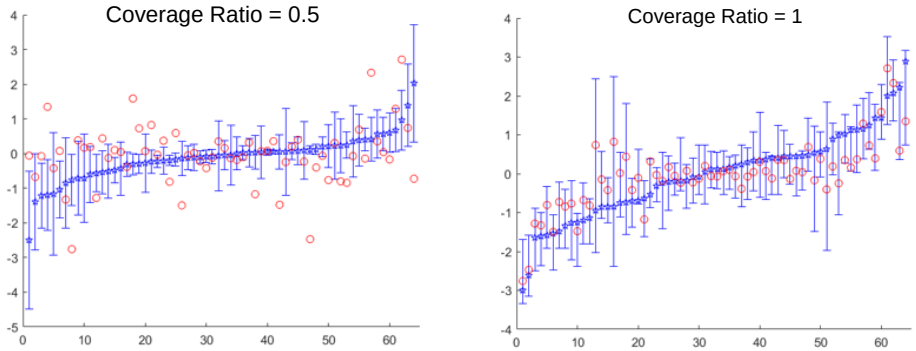
covariance matrix $\sigma^2(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})$. Second, *after* observing the noisy entries $\mathbf{Y}_\Omega$, part (b) shows that the conditional distribution of $\mathbf{X}_{\Omega^c}$ (the unobserved entries in $\mathbf{X}$) given $\mathbf{Y}_\Omega$ is still multivariate Gaussian, with closed-form expressions for its mean vector $\mathbf{X}^P_{\Omega^c}$ and covariance matrix $\mathbf{\Sigma}^P_{\Omega^c}$ in (4).

2.3 Can we directly use the SMG model for UQ?

Lemma 2 provides a closed-form posterior distribution for the low-rank matrix $\mathbf{X}$ *after* observing the noisy observations $\mathbf{Y}_\Omega$. It provides a potential way for computing confidence intervals on each entry in $\mathbf{X}$, assuming the underlying row and column subspaces $\mathcal{U}$ and $\mathcal{V}$ are known. Of course, in practice, such subspaces are never known with certainty. One solution might be to plug in point estimates of $\mathcal{U}$ and $\mathcal{V}$ (estimated from data) within the predictive equations in Lemma 2. We investigate the efficacy of this plug-in approach via a simple numerical example.

The simulation set-up is as follows. Let $m = m_1 = m_2 = 8$ be the row and column dimensions of the matrix, and let $R = 2$ be its rank. We first simulate two random orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ of size $m \times R$, via a truncated SVD on an $m \times m$ matrix with *i.i.d.* $U[0, 1]$ entries. With $\mathcal{P}_\mathcal{U} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{P}_\mathcal{V} = \mathbf{V}\mathbf{V}^T$, the "true" low-rank matrix is then simulated from the SMG model $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2 = 1, R = 2)$. Finally, noisy observations are sampled via (1) with noise variance $\eta^2 = 0.5^2$. In total, 36 entries are observed (56.25% of total entries), with such entries chosen uniformly at random. From this, we can obtain point estimates of the subspaces $\mathcal{U}$ and $\mathcal{V}$, by first estimating $\mathbf{X}$ via nuclear norm minimization (Candès and Plan, 2010), a popular method for matrix completion, and then taking the row and column subspaces for this matrix estimate via SVD. These subspace estimates are then plugged into the expressions in Lemma 2 for UQ.

Figure 1(a) plots the point estimates and 95% plug-in confidence intervals (CIs) for each matrix entry using Lemma 2, with its corresponding true value marked in red. We see that these intervals provide poor coverage performance since many of the true matrix entries are not within these intervals. Indeed, the coverage ratio for these plug-in CIs is only 56%, meaning only around half of the confidence intervals cover the true entries. This poor coverage suggests that this CI approach (with plug-in subspace estimates) can significantly underestimate the underlying uncertainty of point estimates, which is unsurprising since uncertainty for subspace estimation is not incorporated when using Lemma 2. Figure



(a) Coverage ratio using the CIs constructed via Lemma 2 with plug-in subspace estimates.

(b) Coverage ratio using the posterior predictive intervals from the proposed BayeSMG model..

Figure 1: *Plotted are the point estimates (blue points) and 95% Confidence Intervals (blue intervals) for each matrix entry, ordered by increasing point estimates. Red points mark the true matrix values.*

1(b) plots the point estimates and 95% posterior predictive intervals using the proposed BayeSMG method, which accounts for subspace uncertainty by assigning hierarchical priors on subspaces $\mathcal{U}$ and $\mathcal{V}$ from the SMG model. We see that the proposed BayeSMG approach yields much better coverage: the 95% intervals, which are now slightly wider, cover the true matrix entries well. The resulting coverage ratio is now at 100%, which is slightly higher than the nominal coverage rate of 95%, but much closer to this rate than the earlier plug-in approach. It shows the proposed BayeSMG method can provide better uncertainty quantification of $\mathbf{X}$ via a fully-Bayesian model specification on matrix subspaces.

3 The BayeSMG model

3.1 Model specification

We now present the hierarchical specification for the proposed Bayesian SMG model, or BayeSMG for short. We begin by first introducing the matrix von Mises-Fisher (vMF) distribution, which will serve as prior models for the row and column orthonormal frames $\mathbf{U}$ and $\mathbf{V}$. We then present a Gibbs sampling algorithm that makes use of a reparameterization of the SMG model for efficient posterior sampling.

The matrix von Mises-Fisher distribution (Khatri and Mardia, 1977; Mardia and Jupp, 2009) provides a useful class of distributions on the row and column frames, which lie on a so-called Stiefel manifold. A Stiefel manifold (Chikuse, 2012) consists of all orthonormal subspaces of rank R in the space of $\mathbb{R}^m$; this is denoted as $\mathcal{V}_{R,m}$ hereafter. The matrix vMF distribution assumes the following probability density function of matrix $\mathbf{W}$ on $\mathcal{V}_{R,m}$:

$$p(\mathbf{W}; m, R, \mathbf{F}) = \left[{}_0F_1\left(; \frac{m}{2}; \frac{\mathbf{F}^T\mathbf{F}}{4}\right)\right]^{-1} \operatorname{etr}(\mathbf{F}^T\mathbf{W}), \quad \mathbf{W} \in \mathcal{V}_{R,m}, \tag{6}$$

where ${}_0F_1(;\cdot;\cdot)$ is the hypergeometric function, and $\mathbf{F} \in \mathbb{R}^{m\times R}$ is the concentration matrix. We denote this distribution by $\mathbf{W} \sim \mathcal{MF}(m, R, \mathbf{F})$. The matrix vMF distribution provides conditionally conjugate priors for a wide range of multivariate models, including for cluster analysis (Gopal and Yang, 2014) and factor models (Hoff, 2013). One appeal of this class of distribution is that it can be efficiently sampled. Hoff (2009) proposed a rejection sampling algorithm that sequentially samples each column of the matrix $\mathbf{W}$. Recently, Jauch et al. (2020) presented a general simulation framework on the Stiefel manifolds using polar expansions; using such an expansion with Hamiltonian Monte Carlo (Girolami and Calderhead, 2011) provides a better sampling efficiency over competing MCMC methods by an order of magnitude. We will leverage this useful family of priors via the following reparametrization of the BayeSMG model.

The following proposition gives a nice reformulation of the SMG model under uniform subspace priors on $\mathcal{U}$ and $\mathcal{V}$:

Proposition 3 (SVD of BayeSMG)**.** *Suppose* $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, R)$*, with independent uniform priors* $\mathcal{P}_\mathcal{U} \sim U(\mathcal{G}_{R,m_1-R})$*,* $\mathcal{P}_\mathcal{V} \sim U(\mathcal{G}_{R,m_2-R})$*, and fixed* σ^2 *and* R*. Let* $\mathbf{X} = \mathbf{UDV}^T$ *be the SVD of* $\mathbf{X}$*, with singular values* $\operatorname{diag}(\mathbf{D}) = (d_k)_{k=1}^R$ *not necessarily in decreasing order. Then:*

1. *The singular vectors* $\mathbf{U}$ *and* $\mathbf{V}$ *follow independent priors* $\mathcal{MF}(m_1, R, \mathbf{0})$ *and* $\mathcal{MF}(m_2, R, \mathbf{0})$*, respectively.*

2. *The singular values* $\operatorname{diag}(\mathbf{D}) = (d_k)_{k=1}^R$ *follow the repulsed normal distribution, with density:*

$$\frac{1}{Z_R(2\pi\sigma^2)^{R/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^R d_k^2\right\} \prod_{\substack{k,l=1\\k<l}}^R |d_k^2 - d_l^2|,\ d_k > 0,\ k = 1, \cdots, R. \tag{7}$$

The proof of this proposition is provided in the supplementary section. The first part of the proposition shows that the use of uniform priors on the projection matrices $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ corresponds to independent $\mathcal{MF}(m_1, R, \mathbf{0})$ and $\mathcal{MF}(m_2, R, \mathbf{0})$ priors for the singular vectors $\mathbf{U}$ and $\mathbf{V}$, which are uniform priors on the Stiefel manifolds $\mathcal{V}_{R,m_1}$ and $\mathcal{V}_{R,m_2}$, respectively. The second part shows that the singular values in $\mathbf{D}$ follow the repulsed normal distribution, which is closely connected with the distribution of singular values for a Gaussian ensemble (Shen, 2001).

This proposition then motivates the following reparametrization of the BayeSMG model:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \;\; \mathbf{U} \sim \mathcal{MF}(m_1, R, \mathbf{F}_1), \;\; \mathbf{V} \sim \mathcal{MF}(m_2, R, \mathbf{F}_2), \;\; \text{diag}(\mathbf{D}) \sim \mathcal{RN}(\mathbf{0}, \sigma^2), \tag{8}$$

where $\mathcal{RN}(\mathbf{0}, \sigma^2)$ is the repulsed normal distribution in (7), and the priors on $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}$ are independently specified. When little is known a priori on matrix subspaces, one can set the concentration matrices as $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{0}$, which provides non-informative priors on $\mathbf{U}$ and $\mathbf{V}$. In problems where some prior information is available on matrix subspaces, one can elicit a good choice of prior parameters for the vMF priors via a moment matching approach (Wang and Zhou, 2009). We show in the next section that this reparametrization allows for a Gibbs sampling algorithm which makes use of conditionally conjugate priors for efficient posterior sampling.

Finally, we complete the Bayesian specification by assigning the following priors on the variance parameters σ^2 and η^2:

$$[\sigma^2] \sim IG(\alpha_{\sigma^2}, \beta_{\sigma^2}), \quad [\eta^2] \sim IG(\alpha_{\eta^2}, \beta_{\eta^2}), \tag{9}$$

where $IG(\alpha, \beta)$ is the Inverse-Gamma distribution with shape and rate parameters α and β. Table 1 summarizes the full Bayesian model specification for BayeSMG.

3.2 Posterior sampling

Using the reparametrized model (8), we now present a subspace Gibbs sampler for posterior sampling on the BayeSMG model, specifically on the parameters $\Theta = \{\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2\}$ given partial and noisy observations $\mathbf{Y}_\Omega$. We first introduce the sampler under *complete* observation of the noisy matrix $\mathbf{Y}$, then describe a data imputation procedure for posterior sampling under *partial* observations

Model	*Distribution*
Observations	$[\mathbf{Y}_\Omega \vert \mathbf{X}, \eta^2]$: $Y_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(X_{i,j}, \eta^2)$
Low-rank matrix	$[\mathbf{X} \vert \mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2] : \mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, R)$
(equivalently)	$[\mathbf{X} \vert \mathbf{U}, \mathbf{V}, \sigma^2] : \mathbf{X} = \mathbf{UDV}^T$, $\text{diag}\{\mathbf{D}\} \sim \mathcal{RN}(\mathbf{0}, \sigma^2)$
Priors	$[\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{V}, \sigma^2, \eta^2] = [\mathcal{P}_\mathcal{U}]\ [\mathcal{P}_\mathcal{V}]\ [\eta^2][\sigma^2]$
Matrix subspaces	$[\mathcal{P}_\mathcal{U}] \sim U(\mathcal{G}_{R,m_1-R})$
	$[\mathcal{P}_\mathcal{V}] \sim U(\mathcal{G}_{R,m_2-R})$
Matrix variance	$[\sigma^2] \sim IG(\alpha_{\sigma^2}, \beta_{\sigma^2})$
Noise variance	$[\eta^2] \sim IG(\alpha_{\eta^2}, \beta_{\eta^2})$

Table 1: *Model specification for BayeSMG.*

$\mathbf{Y}_\Omega$.

Consider first the setting where complete observations on $\mathbf{Y}$ are obtained. It can then be shown (see supplementary material for a full derivation) that the full conditional distributions of $\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$ and σ^2 take the form:

$$
\begin{aligned}
&[\mathbf{U} \vert \mathbf{D}, \mathbf{V}, \mathbf{Y}, \sigma^2, \eta^2] \sim \mathcal{MF}(m_1, R, \mathbf{YVD}/\eta^2 + \mathbf{F}_1),\\
&[\mathbf{V} \vert \mathbf{D}, \mathbf{U}, \mathbf{Y}, \sigma^2, \eta^2] \sim \mathcal{MF}(m_2, R, \mathbf{Y}^T\mathbf{UD}/\eta^2 + \mathbf{F}_2),\\
&[\mathbf{D} \vert \mathbf{U}, \mathbf{V}, \mathbf{Y}, \sigma^2, \eta^2] \sim \mathcal{RN}\left(\sigma^2 \text{diag}(\mathbf{U}^T\mathbf{YV})/(\eta^2+\sigma^2), \eta^2\sigma^2/(\eta^2+\sigma^2)\right), \quad (10)\\
&[\sigma^2 \vert \mathbf{U}, \mathbf{D}, \mathbf{V}, \mathbf{Y}, \eta^2] \sim IG(\alpha_{\sigma^2} + R/2, \beta_{\sigma^2} + \text{tr}(\mathbf{D}^2)/2),\\
&[\eta^2 \vert \mathbf{U}, \mathbf{D}, \mathbf{V}, \mathbf{Y}, \sigma^2] \sim IG(\alpha_{\eta^2} + m_1 m_2/2, \beta_{\eta^2} + \|\mathbf{Y} - \mathbf{UDV}^T\|_F^2/2).
\end{aligned}
$$

One can then perform the above full conditional updates cyclically for posterior sampling on $[\Theta \vert \mathbf{Y}]$ via Gibbs sampling. As mentioned previously, there are efficient sampling algorithms for the matrix vMF distribution (Hoff, 2009; Jauch et al., 2020), which enable efficient full conditional sampling on $\mathbf{U}$ and $\mathbf{V}$. The full conditional distribution of $\mathbf{D}$ follows the aforementioned repulsed normal distribution with a location shift of $\boldsymbol{\mu}$ (denoted as $\mathcal{RN}(\boldsymbol{\mu}, \delta^2)$), with density:

$$
\frac{1}{Z_R(2\pi\delta^2)^{R/2}} \exp\left\{-\frac{1}{2\delta^2}\sum_{k=1}^{R}(d_k - \mu_k)^2\right\} \prod_{k,l=1;k<l}^{R} |d_k^2 - d_l^2|, \qquad (11)
$$

where $d_k > 0, k = 1, \cdots R$. We have found that this can be quite efficiently sampled via a Metropolis-Hastings sampler (Metropolis et al., 1953), with an "independent" proposal distribution (i.e., independent of the current state) set as a non-central, multivariate t-distribution with mean vector $\boldsymbol{\mu}$ and scale parameter δ.

Consider now the setting where only partial noisy observations $\mathbf{Y}_\Omega$ are

available. We describe a posterior sampling algorithm for $[\Theta|\mathbf{Y}_\Omega]$, which makes use of a modification on the above Gibbs sampler on $[\Theta|\mathbf{Y}]$. The idea is to first sample from the joint distribution $[\Theta, \mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega]$ of both the parameters Θ and unobserved noisy entries $\mathbf{Y}_{\Omega^c}$, then take only the marginal samples of parameters Θ. With an initialization of $\Theta = \Theta'$, the joint distribution $[\Theta, \mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega]$ can be sampled via the following Gibbs sampling steps:

(i) Draw one sample from $[\mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega, \Theta']$. Since the missing entries $\mathbf{Y}_{\Omega^c}$ is assumed to be conditionally independent of the observed entries $\mathbf{Y}_\Omega$ given $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, this is equivalent to sampling $[\mathbf{Y}_{\Omega^c}|\mathbf{X}]$, which amounts to simulating the observation noise in $\mathbf{Y}_{\Omega^c}$ given ground truth $\mathbf{X}_{\Omega^c}$.

(ii) Draw one sample Θ' from the posterior distribution $[\Theta|\mathbf{Y}_{\Omega^c}, \mathbf{Y}_\Omega] = [\Theta|\mathbf{Y}]$ via the Gibbs sampling steps in (10).

Step (i) can be viewed as a data imputation step, which imputes missing entries in the noisy matrix $\mathbf{Y}$. Step (ii) performs the earlier posterior sampling steps for parameters Θ given the full noisy matrix $\mathbf{Y}$.

It is worth noting that step (i) depends on an implicit assumption that the entries are either completely missing at random (CMAR) or missing at random (MAR); see Little and Rubin (2019) for further discussion on missing data modeling. When the entries are missing not at random (MNAR), the sampling of $[\mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega, \Theta']$ can become much complicated, since it would depend on the underlying MNAR mechanism for missing entries. One approach is to adopt a probabilistic model for the MNAR entries (see, e.g., Hernández-Lobato et al., 2014 for one such model), then sample $[\mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega, \Theta']$ given this model. There are, however, several limitations to this approach. (i) The conditional distribution $[\mathbf{Y}_{\Omega^c}|\mathbf{Y}_\Omega, \Theta']$ may be computationally expensive to sample from in the MNAR setting. And (ii) in the case of misspecification for the MNAR model, the resulting recovery of the matrix $\mathbf{X}$ can be highly biased and inaccurate. In the absence of prior information on how the entries are missing (which is the case in many applications), it may be preferable to adopt Algorithm 1 for posterior inference. We will show later (in Section 5.2) that the BayeSMG is empirically robust to mild violations of this implicit MAR assumption for missing entries.

Algorithm 1 summarizes the above steps for the posterior sampling algorithm. The algorithm is first initialized with estimates $\mathbf{U}_{[0]}$, $\mathbf{D}_{[0]}$, and $\mathbf{V}_{[0]}$ obtained from a nuclear-norm completion of $\mathbf{X}$ (Carson et al., 2012). Next, the missing noisy entries $\mathbf{Y}_{\Omega^c}$ are imputed using step (i), then a posterior draw is made using step (ii) via the Gibbs steps in (10). This is then iterated until a desired number

Algorithm 1 `BayeSMG`$(\mathbf{Y}_\Omega, R, \mathbf{F}_1, \mathbf{F}_2, \alpha_{\sigma^2}, \beta_{\sigma^2}, \alpha_{\eta^2}, \beta_{\eta^2})$: Gibbs sampler for BayeSMG

Initialization:

- Complete $\mathbf{X}_{[0]}$ from $\mathbf{Y}_\Omega$ via nuclear-norm minimization.
- Initialize $[\mathbf{U}_{[0]}, \mathbf{D}_{[0]}, \mathbf{V}_{[0]}] \leftarrow \text{svd}(\mathbf{X}_{[0]})$ and $\sigma^2_{[0]} > 0$.

Gibbs sampling: T - total samples

for $t = 1, \ldots, T$ **do**

- Set $\mathbf{X}_{[t]} \leftarrow \mathbf{U}_{[t-1]}\mathbf{D}_{[t-1]}\mathbf{V}^T_{[t-1]}$
- Impute missing entries $\mathbf{Y}_{\Omega^c}$ by sampling $Y_{i,j} \overset{i.i.d.}{\sim} X_{[t],i,j} + \mathcal{N}(0, \eta^2), \quad (i,j) \in \Omega^c$.
- Sample $\mathbf{U}_{[t]} \sim \mathcal{MF}(m_1, R, \mathbf{Y}\mathbf{V}_{[t-1]}\mathbf{D}_{[t-1]}/\eta^2_{[t-1]} + \mathbf{F}_1)$.
- Sample $\mathbf{V}_{[t]} \sim \mathcal{MF}(m_2, R, \mathbf{Y}^T\mathbf{U}_{[t]}\mathbf{D}_{[t-1]}/\eta^2_{[t-1]} + \mathbf{F}_2)$.
- Sample $\mathbf{D}_{[t]} \sim \mathcal{RN}\left(\frac{\sigma^2_{[t-1]}\text{diag}(\mathbf{U}^T_{[t]}\mathbf{Y}\mathbf{V}_{[t]})}{(\eta^2_{[t-1]}+\sigma^2_{[t-1]})}, \frac{\eta^2_{[t-1]}\sigma^2_{[t-1]}}{(\eta^2_{[t-1]}+\sigma^2_{[t-1]})}\right)$.
- Sample $\sigma^2_{[t]} \sim IG(\alpha_{\sigma^2} + R/2, \beta_{\sigma^2} + \text{tr}(\mathbf{D}^2_{[t]})/2)$.
- Sample $\eta^2_{[t]} \sim IG(\alpha_{\eta^2} + m_1 m_2/2, \beta_{\eta^2} + \|\mathbf{Y} - \mathbf{U}_{[t]}\mathbf{D}_{[t]}\mathbf{V}^T_{[t]}\|^2_F/2)$.

Output: Return posterior samples $\{(\mathbf{X}_{[t]}, \mathbf{U}_{[t]}, \mathbf{D}_{[t]}, \mathbf{V}_{[t]}, \sigma^2_{[t]}, \eta^2_{[t]})\}_{t=1}^T$.

of posterior samples is obtained. Using the posterior samples of $(\mathbf{U}_{[t]}, \mathbf{D}_{[t]}, \mathbf{V}_{[t]})$ at each iteration t, we can obtain a sample $\mathbf{X}_{[t]} = \mathbf{U}_{[t]}\mathbf{D}_{[t]}\mathbf{V}^T_{[t]}$ from the desired posterior distribution $[\mathbf{X}|\mathbf{Y}_\Omega]$. These posterior samples $\{\mathbf{X}_{[t]}\}_{t=1}^T$ can then be used for the target goal of uncertainty quantification: the mean of such samples provides a point estimate $\hat{\mathbf{X}}$ for the recovered matrix, and its variability around $\hat{\mathbf{X}}$ provides a measure of uncertainty for this recovery.

While the computational complexity of this algorithm is difficult to establish given the complex manifold sampling steps, we found this posterior sampler to be quite efficient and scalable in practice. For a relatively large 256×256 matrix, the sampler takes around 1 minute to generate $T = 1000$ samples on a standard laptop computer (Intel i7 CPU and 16GB RAM), which is quite efficient given the size of the matrix. We will report computation times for larger matrices in the numerical studies later.

3.3 Inference on matrix rank

The BayeSMG model as presented above assumes the rank of the matrix $\mathbf{X}$ is known, which is often not the case in practice. There has been some literature on this problem of rank estimation for matrix inference. Shapiro et al. (2018) investigates a lower bound of the matrix rank needed for the matrix completion problem to be stable. Hoff (2007) proposes a Bayesian dimension selection method that models the dimension of matrix subspaces via a singular value decomposition (SVD). This allows for a Gibbs sampler which samples the joint posterior distribution of the singular vectors, singular values, and rank. For the BayeSMG setting, one can employ a similar fully Bayesian approach to quantify uncertainty on the matrix rank, by assigning an appropriate prior on R and adopting a similar Gibbs sampler as Hoff (2007). However, in the numerical experiments later (which are quite high-dimensional, with m_1 and m_2 on the order of thousands), we found such an approach to be computationally expensive since Algorithm 1 needs to be performed for each choice of rank R. In this high-dimensional setting, we favor the following maximum a posteriori (MAP) approach for rank inference, which sacrifices a richer quantification of uncertainty for computational efficiency.

Consider the MAP estimate of the unknown matrix $\mathbf{X}$, which can formulated as:

$$\tilde{\mathbf{X}} = \underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\operatorname{argmax}} \; [\mathbf{Y}_\Omega|\mathbf{X}][\mathbf{X}|R][R]. \tag{12}$$

Here, $[\mathbf{X}|R]$ follows the BayeSMG prior specification (8) given rank R, and $[R]$ is a prior distribution assigned on matrix rank. Under uniform subspace priors and a flat prior on R over $\{1, \cdots, m_1 \wedge m_2\}$, it can be shown (see Section 4.1 for a full derivation) that the MAP $\tilde{\mathbf{X}}$ can be well-approximated by the nuclear-norm formulation:

$$\underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\operatorname{argmin}} \left(\sum_{(i,j)\in\Omega} (Y_{i,j} - X_{i,j})^2 + \lambda \|\mathbf{X}\|_* \right). \tag{13}$$

Here, $\|\mathbf{X}\|_*$ is the nuclear norm of $\mathbf{X}$ (the sum of its singular values, see Candès and Tao, 2010), and λ is a regularization parameter. The optimization problem (13) can be efficiently solved via convex optimization algorithms (see Section 1.1 for further details).

In practice, λ can be estimated via cross-validation (Friedman et al., 2017) on the observed entries $\mathbf{Y}_\Omega$. We first divide these entries into multiple folds. For each fold, we first use nuclear-norm minimization (13) to estimate the entries

of the particular fold. Then we compute the cross-validation error for these estimates. We then select the optimal tuning parameter λ^* such that it is the λ that minimizes the sum of these cross-validation errors for all folds.

With this estimate λ^*, an (approximate) MAP estimate $\tilde{\mathbf{X}}$ can be obtained by solving (13) with $\lambda = \lambda^*$. This in turn yields an approximate MAP estimate of R via the rank of the matrix estimate $\tilde{\mathbf{X}}$. Finally, this rank estimate can be plugged into Algorithm 1 for uncertainty quantification on matrix $\mathbf{X}$. For high-dimensional problems with either m_1 or m_2 large, this plug-in MAP approach for rank estimation can yield significant computational savings over a fully Bayesian treatment.

4 Theoretical insights

We now provide some theoretical insights on the BayeSMG model. We first discuss an interesting link between the maximum-a-posterior (MAP) estimator and regularized estimators in the literature, then present a connection between model uncertainty from the BayeSMG model and coherence, which is then used to prove an error monotonicity result on uncertainty quantification.

4.1 Connection to Regularized Estimators

The following lemma reveals a connection between the BayeSMG model and existing completion methods:

Lemma 4 (MAP estimator)**.** *Assume the BayeSMG model in Table 1, with* $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{0}$*,* η^2 *and* σ^2 *fixed, and a uniform prior on rank* R*. Conditional on* $\mathbf{Y}_\Omega$*, the MAP estimator for* $\mathbf{X}$ *becomes*

$$\underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\operatorname{argmin}} \left(\frac{\|\mathbf{Y}_\Omega - \mathbf{X}_\Omega\|_2^2}{\eta^2} + \log(2\pi\sigma^2)\operatorname{rank}^2(\mathbf{X}) + \frac{\|\mathbf{X}\|_F^2}{\sigma^2} \right), \tag{14}$$

where $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{i,j}^2}$ *is the Frobenius norm of* $\mathbf{X}$*.*

The MAP estimator $\tilde{\mathbf{X}}$ in (14) connects the proposed model with existing deterministic matrix completion methods (see Davenport and Romberg, 2016 and references therein). Consider the following approximation to the MAP formulation (14). Treating $\log(2\pi\sigma^2)\operatorname{rank}^2(\mathbf{X})$ as a Lagrange multiplier, and replace the rank function $\operatorname{rank}(\mathbf{X})$ by its nuclear norm $\|\mathbf{X}\|_*$ (its tightest convex

relaxation, see Keshavan et al., 2010), the optimization in (14) becomes:

$$\underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\operatorname{argmin}} \|\mathbf{Y}_\Omega - \mathbf{X}_\Omega\|_2^2 + \lambda\left\{\alpha\|\mathbf{X}\|_* + (1-\alpha)\|\mathbf{X}\|_F^2\right\}, \tag{15}$$

for some choice of $\lambda > 0$ and $\alpha \in (0, 1)$. Using (15) to approximate (14), we can then view the MAP estimator as an analogue of the *elastic net* estimator (Zou and Hastie, 2005) from linear regression for noisy matrix completion.

To see the connection between the MAP estimator $\tilde{\mathbf{X}}$ and existing matrix completion methods, set $\alpha = 1$ in (15). The problem then reduces to the nuclear-norm formulation in (13), which is widely used for deterministic matrix completion (Candès and Recht, 2009; Candès and Tao, 2010; Recht, 2011). This provides an intuitive connection between the proposed Bayesian model and existing completion methods, which we leveraged earlier for efficient inference on matrix rank.

4.2 Uncertainty and coherence

Consider next the following definition of subspace coherence from Candès and Recht (2009), ignoring scaling factors:

Definition 5 (Coherence, Definition 1.2 of Candès and Recht, 2009)**.** *Let $\mathcal{U} \in \mathcal{G}_{R,m-R}$ be an R-plane in $\mathbb{R}^m$, and let $\mathcal{P}_\mathcal{U}$ be the orthogonal projection onto $\mathcal{U}$. The* coherence *of subspace $\mathcal{U}$ with respect to the i-th basis vector, $\mathbf{e}_i$, is defined as $\mu_i(\mathcal{U}) := \|\mathcal{P}_\mathcal{U}\mathbf{e}_i\|_2^2$, and the* coherence *of $\mathcal{U}$ is defined as $\mu(\mathcal{U}) = \max_{i=1,\dots,m} \mu_i(\mathcal{U})$.*

In words, coherence measures how *correlated* a subspace $\mathcal{U}$ is with the basis vectors $\{\mathbf{e}_i\}_{i=1}^m$. A large $\mu_i(\mathcal{U})$ suggests that $\mathcal{U}$ is highly correlated with the i-th basis vector $\mathbf{e}_i$, in that the projection of $\mathbf{e}_i$ onto $\mathcal{U}$ preserves much of its original length; a small value of $\mu_i(\mathcal{U})$ suggests that $\mathcal{U}$ is nearly orthogonal with $\mathbf{e}_i$, so a projection of $\mathbf{e}_i$ onto $\mathcal{U}$ loses most of its length. Figure 2 visualizes these two cases using the projection of three basis vectors on a two-dimensional subspace $\mathcal{U}$. Note that the projection of the red vector onto $\mathcal{U}$ retains nearly unit length, so $\mathcal{U}$ has near-maximal coherence for this basis. The projection of the black vector onto $\mathcal{U}$ results in a considerable length reduction, so $\mathcal{U}$ has near-minimal coherence for this basis. The overall coherence of $\mathcal{U}$, $\mu(\mathcal{U})$, is largely due to the high coherence of the red basis vector.

In matrix completion literature, coherence is widely used to quantify the *recoverability* of a low-rank matrix $\mathbf{X}$. Here, the same notion of coherence arises

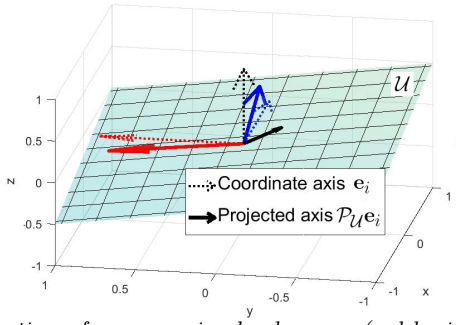


Figure 2: *A visualization of near-maximal coherence (red basis vector) and minimal coherence (black basis vector) for subspace* $\mathcal{U}$*.*

in a different context within the proposed model's uncertainty quantification. Lemma 2 provides the basis for this connection. Consider first the case where no matrix entries have been observed. From Lemma 2(a), vec($\mathbf{X}$) follows the degenerate Gaussian distribution $\mathcal{N}\{\mathbf{0}, \sigma^2(\mathcal{P}_{\mathcal{V}} \otimes \mathcal{P}_{\mathcal{U}})\}$. The variance of the (i, j)-th entry in $\mathbf{X}$ can then be shown to be:

$$\text{Var}(X_{i,j}) = \sigma^2(\mathbf{e}_i^T \mathcal{P}_{\mathcal{U}} \mathbf{e}_i)(\mathbf{e}_j^T \mathcal{P}_{\mathcal{V}} \mathbf{e}_j) = \sigma^2 \mu_i(\mathcal{U}) \mu_j(\mathcal{V}). \tag{16}$$

Hence, before observing data, the model uncertainty for entry $X_{i,j}$ is proportional to the product of coherences for the row and column spaces $\mathcal{U}$ and $\mathcal{V}$, corresponding to the i-th and the j-th basis vectors. Put another way, BayeSMG assigns *greater variation* to matrix entries with *higher* subspace coherence in either its row or column index. It is quite appealing given the original role of coherence in matrix completion, where larger row (or column) coherences imply greater "spikiness" for entries; our framework accounts for this by assigning greater model uncertainty to such entries.

Consider next the case where noisy entries $\mathbf{Y}_\Omega$ have been observed. Let us adopt a slightly generalized notion of coherence:

Definition 6 (Cross-coherence)**.** *The* cross-coherence *of subspace* $\mathcal{U}$ *with respect to the basis vectors* $\mathbf{e}_i$ *and* $\mathbf{e}_{i'}$ *is defined as* $\nu_{i,i'}(\mathcal{U}) = \mathbf{e}_{i'}^T \mathcal{P}_{\mathcal{U}} \mathbf{e}_i$*.*

The cross-coherence $\nu_{i,i'}(\mathcal{U})$ quantifies how correlated the basis vectors $\mathbf{e}_i$ and $\mathbf{e}_{i'}$ are, *after* a projection onto $\mathcal{U}$. For example, in Figure 2, the pair of red / blue projected basis vectors have negative cross-coherence for $\mathcal{U}$, whereas the pair of blue / black projected vectors have positive cross-coherence. When $i = i'$, this cross-coherence reduces to the original coherence in Definition 5.

Define now the cross-coherence vector $\boldsymbol{\nu}_i(\mathcal{U}) = [\nu_{i,i_n}(\mathcal{U})]_{n=1}^N \in \mathbb{R}^N$, where again $\Omega = \{(i_n, j_n)\}_{n=1}^N$. From equation (4) in Lemma 2, the conditional variance

of entry $X_{i,j}$ for an unobserved index $(i,j) \in \Omega^c$ becomes:

$$\mathrm{Var}(X_{i,j}|\mathbf{Y}_\Omega) = \sigma^2\mu_i(\mathcal{U})\mu_j(\mathcal{V}) - \sigma^2\boldsymbol{\nu}_{i,j}^T\left[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}\right]^{-1}\boldsymbol{\nu}_{i,j}, \qquad (17)$$

where $\boldsymbol{\nu}_{i,j} := \boldsymbol{\nu}_i(\mathcal{U}) \circ \boldsymbol{\nu}_j(\mathcal{V})$, and $\circ$ denotes the entry-wise (Hadamard) product. The expression in (17) yields a nice interpretation. From a UQ perspective, the first term in (17), $\mu_i(\mathcal{U})\mu_j(\mathcal{V})$, is simply the unconditional uncertainty for entry $X_{i,j}$, *prior* to observing data. The second term, $\boldsymbol{\nu}_{i,j}^T[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}]^{-1}\boldsymbol{\nu}_{i,j}$, can be viewed as the *reduction* in uncertainty, *after* observing the noisy entries $\mathbf{Y}_\Omega$. This uncertainty reduction is made possible by the correlation structure imposed on $\mathbf{X}$, via the SMG model; (17) also yields valuable insight in terms of subspace correlation. The first term $\mu_i(\mathcal{U})\mu_j(\mathcal{V})$ can be seen as the joint correlation between (i) row space $\mathcal{U}$ to row index i, and (ii) column space $\mathcal{V}$ to column index j, *prior* to any observations. The second term can be viewed as the portion of this correlation *explained* by observed indices Ω.

4.3 Error monotonicity

This link between coherence and uncertainty then sheds insight on expected error decay. This is based on the following proposition:

Proposition 7 (Variance reduction)**.** *Suppose* $\mathbf{X}$ *follows the BayeSMG model in Table 1, with* $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{0}$ *and fixed* σ^2 *and* η^2*. Let* $\mathbf{Y}_\Omega$ *contain the noisy entries at* $\Omega \subseteq [m_1] \times [m_2]$*, and let* $\mathbf{Y}_{\Omega\cup(i,j)}$ *contain an additional noisy observation at* $(i,j) \in \Omega^c$*. For any index* $(k,l) \in [m_1] \times [m_2]$*, the expected variance of* $X_{k,l}$ *can be decomposed as*

$$\mathbb{E}_{\mathcal{U},\mathcal{V}}[\mathrm{Var}(X_{k,l}|\mathbf{Y}_{\Omega\cup(i,j)})] = \mathbb{E}_{\mathcal{U},\mathcal{V}}[\mathrm{Var}(X_{k,l}|\mathbf{Y}_\Omega)] - \mathbb{E}_{\mathcal{U},\mathcal{V}}\Big[\frac{\mathrm{Cov}^2(X_{k,l}, X_{i,j}|\mathbf{Y}_\Omega)}{\mathrm{Var}(X_{i,j}|\mathbf{Y}_\Omega) + \eta^2}\Big], \qquad (18)$$

where $\mathrm{Var}(X_{k,l}|\mathbf{Y}_{\Omega\cup(i,j)})$ *is provided in* (17)*, and*

$$\mathrm{Cov}(X_{i,j}, X_{k,l}|\mathbf{Y}_\Omega) = \sigma^2\{\nu_{i,k}(\mathcal{U})\nu_{j,l}(\mathcal{V}) - \boldsymbol{\nu}_{i,j}^T\left[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}\right]^{-1}\boldsymbol{\nu}_{k,l}\}.$$

Remark: Proposition 7 shows, given observed indices Ω, the reduction in uncertainty (as measured by variance) for an unobserved entry $X_{k,l}$, after observing an additional entry at index (i,j). The last term in (18) quantifies this reduction, and can be interpreted as follows. For an unobserved index $(k,l) \notin \Omega \cup (i,j)$, the amount of uncertainty reduction is related to the "signal-to-noise" ratio,

where the signal is the conditional squared-covariance between the "unobserved" entry $X_{k,l}$ and the "to-be-observed" entry $X_{i,j}$, and the noise is the conditional variance of the "to-be-observed" entry.

The insight of *error monotonicity* then follows:

Corollary 1 (Error monotonicity)**.** *Suppose* $\mathbf{X}$ *follows the BayeSMG model in Table 1, with* $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{0}$ *and fixed* σ^2 *and* η^2*. Let* $[(i_n, j_n)]_{n=1}^{m_1 m_2} \subseteq [m_1] \times [m_2]$ *be an arbitrary sampling sequence, where* $(i_n, j_n) \neq (i_{n'}, j_{n'})$ *for* $n \neq n'$*. Let* $X^P_{k,l}$ *be the* (k,l)*-th entry of the conditional mean in* (4)*. Define the error term*

$$\epsilon^2_N(k,l) := \mathbb{E}_{\mathbf{X}}\left[\left(X_{k,l} - X^P_{k,l}\right)^2 \middle| \mathbf{Y}_{\Omega_{1:N}}\right],\ (k,l) \in [m_1] \times [m_2].$$

Then $\epsilon^2_{N+1}(k,l) \leq \epsilon^2_N(k,l)$ *for any* $(k,l) \in [m_1] \times [m_2]$ *and* $N = 1, 2, \cdots$.

Remark: This corollary shows that, for any sampling sequence and any index (k,l), the expected squared-error in estimating $X_{k,l}$ with the conditional mean $X^P_{k,l}$ is always monotonically decreasing as more samples are collected. This is intuitive since one expects to gain greater accuracy and precision on the unknown matrix $\mathbf{X}$ as more entries are observed. The fact that the proposed model quantifies this monotonicity property provides a reassuring check on our UQ approach.

5 Numerical experiments

We now investigate the performance of the proposed BayeSMG method in numerical simulations and compare it to the BPMF method (Salakhutdinov and Mnih, 2008), a popular Bayesian matrix completion method in the literature.

5.1 Synthetic data

For the first numerical study, we assume the true matrix $\mathbf{X}$ is generated from the SMG distribution, i.e., as $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{V}}, \sigma^2 = 1, R = 2)$, with uniformly sampled subspaces $\mathcal{U}$ and $\mathcal{V}$. The matrix $\mathbf{X}$ entries are assumed to be missing-at-random and the observed entries are contaminated by noise with a variance $\eta^2 = 0.05^2$, which we presume to be known. The prior specifications are as follows. For BayeSMG, we assign a weakly-informative prior $\sigma^2 \sim IG(0.01, 0.01)$ on the variance parameter σ^2, with non-informative manifold hyperparameters $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{0}$. For BPMF, we assign the recommended weak Inverse-Wishart

priors on covariance matrices $\mathbf{\Sigma}_M \sim \text{IW}(R = 2, \mathbf{I})$, $\mathbf{\Sigma}_N \sim \text{IW}(R = 2, \mathbf{I})$. We then ran 10,000 MCMC iterations for both methods, with the first 2,000 samples taken as burn-in. Standard MCMC convergence checks were performed via trace plot inspection (see Figure 3 (b)) and the Gelman-Rubin statistic (Gelman and Rubin, 1992).

We employ two metrics to compare the posterior contraction and UQ performance of these two methods. The first is the Mean Frobenius Error (MFE), defined as

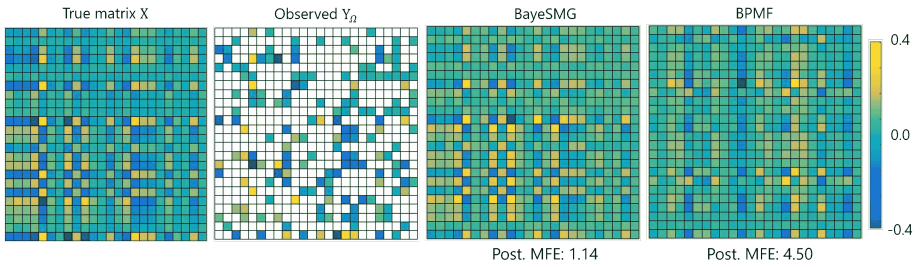
$$\text{MFE} = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{X} - \mathbf{X}_{[t]}\|_F.$$

The MFE calculates the Frobenius norm of the difference between the posterior predictive samples $\{\mathbf{X}_{[t]}\}_{t=1}^T$ and the original matrix $\mathbf{X}$. A smaller MFE suggests better recovery and faster posterior contraction for the desired matrix $\mathbf{X}$. The second metric is the Mean Spectral Distance (MSD), defined as

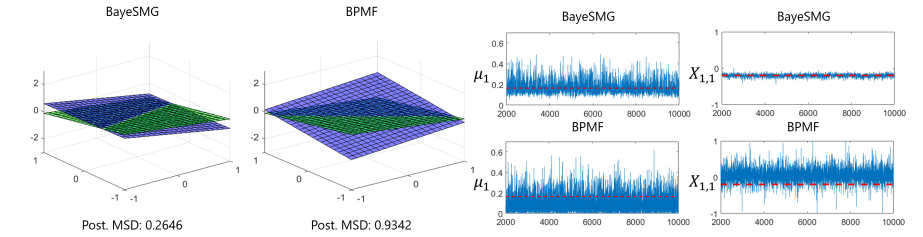
$$\text{MSD} = \frac{1}{T}\sum_{t=1}^{T} d_S(\mathcal{U},\mathcal{U}_{[t]}), \quad d_S(\mathcal{U},\mathcal{U}') := \sqrt{1 - \|\mathbf{U}^T\mathbf{U}'\|_2^2},$$

where $\mathbf{U}$ (or $\mathbf{U}'$) is any frame in subspace $\mathcal{U}$ (or $\mathcal{U}'$). The MSD calculates the spectral distance (Calderbank et al., 2015) between the posterior samples $\{\mathcal{U}_{[t]}\}_{t=1}^T$ for the row subspaces (equivalently, $\{\mathcal{V}_{[t]}\}_{t=1}^T$ for the column subspaces) and the true row subspace $\mathcal{U}$ (equivalently, the true column subspace $\mathcal{V}$). A smaller MSD suggests better recovery and posterior contraction for matrix subspaces.

The first two plots in Figure 3(a) visualize the true matrix $\mathbf{X}$ and the observed $\mathbf{Y}_\Omega$, with 20% of the entries observed uniformly-at-random. Here, the rank R is estimated via the approximate MAP approach in Section 3.3. The two subsequent plots visualize the posterior mean estimates for $\mathbf{X}$ using BayeSMG and BPMF. We see that the BayeSMG method provides visually better recovery of the matrix $\mathbf{X}$, with a lower posterior MFE than the BPMF method. The first two plots in Figure 3(b) visualize the true and estimated row spaces using BayeSMG and BPMF. We again see that BayeSMG gives a visually better recovery of the row space of $\mathbf{X}$ (the same holds for its column space), with a lower posterior MSD than BPMF. The next two plots show the trace plots for the first-row coherence μ_1 and the first matrix entry $X_{1,1}$, which is unobserved. We see that the posterior samples for BayeSMG concentrate tightly around the true coherence and matrix values, whereas the posterior samples for BPMF fluctuate much



(a) The four plots show (from left to right) the true matrix $\mathbf{X}$, observations $\mathbf{Y}_\Omega$, and the posterior mean estimates from BayeSMG and BPMF.



(b) The left plots visualize the true row space (green) and estimated row space (blue) from BayeSMG and BPMF for the first two dimensions, with posterior MSD calculated. The right plots show the trace plots for row coherence μ_1 and an unobserved entry $X_{1,1}$, for BayeSMG and BPMF, with true values dotted in red.

Figure 3: *Recovery and UQ performance for a simulated 25×25 matrix.*

more around the truth. The above observations suggest that when the matrix is generated from the assumed prior model, BayeSMG yields much faster posterior contraction than BPMF, leading to more accurate and precise estimates of $\mathbf{X}$ and its subspaces. Next, we will show in the following image recovery and seismic sensor applications that the BayeSMG method provides similar improvements over BPMF via modeling and integrating subspace information.

5.2 Image inpainting

Image inpainting is a fundamental problem in image processing (Bertalmio et al., 2000; Cai et al., 2010), which aims to recover and reconstruct images with missing pixels and noise corruption. It appears in numerous applications where image data are susceptible to unreliable data transmission and scratches. Take, for example, the problem of solar imaging (Xie et al., 2012). When a satellite transmits an image of the sun back to the earth, many pixels will inevitably be lost or corrupted due to the instabilities in the transmission process. The

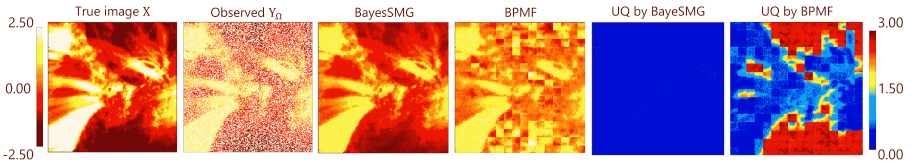


Figure 4: *Performance comparison between BayeSMG and BPMF on a* 256×256 *solar flare image. The plots (from left to right) show the original image, the partially observed image with noise, the recovered images using BayeSMG and BPMF, and the widths of the entry-wise 95% HPD intervals from BayeSMG and BPMF.*

missing pixels would become a problem when the image is scaled up. In this case, the quantification of image uncertainty can be as important as the recovery, since this UQ provides insight into the quality of recovered image features in different regions. There has been some work on applying deterministic matrix completion methods for image in-painting (e.g., Xue et al., 2017), but little has been done on uncertainty quantification. Our method addresses the latter goal.

We consider the aforementioned solar imaging problem, where the matrix $\mathbf{X}$ is a 256×256 image solar flare. The pixel intensity value is encoded from 0 to 255 and represents the use of pseudo-color in the images. We then normalize pixel intensities to have zero mean and unit variance. Half of the pixels in this image are observed uniformly at random, then corrupted by Gaussian noise $\eta^2 = 0.05^2$. We note that, for this problem, the recovery and UQ of the row and column subspaces are of interest as well. This is because image features are often represented in the row and column spaces. Here, these subspaces may represent domain-specific, interpretable phenomena, such as different classes of solar flares, certain shapes, and sunspots. Furthermore, human eyes are typically not as sensitive to high-frequency image features; therefore, a few SVD components can often capture the vital features of an image, making its rank low. For BayeSMG and BPMF, we estimate the rank R following the approximate MAP approach in Section 3.3, and perform 1,000 iterations of MCMC, with a burn-in period of 200. As before, MCMC convergence checks were performed via trace plot inspection and standard diagnostics.

Figure 4 shows the original solar image, its partial observations, and the recovered image using BayeSMG and BPMF via its posterior predictive mean, as well as its corresponding uncertainties via its 95% highest posterior density (HPD) interval width (Hyndman, 1996). We see that the BayeSMG method provides a much better recovery, with a noticeably lower MFE of 31.0 compared to the BPMF method (350.8). Visually, we see that the BayeSMG recovery

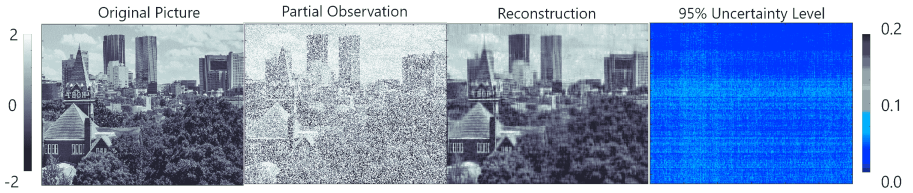


Figure 5: *Performance of BayeSMG on recovering a large* 1911×3000 *image of the Georgia Tech campus. The four plots show (from left to right) the original image, the partial observations, the recovered image using BayeSMG, and the widths of the entry-wise 95% HPD intervals from BayeSMG.*

captures the key features of the image, e.g., different types of solar flares. The BPMF recovery, on the other hand, loses much of the smaller-scale features and contains significant blocking defects. One plausible explanation is when a low-rank subspace structure is present in $\mathbf{X}$ (as is the case here), the proposed method can better learn and integrate this structure for improved recovery. Apart from that, an inspection of the HPD plots shows that the BayeSMG provides more accurate estimates of the recovered image, with narrow posterior HPD intervals across the whole matrix. In contrast, the BPMF is much more uncertain of its recovery: its entry-wise posterior density intervals are considerably larger, particularly for pixels with low intensities. Computation-wise, the posterior sampling for BayeSMG can be carried out within one minute on a standard laptop (Intel i7 processor with 16GB RAM), which is quite fast considering the relatively large image size.

To demonstrate the scalability of BayeSMG, we consider next a much higher-dimensional image of the Georgia Tech campus. This image is converted to a gray-scale matrix of size 1911×3000 and standardized to zero mean and unit variance. As before, half of the pixels are observed uniformly at random, then corrupted by a Gaussian noise $\eta^2 = 0.05^2$. To reduce computation time for posterior sampling, we fix the rank as $R = 30$ for both BayeSMG and BPMF, instead of estimating the rank using the procedure in Section 3.3. We run the MCMC sampler for 500 iterations after a burn-in period of 100.

Figure 5 shows the true image, its partial observations, and the recovered image from BayeSMG as well as its corresponding uncertainty. The MFE of this recovery is 1005.1, which is again noticeably smaller than that for the BPMF recovery (3004.8). We see that the recovered BayeSMG image captures the original image's main features, which shows that the proposed method can learn and integrate the subspace structure for recovery. As before, the BayeSMG is

quite confident of this completion, with narrow posterior HPD intervals over all pixels. Despite this being a much larger image, we can still carry out BayeSMG on the same standard laptop, albeit with a time of close to two hours. It suggests that the proposed method can yield effective probabilistic matrix recovery in high-dimensional settings.

Recall from Section 3.2 that the proposed posterior sampler for BayeSMG implicitly assumes the matrix entries are missing at random. To see how robust BayeSMG is to slight deviations from this MAR assumption, we investigate the recovery performance of BayeSMG for a 256×256 lighthouse image, where the entries are missing in a not-at-random setting. In particular, we consider the MNAR case where image pixels with a higher intensity value (i.e., darker) are more likely to be observed, and pixels with a lower intensity value (i.e., lighter) are more likely to be missing. Here, 40% of the entries with intensities higher than the population median are observed randomly, 25% of entries with intensities equal to the median are observed randomly, and 10% of remaining entries are observed randomly. Overall, around 25.1% of image pixels are observed using this scheme, but the probability of missing for a single pixel depends on the true pixel intensity.

Figure 6 shows the sampled image pixels for this MNAR setting with its corresponding image recovery via the posterior mean of the BayeSMG method. For comparison, we also illustrate the sampled pixels under an MCAR setting (where every entry is observed independently with probability 25%), with its corresponding image recovery via BayeSMG. We estimate the ranks in both scenarios via the procedure in Section 3.3. For the MNAR case, the MFE is 154.35, compared to an MFE of 148.33 for the MCAR case. While the error is slightly higher for the MNAR case (around 4% larger), we see from Figure 6 that there is little discernible difference visually between the recovered images in both cases. It suggests that the proposed BayeSMG sampler appears to be quite robust to mild violations of the implicit missing-at-random assumption for Algorithm 1. However, if prior information on the MNAR nature of the missing entries is known, then we can integrate such information within BayeSMG, yielding further improvements in recovery performance (see Section 3.2).

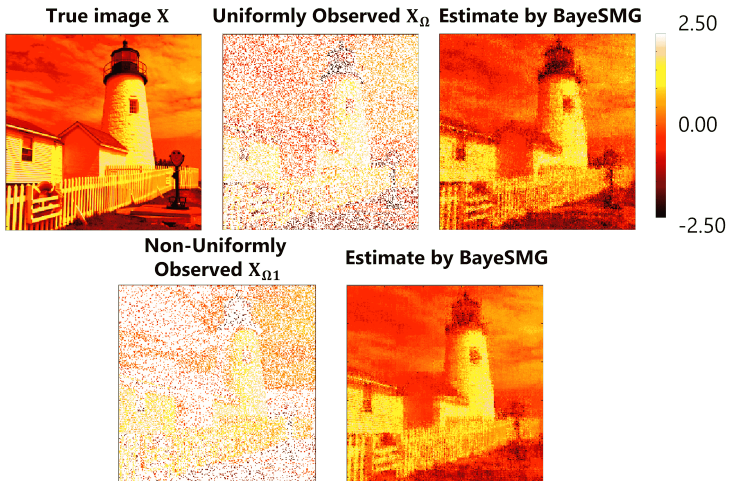
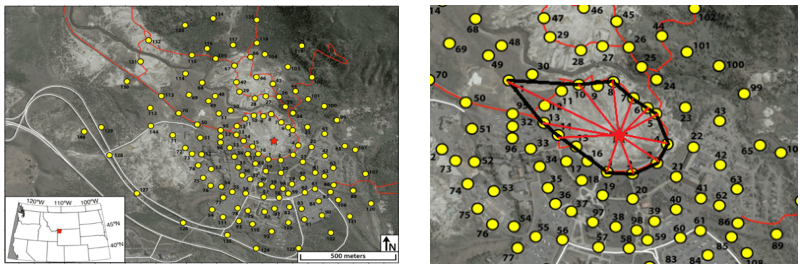


Figure 6: *Performance of BayeSMG on MNAR image pixels. In the first row, the first image is the original matrix, the second is the noisy matrix with entries sampled uniformly at random (MAR), and the third is its recovery estimate via the posterior mean of BayeSMG. In the second row, the first image is the noisy matrix with entries sampled MNAR, and the second image is its recovery estimate via BayeSMG.*

6 Seismic sensor network recovery

Seismic imaging is applied widely for finding oil and natural gas beneath the surface of the earth. Ambient Noise Seismic Imaging (Bensen et al., 2007) is a relatively new technique for seismic imaging with great potential. It uses "ambient noises" instead of actively collected signals and is non-invasive to the environment (compared to the traditional active imaging techniques). ANSI has proved useful for imaging shallow earth structures; it utilizes pairwise cross-correlation function between signals recorded by seismic sensors followed by time-frequency analysis. In a recent study (Xu et al., 2019) on the Old Faithful geyser at Yellowstone National Park, 133 sensors were deployed in its vicinity to collect ambient noise signals for investigating geological structures. Figure 7(a) shows the locations of these sensors.

One shortcoming of ANSI, however, is that many pairwise cross-correlations do not contain identifiable signals. In other words, the peak in the cross-correlation is unobserved since ANSI works on weak ambient noises. This missing data then results in missing entries in the cross-correlation matrix. To determine whether a cross-correlation is "missing", we first identify which correlations have an unsatisfactory signal-to-noise ratio (SNR), by inspecting the



(a) (b)

Figure 7: *The location of all 133 sensors near the geyser in Yellowstone National Park. The yellow circles indicate the sensors and the red pentagram indicates the location of the geyser. (a) shows the distribution of all 133 sensors over the region close to the geyser (see Wu et al., 2017 for details); (b) shows the locations of the 12 most significant sensors and their relative direction from each other.*

standard deviation ξ outside of the main wave lobe relative to the magnitude of the wave peak g. The correlation is deemed missing if $g/\xi < 20$. We note that entries on this cross-correlation matrix $\mathbf{X}$ are observed with noise due to background vibrations caused by bubble collapse and boiling water. Here, the standard deviation of the noise is estimated to be $\eta = 0.05$ from an inspection of sensor readings during the period when only noise signals are present. Figure 8 shows the observed noisy matrix entries $\mathbf{Y}_\Omega$.

To proceed with ANSI analysis, one would then need to estimate missing entries in the cross-correlation matrix $\mathbf{X}$. Bensen et al. (2007) shows that such a matrix is indeed low-rank. Sensors No. 1 to 16 record the tremor signals from the geyser most accurately and create strong signals with high SNR. These 16 sensors provide the dominant modes in the pairwise correlation signals across all sensors. Therefore we designate the matrix rank as 16, guided by prior geological knowledge. Here, uncertainty quantification is crucial for estimating geologic structure and identifying source of activities. With this uncertainty, engineers can better interpret the wave tomography generated from time delay estimates, and identify parts where estimates are accurate and where they are not. This in turn impacts the accuracy of analysis downstream, which subsequently provides greater insight on reconstruction quality.

Figure 8 visualizes the recovery and UQ performance from BayeSMG and BPMF, using an estimated rank of $R = 15$ via the approach in Section 3.3. We see that the BayeSMG yields much more precise estimates (i.e., narrower HPD interval widths) compared to the BPMF. In particular, when an entire

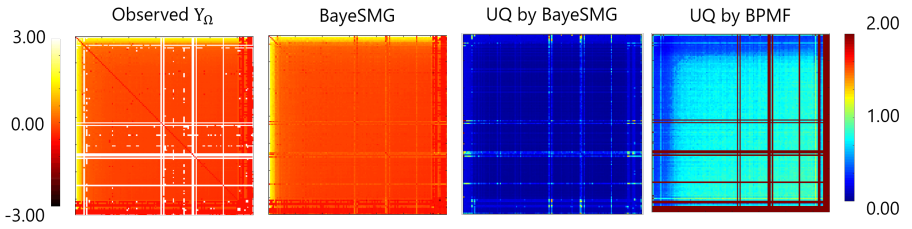


Figure 8: *Performance comparison between BayeSMG and BPMF on the ambient noise cross-correlation data matrix. The first plot (from the left) shows the observed entries in the cross-correlation matrix, with missing entries in white. The second plot shows the completed matrix via the posterior mean from BayeSMG. The third and fourth plots visualize the widths of the entry-wise 95% HPD intervals from BayeSMG and BPMF.*

row or column of $\mathbf{X}$ is missing, the uncertainties returned by BPMF can be very high, which reduces the usefulness of its recovered entries. On the contrary, the proposed BayeSMG method, by leveraging subspace information, can yield more precise inference on these missing rows and columns. One underlying reason is that the BayeSMG approach explicitly integrates subspace modeling for recovery and UQ. From the visualization of $\mathbf{Y}_\Omega$ in Figure 8, we see that there are clearly-seen bright stripes in the left and top edges of the plot, which strongly suggests the presence of low-rank subspaces in $\mathbf{X}$. It is not a surprise since we know several sensors have highly correlated signals due to their proximity. The BayeSMG appears to exploit this subspace structure to provide more confident predictions. The BPMF yields much higher uncertainty in inference, particularly in rows and columns with little to no observations. While the ground truth for the entire matrix $\mathbf{X}$ is not known for this sensor network, we would expect from previous experiments that the BayeSMG yields improved recovery performance over the BPMF, particularly in rows and columns with few observations.

With posterior samples on $\mathbf{X}$ in hand, we can then use its subspace information to locate (or match) a few sensors that contain highly correlated signals with each other. This sensor matching is helpful in seismology studies since we can use it to estimate the dimension and the capacity of the hydrothermal reservoir of the geyser (Wu et al., 2017). We first perform an SVD step on the posterior mean $\hat{\mathbf{X}}$, and find the singular vector with the largest singular value. We then inspect all the rows of the matrix $\hat{\mathbf{X}}$, and select the rows most aligned with this vector. We check these rows to locate the most significantly correlated sensors. Figure 7(b) shows the locations of the 12 most correlated sensors and their

relative directions from each other. The identified sensors are among the closest to the Old Faithful geyser, and their related observations are dominated by the highly fractured and porous geological structure underground adjacent to the geyser. Using readings from these sensors, researchers can identify a different pattern of the waveform in tremor signals, which suggests a variety of geological structures underneath the geyser.

7 Conclusions

We proposed a new BayeSMG model for uncertainty quantification in low-rank matrix completion. A key novelty of the BayeSMG model is that it parametrizes the unknown matrix $\mathbf{X}$ via manifold prior distributions on its row and column subspaces. This Bayesian subspace parametrization allows for direct posterior inference on matrix subspaces, which we can use for improved matrix recovery. We introduced a Gibbs sampler for posterior inference, which provides efficient posterior sampling even for matrices with dimensions on the order of thousands. Additionally, we showed that BayeSMG provides a probabilistic interpretation for subspace coherence, which we can use to show an error monotonicity result for UQ. We then showed the effective recovery and UQ performance of BayeSMG on simulated data, image data, and an application for seismic sensor network recovery.

For future work, it would be interesting to design locations for observations to control the uncertainties, exploring the connection with experimental design literature, e.g., integrated mean-squared error designs (Sacks et al., 1989) or distance-based designs (Mak and Joseph, 2018).

Acknowledgment

Henry Shaowu Yuchi and Yao Xie are supported by NSF CCF-1650913, NSF DMS-1938106, and NSF DMS-1830210. Simon Mak is supported by NSF CSSI Frameworks grant 2004571. The data and picture used in the seismic sensor network recovery are provided by Sin-Mei Wu and Fan-Chi Lin.

References

Alquier, P., Cottet, V., Chopin, N., and Rousseau, J. (2014). Bayesian matrix completion: prior specification. *arXiv preprint arXiv:1406.1440*.

Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2011). Low-rank matrix completion by variational sparse Bayesian learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2188–2191.

Bensen, G., Ritzwoller, M., Barmin, M., Levshin, A. L., Lin, F., Moschetti, M., Shapiro, N., and Yang, Y. (2007). Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International*, 169(3):1239–1260.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Calderbank, R., Thompson, A., and Xie, Y. (2015). On block coherence of frames. *Applied and Computational Harmonic Analysis*, 38(1):50–71.

Candès, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Candès, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Carson, W. R., Chen, M., Rodrigues, M. R., Calderbank, R., and Carin, L. (2012). Communications-inspired projection design with application to compressive sensing. *SIAM Journal on Imaging Sciences*, 5(4):1185–1212.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

Chikuse, Y. (2012). *Statistics on Special Manifolds*. Springer Science & Business

Media.

Davenport, M. A. and Romberg, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622.

Eriksson, B., Balzano, L., and Nowak, R. (2012). High-rank matrix completion. In *Artificial Intelligence and Statistics*, pages 373–381. PMLR.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2017). *The Elements of Statistical Learning*. Springer.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Gopal, S. and Yang, Y. (2014). Von Mises-Fisher clustering models. In *International Conference on Machine Learning*, pages 154–162.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. CRC Press.

Hernández-Lobato, J. M., Houlsby, N., and Ghahramani, Z. (2014). Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520. PMLR.

Hoff, P. D. (2007). Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685.

Hoff, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Hoff, P. D. (2013). Bayesian analysis of matrix data with rstiefel. *arXiv preprint arXiv:1304.3673*.

Hoffman, K. and Kunze, R. (1971). *Linear Algebra*. Englewood Cliffs, New Jersey.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

Jauch, M., Hoff, P. D., and Dunson, D. B. (2020). Monte Carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, pages 1–23.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998.

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

Koltchinskii, V., Lounici, K., Tsybakov, A. B., et al. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Lawrence, N. D. and Urtasun, R. (2009). Non-linear matrix factorization with Gaussian processes. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 601–608.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Mak, S. and Joseph, V. R. (2018). Support points. *Annals of Statistics*, 46(6A):2562–2592.

Mak, S. and Xie, Y. (2018). Maximum entropy low-rank matrix recovery. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):886–901.

Mardia, K. V. and Jupp, P. E. (2009). *Directional Statistics*, volume 494. John Wiley & Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 880–887.

Shapiro, A., Xie, Y., and Zhang, R. (2018). Matrix completion with deterministic pattern: A geometric perspective. *IEEE Transactions on Signal Processing*, 67(4):1088–1103.

Shen, J. (2001). On the singular values of Gaussian random matrices. *Linear Algebra and its Applications*, 326(1-3):1–14.

Smith, R. C. (2013). *Uncertainty Quantification: Theory, Implementation, and Applications*. SIAM.

Wang, Z. and Zhou, H. (2009). A general method of prior elicitation in bayesian reliability analysis. In *2009 8th International Conference on Reliability, Maintainability and Safety*, pages 415–419. IEEE.

Wu, S.-M., Ward, K. M., Farrell, J., Lin, F.-C., Karplus, M., and Smith, R. B. (2017). Anatomy of Old Faithful from subsurface seismic imaging of the Yellowstone Upper Geyser Basin. *Geophysical Research Letters*, 44(20):10–240.

Xie, Y., Huang, J., and Willett, R. (2012). Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27.

Xu, D., Song, B., Xie, Y., Wu, S.-M., Lin, F.-C., and Song, W. (2019). Low-rank matrix completion for distributed ambient noise imaging systems. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1059–1065. IEEE.

Xue, H., Zhang, S., and Cai, D. (2017). Depth image inpainting: Improving low rank matrix completion with low gradient regularization. *IEEE Transactions on Image Processing*, 26(9):4311–4320.

Zhang, X., Cui, W., and Liu, Y. (2020). Matrix completion with prior subspace information via maximizing correlation. *arXiv preprint arXiv:2001.01152*.

Zhou, M., Wang, C., Chen, M., Paisley, J., Dunson, D., and Carin, L. (2010). Nonparametric bayesian matrix completion. In *2010 IEEE Sensor Array and Multichannel Signal Processing Workshop*, pages 213–216. IEEE.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.

A Proofs

A.1 Proof of Lemma 2

Proof. We first prove part (a) of the lemma. To show that $\mathbf{X} \in \mathcal{T}$ almost surely, let $\mathbf{Z}$ be an arbitrary matrix in $\mathbb{R}^{m_1 \times m_2}$, with SVD $\mathbf{Z} = \widetilde{\mathbf{U}}\mathbf{D}\widetilde{\mathbf{V}}^T$, $\mathbf{D} = \text{diag}(\{d_k\}_{k=1}^R)$. Letting $\mathbf{u}_k = \mathcal{P}_\mathcal{U}\widetilde{\mathbf{u}}_k$ and $\mathbf{v}_k = \mathcal{P}_\mathcal{V}\widetilde{\mathbf{v}}_k$, where $\widetilde{\mathbf{u}}_k$ and $\widetilde{\mathbf{v}}_k$ are column vectors for $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ respectively, we have $\mathbf{u}_k \in \mathcal{U}$ and $\mathbf{v}_k \in \mathcal{V}$ for $k = 1, \cdots, R$. From Definition 1, $\mathbf{X}$ can then be written as $\mathbf{X} = \mathcal{P}_\mathcal{U}\mathbf{Z}\mathcal{P}_\mathcal{V} = (\mathcal{P}_\mathcal{U}\widetilde{\mathbf{U}})\mathbf{D}(\mathcal{P}_\mathcal{V}\widetilde{\mathbf{V}})^T = \sum_{k=1}^R d_k \mathbf{u}_k \mathbf{v}_k^T$, as desired. Next, note that the pseudo-inverse of $\mathcal{P}_\mathbf{u}$, $(\mathcal{P}_\mathbf{u})^+$, is simply $\mathcal{P}_\mathbf{u}$, since $\mathcal{P}_\mathbf{u}(\mathcal{P}_\mathbf{u})^+\mathcal{P}_\mathbf{u} = (\mathcal{P}_\mathbf{u})^+\mathcal{P}_\mathbf{u}(\mathcal{P}_\mathbf{u})^+ = \mathcal{P}_\mathbf{u}$ by the idempotency of $\mathcal{P}_\mathbf{u}$, and $\mathcal{P}_\mathbf{u}(\mathcal{P}_\mathbf{u})^+ = (\mathcal{P}_\mathbf{u})^+\mathcal{P}_\mathbf{u}$ are both symmetric. Moreover, letting $\det^*$ be the pseudo-determinant operator, we have $\det^*(\mathcal{P}_\mathcal{U}) = \det^*(\mathbf{U}\mathbf{U}^T) = \det(\mathbf{U}^T\mathbf{U}) = 1$, and $\det^*(\mathcal{P}_\mathcal{V}) = 1$ by the same argument. Using this along with Theorem 2.2.1 in Gupta and Nagar (1999), the density function $f(\mathbf{X})$ and the distribution of $\text{vec}(\mathbf{X})$ follow immediately.

We now prove part (b) of the lemma. From part (a), we have $\text{vec}(\mathbf{X}) \sim \mathcal{N}\{\mathbf{0}, \sigma^2(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})\}$, so:

$$[\mathbf{Y}_\Omega, \mathbf{X}_{\Omega^c}] \sim \mathcal{N}\left\{\mathbf{0}, \begin{bmatrix} \sigma^2\mathbf{R}_N(\Omega) + \eta^2\mathbf{I} & \sigma^2(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega,\Omega^c} \\ \sigma^2(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega,\Omega^c}^T & \sigma^2(\mathcal{P}_\mathcal{V} \otimes \mathcal{P}_\mathcal{U})_{\Omega^c} \end{bmatrix}\right\}.$$

The expressions for $\mathbf{X}_{\Omega^c}^P$ and $\mathbf{\Sigma}_{\Omega^c}^P$ in (4) then follow from the conditional density of the multivariate Gaussian distribution. Part (c) of the lemma can be shown in a similar way as for part (b). □

A.2 Proof of Proposition 3

Proof. For fixed $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$, $\mathbf{X}$ can be written as:

$$\mathbf{X} = \mathcal{P}_\mathcal{U}\mathbf{Z}\mathcal{P}_\mathcal{V} = \mathbf{U}(\mathbf{U}^T\mathbf{Z}\mathbf{V})\mathbf{V}^T, \tag{19}$$

where $Z_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $\mathcal{P}_\mathcal{U} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{P}_\mathcal{V} = \mathbf{V}\mathbf{V}^T$. By Theorem 2.3.10 in Gupta and Nagar (1999), each entry of $\tilde{\mathbf{Z}} = \mathbf{U}^T\mathbf{Z}\mathbf{V} \in \mathbb{R}^{R \times R}$ follows $\tilde{Z}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Note that the distribution of $\tilde{\mathbf{Z}}$ is independent of the initial choice of $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$ (and thereby $\mathbf{U}$ and $\mathbf{V}$). By Theorem 1 of Shen (2001), $\tilde{\mathbf{Z}}$ can be further factorized via its SVD:

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{V}}^T, \tag{20}$$

with independent $\tilde{\mathbf{U}} \sim U(\mathcal{V}_{R,R})$, $\tilde{\mathbf{V}} \sim U(\mathcal{V}_{R,R})$ and $\text{diag}(\mathbf{D})$ following the repulsed normal distribution (7).

Next, assign independent uniform priors $U(\mathcal{G}_{R,m_1-R})$ and $U(\mathcal{G}_{R,m_2-R})$ on projection matrices $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$, which induces independent uniform priors $U(\mathcal{V}_{R,m_1-R})$ and $U(\mathcal{V}_{R,m_2-R})$ on frames $\mathbf{U}$ and $\mathbf{V}$. From (19), we have:

$$\mathbf{X} = \mathbf{U}(\tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{V}}^T)\mathbf{V}^T = (\mathbf{U}\tilde{\mathbf{U}})\mathbf{D}(\mathbf{V}\tilde{\mathbf{V}})^T =: \tilde{\tilde{\mathbf{U}}}\mathbf{D}\tilde{\tilde{\mathbf{V}}}^T. \tag{21}$$

Note that $\tilde{\tilde{\mathbf{U}}} = \mathbf{U}\tilde{\mathbf{U}}$ is an orthonormal frame, since $(\mathbf{U}\tilde{\mathbf{U}})^T(\mathbf{U}\tilde{\mathbf{U}}) = \tilde{\mathbf{U}}^T(\mathbf{U}^T\mathbf{U})\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \mathbf{I}$. Moreover, $\tilde{\tilde{\mathbf{U}}} \sim U(\mathcal{V}_{R,m_1-R})$, since $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are independent and uniformly distributed. Similarly, one can show $\tilde{\tilde{\mathbf{V}}} = \mathbf{V}\tilde{\mathbf{V}} \sim U(\mathcal{V}_{R,m_2-R})$ as well, which proves the proposition. □

A.3 Proof of Lemma 4

Proof. Since $U(\mathcal{G}_{R,m-R})$ is a special case of the matrix Langevin distribution (Section 2.3.2 in Chikuse (2012)), it follows from (2.3.22) of Chikuse (2012) that $[\mathcal{P}_\mathcal{U}|R] \propto 1$ and $[\mathcal{P}_\mathcal{V}|R] \propto 1$. For fixed η^2 and σ^2, the MAP estimator for $\mathbf{X}$ then becomes:

$$\begin{aligned}
\tilde{\mathbf{X}} \in & \underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\text{Argmax}} \; [\mathbf{Y}_\Omega|\mathbf{X},\eta^2][\mathbf{X}|\mathcal{P}_\mathcal{U},\mathcal{P}_\mathcal{V},\sigma^2,R]\cdot \\
& \qquad [\mathcal{P}_\mathcal{U}|R]\;[\mathcal{P}_\mathcal{V}|R]\;[R] \\
& \quad \text{s.t. } \mathcal{P}_\mathcal{U} \in \mathcal{G}_{R,m_1-R}, \mathcal{P}_\mathcal{V} \in \mathcal{G}_{R,m_2-R}, R \le m_1 \wedge m_2 \\
\in & \underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\text{Argmax}} \; \exp\left\{-\frac{1}{2\eta^2}\|\mathbf{Y}_\Omega - \mathbf{X}_\Omega\|_2^2\right\}\cdot \\
& \qquad \left[\frac{1}{(2\pi\sigma^2)^{R^2/2}}\exp\left\{-\frac{1}{2\sigma^2}\text{tr}\left[(\mathbf{X}\mathcal{P}_\mathcal{V})^T(\mathcal{P}_\mathcal{U}\mathbf{X})\right]\right\}\right]\cdot \\
& \quad \text{s.t. } \mathcal{P}_\mathcal{U} \in \mathcal{G}_{R,m_1-R}, \mathcal{P}_\mathcal{V} \in \mathcal{G}_{R,m_2-R}, R \le m_1 \wedge m_2 \\
\in & \underset{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}}{\text{Argmin}} \; \left[\frac{1}{\eta^2}\|\mathbf{Y}_\Omega - \mathbf{X}_\Omega\|_2^2 + \log(2\pi\sigma^2)R^2 + \right. \\
& \qquad \left. \frac{1}{\sigma^2}\text{tr}\left[(\mathbf{X}\mathcal{P}_\mathcal{V})^T(\mathcal{P}_\mathcal{U}\mathbf{X})\right]\right] \\
& \quad \text{s.t. } \mathcal{P}_\mathcal{U} \in \mathcal{G}_{R,m_1-R}, \mathcal{P}_\mathcal{V} \in \mathcal{G}_{R,m_2-R}, R \le m_1 \wedge m_2.
\end{aligned}$$

Since $\mathbf{X} = \mathcal{P}_\mathcal{U}\mathbf{Z}\mathcal{P}_\mathcal{V}$, we have $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ for some $\mathbf{D} = \text{diag}(\{d_k\}_{k=1}^R)$, $\mathbf{U} \in \mathbb{R}^{m_1\times R}$ and $\mathbf{V} \in \mathbb{R}^{m_2\times R}$, where $\mathbf{U}$ and $\mathbf{V}$ are R-frames satisfying $\mathcal{P}_\mathcal{U} = \mathbf{U}\mathbf{U}^T$

and $\mathcal{P}_{\mathcal{V}} = \mathbf{V}\mathbf{V}^T$. Hence:

$$\begin{aligned}
&\operatorname{tr}\left[(\mathbf{X}\mathcal{P}_{\mathbf{V}})^T(\mathcal{P}_{\mathbf{U}}\mathbf{X})\right] \\
&= \operatorname{tr}\left[(\mathbf{V}\mathbf{V}^T)(\mathbf{V}\mathbf{D}\mathbf{U}^T)(\mathbf{U}\mathbf{U}^T)(\mathbf{U}\mathbf{D}\mathbf{V}^T)\right] \\
&= \operatorname{tr}\left[(\mathbf{V}^T\mathbf{V})^2\mathbf{D}(\mathbf{U}^T\mathbf{U})^2\mathbf{D}\right] && \text{(cyclic invariance of trace)} \\
&= \operatorname{tr}\left[\mathbf{D}^2\right] && (\mathbf{V}^T\mathbf{V} = \mathbf{I} \text{ and } \mathbf{U}^T\mathbf{U} = \mathbf{I}) \\
&= \|\mathbf{X}\|_F^2, && \text{(Frob. norm is equal to Schatten 2-norm)}
\end{aligned}$$

which proves the expression in (14).

□

A.4 Proof of Theorem 7

Proof. Consider the following block decomposition:

$$\mathbf{R}_{N+1}(\Omega \cup (i,j)) + \gamma^2\mathbf{I} = \begin{pmatrix} \mathbf{R}_N(\Omega) + \gamma^2\mathbf{I} & \boldsymbol{\nu}_i(\mathcal{U}) \circ \boldsymbol{\nu}_j(\mathcal{V}) \\ [\boldsymbol{\nu}_i(\mathcal{U}) \circ \boldsymbol{\nu}_j(\mathcal{V})]^T & \mu_i(\mathcal{U})\mu_j(\mathcal{V}) + \gamma^2 \end{pmatrix}.$$

Using the Schur complement identity for matrix inverses Hoffman and Kunze (1971), we have:

$$\left[\mathbf{R}_{N+1}(\Omega \cup (i,j)) + \gamma^2\mathbf{I}\right]^{-1} = \begin{pmatrix} \boldsymbol{\Gamma} + \tau^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}\boldsymbol{\xi}^T\boldsymbol{\Gamma} & -\tau^{-1}\boldsymbol{\xi}^T\boldsymbol{\Gamma} \\ -\tau^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi} & \tau^{-1} \end{pmatrix}, \tag{22}$$

where $\boldsymbol{\xi} = \boldsymbol{\nu}_i(\mathcal{U}) \circ \boldsymbol{\nu}_j(\mathcal{V})$, $\boldsymbol{\Gamma} = \left[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}\right]^{-1}$ and $\tau = \mu_i(\mathcal{U})\mu_j(\mathcal{V}) - \boldsymbol{\xi}^T\Gamma\boldsymbol{\xi} + \gamma^2$. Using the conditional variance expression in (17), $\tau = \operatorname{Var}(X_{i,j}|\mathbf{Y}_\Omega)/\sigma^2 + \gamma^2$. Letting $\widetilde{\boldsymbol{\xi}} = \boldsymbol{\nu}_k(\mathcal{U}) \circ \boldsymbol{\nu}_l(\mathcal{V})$ and applying (17) again, it follows that:

$$\begin{aligned}
&\operatorname{Var}(X_{k,l}|\mathbf{Y}_{\Omega\cup(i,j)}) \\
&= \sigma^2\left\{\mu_k(\mathcal{U})\mu_l(\mathcal{V}) - \widetilde{\boldsymbol{\xi}}^T\Gamma\widetilde{\boldsymbol{\xi}}\right\} \\
&\quad - \tau^{-1}\sigma^2\left\{\boldsymbol{\nu}_{i,j}^T\left[\mathbf{R}_N(\Omega) + \gamma^2\mathbf{I}\right]^{-1}\boldsymbol{\nu}_{k,l} - \nu_{i,k}(\mathcal{U})\nu_{j,l}(\mathcal{V})\right\}^2 \\
&\qquad\qquad \text{(using (22) and algebraic manipulations)} \\
&= \operatorname{Var}(X_{k,l}|\mathbf{Y}_\Omega) - \frac{\operatorname{Cov}^2(X_{i,j}, X_{k,l}|\mathbf{Y}_\Omega)}{\operatorname{Var}(X_{i,j}|\mathbf{Y}_\Omega) + \eta^2}, && \text{(from (4))}
\end{aligned}$$

which proves the theorem. □

A.5 Proof of Corollary 1

Proof. This follows directly from Theorem 7 and the fact that:

$$\mathrm{Cov}^2(X_{i,j}, X_{k,l}|\mathbf{Y}_{\Omega_{1:N}})/\{\mathrm{Var}(X_{i,j}|\mathbf{Y}_{\Omega_{1:N}}) + \eta^2\} > 0.$$

□

A.6 Proof of full conditional distributions

Proof. For fixed rank R, the posterior distribution $[\Theta|\mathbf{Y}]$ can be written as:

$$\begin{aligned}
[\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2|\mathbf{Y}] &\propto [\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2] \cdot [\mathbf{U}] \cdot [\mathbf{V}] \cdot [\mathbf{D}|\sigma^2] \cdot [\sigma^2] \\
&\propto \frac{1}{(\eta^2)^{(m_1 m_2)/2}} \exp\left\{-\frac{1}{2\eta^2}\|\mathbf{Y} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2\right\} \cdot \frac{1}{(\sigma^2)^{R/2}} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{R} d_k^2\right\} \cdot \prod_{\substack{k,l=1 \\ k<l}}^{R} |d_k^2 - d_l^2| \\
&\quad \cdot \frac{1}{(\sigma^2)^{\alpha_{\sigma^2}+1}} \exp\left\{-\frac{\beta_{\sigma^2}}{\sigma^2}\right\} \cdot \frac{1}{(\eta^2)^{\alpha_{\eta^2}+1}} \exp\left\{-\frac{\beta_{\eta^2}}{\eta^2}\right\}.
\end{aligned}$$

From this, the full conditional distributions can then be derived as follows:

$$\begin{aligned}
&[\mathbf{U}|\mathbf{Y}, \mathbf{D}, \mathbf{V}, \sigma^2, \eta^2] \propto \mathrm{etr}\{(\mathbf{Y}\mathbf{V}\mathbf{D})^T\mathbf{U}/\eta^2\} \sim vMF(m_1, R, \mathbf{Y}\mathbf{V}\mathbf{D}/\eta^2), \\
&[\mathbf{V}|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \sigma^2, \eta^2] \propto \mathrm{etr}\{(\mathbf{Y}^T\mathbf{U}\mathbf{D})^T\mathbf{V}/\eta^2\} \sim vMF(m_2, R, \mathbf{Y}^T\mathbf{U}\mathbf{D}/\eta^2), \\
&[\mathbf{D}|\mathbf{Y}, \mathbf{U}, \mathbf{V}, \sigma^2, \eta^2] \propto \exp\left\{-\frac{1}{2\eta^2}\|\mathbf{Y} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{R} d_k^2\right\} \prod_{\substack{k,l=1 \\ k<l}}^{R} |d_k^2 - d_l^2| \\
&\qquad\qquad \sim \mathcal{RN}\left(\sigma^2\mathrm{diag}(\mathbf{U}^T\mathbf{Y}\mathbf{V})/(\eta^2+\sigma^2), \eta^2\sigma^2/(\eta^2+\sigma^2)\right) \\
&[\sigma^2|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}, \eta^2] \sim IG(\alpha + R/2, \beta + \mathrm{tr}(\mathbf{D}^2)/2) \\
&[\eta^2|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2] \sim IG(\alpha_{\eta^2} + m_1 m_2/2, \beta_{\eta^2} + \|\mathbf{Y} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2/2).
\end{aligned}$$

□