
Information-Guided Sampling for Low-Rank Matrix Completion

Simon Mak¹ Henry Shaowu Yuchi² Yao Xie²

Abstract

The matrix completion problem, which aims to recover a low-rank matrix \mathbf{X} from partial, noisy observations of its entries, arises in many machine learning applications. In this work, we present a novel information-theoretic framework for initial and sequential sampling of matrix entries for noisy matrix completion, based on the maximum entropy sampling principle in Shewry & Wynn (1987). The key novelty in our approach is that it makes use of uncertainty quantification (UQ) – a measure of uncertainty for unobserved entries – to guide the sampling procedure. Our framework reveals new insights on the role of coherence and coding design on sampling for matrix completion.

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 (e.g., in gene studies (Natarajan & Dhillon, 2014)); (ii) there can be missing observations at individual entries. The *matrix completion* problem aims to complete the missing entries of \mathbf{X} from partial (and often noisy) observations.

Matrix completion has attracted much attention since the seminal works of Candès & Tao (2010), Candès & Recht (2009), and Recht (2011). This is then extended to the *noisy* matrix completion setting, where entries are observed with noise; important results include Candès & Plan (2010), Keshavan et al. (2010), Koltchinskii et al. (2011), and Negahban & Wainwright (2012), among others. There is now a

This work is supported by NSF CSSI Frameworks Grant 2004571. ¹Department of Statistical Science, Duke University, North Carolina, USA. ²H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Georgia, USA.. Correspondence to: Henry Shaowu Yuchi <shaowu.yuchi@gatech.edu>.

rich body of work on matrix completion with various convex and non-convex optimization algorithms; recent overviews include Davenport & Romberg (2016) and Chi et al. (2019). However, much of the literature has focused on the *point estimation* of \mathbf{X} only, with little work on quantifying the uncertainty of such estimates and how this can help guide the sampling of matrix entries.

This work presents a novel approach for *designing* the observed entries in \mathbf{X} for low-rank matrix completion, with the goal of maximizing information on \mathbf{X} . While most of the literature assumes entries are sampled uniformly at random, there has been some work on adaptive sampling schemes (Sutherland et al., 2013; Lan et al., 2016; Bhargava et al., 2017; Ruchansky et al., 2015). There is also related work from binary relation learning (Goldman & Warmuth, 1995; Nakamura & Abe, 2002), but such methods do not consider low-rank structure. This paper makes novel contributions to the topic of *information-theoretic* sampling for low-rank matrix completion. Our work differs from the large body of literature on information-theoretic design (Palomar & Verdú, 2006; Carson et al., 2012; Wang et al., 2014; Shlezinger et al., 2017) in that, instead of maximizing the mutual information between signal (i.e., \mathbf{X}) and observed entries (denoted as \mathbf{Y}_Ω), we study a dual but equivalent problem of maximizing the *entropy* of observations \mathbf{Y}_Ω . Using the maximum entropy sampling principle from (Shewry & Wynn, 1987), this dual view sheds new insights into matrix completion sampling and provides simple closed-form criteria for initial and active learning.

2. Low Rank Matrix Modeling

2.1. The completion problem

Let $\mathbf{X} = (X_{i,j}) \in \mathbb{R}^{m_1 \times m_2}$ be the desired low-rank matrix. Suppose \mathbf{X} is observed with noise at N indices $\Omega_{1:N} = \{(i_n, j_n)\}_{n=1}^N \subseteq [m_1] \times [m_2]^1$ (sometimes denoted as Ω for brevity). Let $Y_{i,j}$ be the noisy observation at index $(i, j) \in \Omega$ under the Gaussian noise model:

$$Y_{i,j} = X_{i,j} + \epsilon_{i,j}, \quad \epsilon_{i,j} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \eta^2). \quad (1)$$

Further let $\mathbf{X}_\Omega \in \mathbb{R}^N$ and $\mathbf{Y}_\Omega \in \mathbb{R}^N$ denote the vectorized entries of \mathbf{X} and \mathbf{Y} at observed indices Ω , and let $\mathbf{X}_{\Omega^c} \in$

¹ $[m] := \{1, \dots, m\}$, $m_1 \wedge m_2 := \min(m_1, m_2)$, $m_1 \vee m_2 := \max(m_1, m_2)$.

$\mathbb{R}^{m_1 m_2 - N}$ and $\mathbf{Y}_{\Omega^c} \in \mathbb{R}^{m_1 m_2 - N}$ denote the vectorized entries of \mathbf{X} and \mathbf{Y} at unobserved indices $\Omega^c = ([m_1] \times [m_2]) \setminus \Omega$.

2.2. Singular Matrix-variate Gaussian model

Consider the following model for the low-rank matrix \mathbf{X} (assumed to be normalized with zero mean):

Definition 1 (Singular matrix-variate Gaussian (SMG); Definition 2.4.1, (Gupta & Nagar, 1999)). Let $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix with entries $Z_{i,j} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $(i, j) \in [m_1] \times [m_2]$. The random matrix \mathbf{X} has a singular matrix-variate Gaussian (SMG) distribution if $\mathbf{X} \stackrel{d}{=} \mathcal{P}_U \mathbf{Z} \mathcal{P}_V$ for some projection matrices $\mathcal{P}_U = \mathbf{U} \mathbf{U}^T$ and $\mathcal{P}_V = \mathbf{V} \mathbf{V}^T$, 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 $R < m_1 \wedge m_2$ ¹.

The projection matrices \mathcal{P}_U and \mathcal{P}_V provide a parametrization of the row and column subspaces of \mathbf{X} . Since such parameters are unknown in practice, we adopt a Bayesian approach (Gelman et al., 2014) and assign the following non-informative priors as in Yuchi et al. (2021):

$$\begin{aligned} [\mathcal{P}_U] &\sim \text{Unif}(\mathcal{G}_{R, m_1 - R}), & [\mathcal{P}_V] &\sim \text{Unif}(\mathcal{G}_{R, m_2 - R}), \\ [\eta^2] &\sim IG(\alpha_{\eta^2}, \beta_{\eta^2}), & [\sigma^2] &\sim IG(\alpha_{\sigma^2}, \beta_{\sigma^2}) \end{aligned} \quad (2)$$

where $\mathcal{G}_{R, m-R}$ is the Grassmann manifold (Chikuse, 2012) containing all $m \times m$ projection matrices of rank R .

Under such priors, the maximum-a-posteriori (MAP) estimator of matrix \mathbf{X} is closely connected to the nuclear-norm formulation, widely used for matrix completion (Candès & Recht, 2009; Candès & Tao, 2010):

$$\hat{\mathbf{X}} = \underset{\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}}{\text{Argmin}} \left[\sum_{(i,j) \in \Omega} (Y_{i,j} - X_{i,j})^2 + \lambda \|\mathbf{X}\|_* \right]. \quad (3)$$

Here, $\|\mathbf{X}\|_*$ is the nuclear norm which sums up the singular values of \mathbf{X} . Further details on this connection can be found in Yuchi et al. (2021).

3. Maximum entropy sampling for matrix completion

With this modeling framework, we present a information-theoretic method for sampling matrix entries, using the *maximum entropy principle*, which was first introduced in Shewry & Wynn (1987) for experimental design of spatio-temporal models. This has two parts: (a) an *initial sampling* strategy for preliminary learning on \mathbf{X} , and (b) a *sequential sampling* strategy to greedily maximize information gain.

3.1. The maximum entropy sampling principle

In the following, we use the definitions of entropy as presented in Cover & Thomas (2012). For two random variables (X, Y) , the chain rule (Theorem 2.2.1, Cover

& Thomas (2012)) connects the *joint entropy* $H(X, Y)$ with the *conditional entropy* $H(Y|X)$ via the identity $H(X, Y) = H(X) + H(Y|X)$. Consider now the noisy matrix completion problem. Applying this chain rule, we get the decomposition:

$$H(\mathbf{Y}_{\Omega}, \mathbf{X}) = H(\mathbf{Y}_{\Omega}) + H(\mathbf{X}|\mathbf{Y}_{\Omega}). \quad (4)$$

Here, $H(\mathbf{Y}_{\Omega}, \mathbf{X})$ is the joint entropy of observations \mathbf{Y}_{Ω} and matrix \mathbf{X} , $H(\mathbf{Y}_{\Omega})$ is the entropy of \mathbf{Y}_{Ω} , and $H(\mathbf{X}|\mathbf{Y}_{\Omega})$ is the conditional entropy of \mathbf{X} after observing \mathbf{Y}_{Ω} . To maximize information gain on \mathbf{X} from observations, we wish to sample indices Ω which minimize conditional entropy $H(\mathbf{X}|\mathbf{Y}_{\Omega})$.

Next, for the sampling model (1), it can be shown that the joint entropy $H(\mathbf{Y}_{\Omega}, \mathbf{X})$ does *not* depend on sampled indices Ω (this follows by reversing \mathbf{X} and \mathbf{Y}_{Ω} in (4), see Mak & Xie (2018) for details). Hence, a sampling scheme Ω which maximizes the entropy of *observations* \mathbf{Y}_{Ω} (i.e., $H(\mathbf{Y}_{\Omega})$) also minimizes the conditional entropy $H(\mathbf{X}|\mathbf{Y}_{\Omega})$, which yields maximum *information gain* on \mathbf{X} . The maximum entropy sampling principle for noisy matrix completion therefore aims to maximize $H(\mathbf{Y}_{\Omega})$. The key advantage of this principle is that it allows us to work with a simple, closed-form expression for $H(\mathbf{Y}_{\Omega})$ (derived below) as an efficient proxy for the desired entropy term $H(\mathbf{X}|\mathbf{Y}_{\Omega})$, which is more complicated and difficult to optimize.

From the SMG model, the exp-entropy² term $E(\Omega_{1:N}) := \exp\{H(\mathbf{Y}_{\Omega})\}$ admits a closed form:

Lemma 1 (Observational entropy). For fixed \mathcal{P}_U and \mathcal{P}_V ,

$$E(\Omega_{1:N}) := \exp\{H(\mathbf{Y}_{\Omega})\} = C \det\{\sigma^2 \mathbf{R}_N(\Omega_{1:N}) + \eta^2 \mathbf{I}\}, \quad (5)$$

for some constant C not depending on indices $\Omega_{1:N}$.

Our strategy is as follows. For *initial* sampling (Section 3.2), we will first derive a lower bound on (5), then generalize this bound under uniform (non-informative) priors on $(\mathcal{P}_U, \mathcal{P}_V)$. The maximization of this generalized bound yields an intuitive initial sampling strategy, with connections to coding design. For *sequential* sampling (Section 3.3), we will first use the nuclear-norm minimization in (3) to obtain subspace estimates, then plug in these estimates into (5) to derive an efficient, sequential sampling algorithm.

3.2. Initial sampling: Latin square design

For simplicity, assume $m_1 = m_2 = m$ (generalized later), with total initial samples $N = m$. Following Yuchi et al. (2021), we define the *coherence* of subspace \mathcal{U} for the i -th basis vector \mathbf{e}_i as $\mu_i(\mathcal{U}) := \|\mathcal{P}_U \mathbf{e}_i\|_2^2$, and the *cross-*

²Since $\exp(\cdot)$ is monotone, the maximum entropy principle holds when maximizing exp-entropy $E(\Omega_{1:N})$.

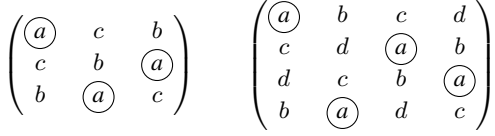


Figure 1. A 3×3 and a 4×4 Latin square. A balanced sampling scheme is obtained by sampling the entries with ‘a’ (circled).

coherence of \mathcal{U} for bases \mathbf{e}_i and $\mathbf{e}_{i'}$ as $\nu_{i,i'}(\mathcal{U}) = \mathbf{e}_{i'}^T \mathcal{P}_{\mathcal{U}} \mathbf{e}_i$. The lemma below gives a lower bound on $\mathbb{E}(\Omega_{1:N})$:

Proposition 1 (Lower bound on observation entropy). *For fixed $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$, we have*

$$\mathbb{E}^{1/N}(\Omega_{1:N}) \geq \min_{n=1, \dots, N} C \left[\sigma^2 \mu_{i_n}(\mathcal{U}) \mu_{j_n}(\mathcal{V}) + \eta^2 - \Psi(\Omega_{1:N}) \right], \quad (6)$$

where:

$$\Psi(\Omega_{1:N}) = \frac{\sigma^2(N-1)}{2R} \left\{ \max_{n':n' \neq n} \nu_{i_n, i_{n'}}^2(\mathcal{U}) + \max_{n':n' \neq n} \nu_{j_n, j_{n'}}^2(\mathcal{V}) \right\} \quad (7)$$

and C is a constant not depending on $\Omega_{1:N}$.

Using the right side of (6) as a proxy for $\mathbb{E}^{1/N}(\Omega_{1:N})$, the maximization of $\mathbb{E}(\Omega_{1:N})$ under the non-informative priors (2) can be approximated by the minimization of (7), since the coherence terms $\mu_{i_n}(\mathcal{U})$ and $\mu_{j_n}(\mathcal{V})$ are on average equal for any i_n and j_n , under uniform priors on $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$. This minimization amounts to jointly minimizing $\max_{n \neq n'} (\mathbf{e}_{i_n}^T \mathbf{e}_{i_{n'}})^2$ and $\max_{n \neq n'} (\mathbf{e}_{j_n}^T \mathbf{e}_{j_{n'}})^2$. Clearly, if $i_n = i_{n'}$ for some $n \neq n'$ (i.e., same row is sampled twice), then the first term attains the maximum value of 1. Likewise, if $j_n = j_{n'}$ for some $n \neq n'$ (i.e., same column is sampled twice), then the second term becomes 1 as well. Both are undesirable, since we wish to minimize the two terms in (7).

With little knowledge on \mathbf{X} , an initial sampling scheme achieving maximum entropy should therefore be *balanced*, in that all rows and columns should be sampled exactly once (in specific case of $N = m_1 = m_2$) or an equal number of times (for general matrices). This balanced sampling scheme can be nicely represented by a *Latin square* (Keedwell & Dénes, 2015): an $m \times m$ array with m distinct symbols, each occurring exactly once in each row and column. Figure 1 shows an example of a 3×3 and 4×4 Latin square; a balanced sample is obtained by choosing indices with a specific letter (e.g., ‘a’). Latin squares are widely used in error-correcting codes (Colbourn et al., 2004; Huczynska, 2006) and experimental design (Fisher, 1937), and there are efficient algorithms (Jacobson & Matthews, 1996) for generating randomized Latin squares.

3.3. Sequential sampling

Consider next the setting where the noisy entries \mathbf{Y}_{Ω} have been observed at $\Omega_{1:N}$, and suppose informed estimates are on subspaces \mathcal{U} and \mathcal{V} from \mathbf{Y}_{Ω} (more on this in Section 3.4). Fixing the observed indices $\Omega_{1:N}$, the sequential problem of sampling the next index $(i, j) \notin \Omega_{1:N}$ maximizing entropy $\mathbb{E}(\Omega_{1:N} \cup (i, j))$ can be formulated as:

Lemma 2. *For fixed $\mathcal{P}_{\mathcal{U}}$, $\mathcal{P}_{\mathcal{V}}$ and observed indices $\Omega_{1:N}$,*

$$\begin{aligned} & \underset{(i,j) \in \Omega_{1:N}^c}{\text{Argmax}} \mathbb{E}(\Omega_{1:N} \cup (i, j)) \\ &= \underset{(i,j) \in \Omega_{1:N}^c}{\text{Argmax}} \{ \mu_i(\mathcal{U}) \mu_j(\mathcal{V}) - \nu_{i,j}^T [\mathbf{R}_N(\Omega_{1:N}) + \gamma^2 \mathbf{I}]^{-1} \nu_{i,j} \}. \end{aligned} \quad (8)$$

Lemma 2 provides an *easy-to-evaluate* criterion for greedily maximizing information on \mathbf{X} . We give further heuristics for reducing computation time for this optimization in Section 3.4.

There are two useful interpretations of the sequential sampling scheme (8). First, it samples the index yielding greatest *information gain* on \mathbf{X} , given prior observations \mathbf{Y}_{Ω} . Second, it samples the index with greatest posterior *uncertainty* from the probabilistic model, given observations \mathbf{Y}_{Ω} . This links the objective of information-greedy active learning with the uncertainty quantification of matrix \mathbf{X} (this connection was noted earlier in MacKay (1992), but in a different active learning context).

3.4. MaxEnt: Information-theoretic sampling algorithm

We now combine these insights into an information-theoretic sampling algorithm MaxEnt for low-rank matrix completion. Assume $m_1 \geq m_2$. For *initial* sampling, one way to guarantee a balanced design is to (a) generate $\lfloor m_1/m_2 \rfloor$ random Latin squares (Jacobson & Matthews, 1996) of size $m_2 \times m_2$, (b) vertically stacking these squares to form an $(m_2 \lfloor m_1/m_2 \rfloor) \times m_2$ rectangle, and (c) sampling the entries labeled ‘a’ from this rectangle. Using these initial samples, the projection matrices $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ can then be estimated via the posterior sampling algorithm BayeSMG in Yuchi et al. (2021) (this is discussed in detail in Appendix A.1). From this, we sample the unobserved entry yielding the greatest expected posterior information gain from (8). These steps are repeated until a desired error is achieved. Algorithm 1 summarizes this procedure.

While the sequential sampling procedure makes use of an easy-to-evaluate acquisition function (8), there are several heuristics that can further speed up computation when \mathbf{X} is high-dimensional. First, the exhaustive search in (8) over all unobserved indices $\Omega_{1:N}^c$ can be time-consuming. One remedy is to screen out indices with small coherences $\mu_i(\mathcal{U})$ and $\mu_j(\mathcal{V})$ (which are likely poor entries to sample by (8)), then perform the search over a much smaller index set. Second,

Algorithm 1 MaxEnt

Input: Total samples $N_{\max} = N_{\text{ini}} + N_{\text{seq}}$
Initial sampling ($N_{\text{ini}} = m_1 \vee m_2$ samples):

- Generate & stack $\lfloor m_1/m_2 \rfloor$ random $m_2 \times m_2$ Latin squares.
- Set Ω as the entries labeled ‘a’.

Sequential sampling (N_{seq} samples):

for $n = N_{\text{ini}+1}$ **to** N_{max} **do**

- Run the posterior sampler BayeSMG (Yuchi et al., 2021).
- Obtain projection matrix estimates $(\hat{\mathcal{P}}_U, \hat{\mathcal{P}}_V)$ via posterior means.
- Sample next entry (i_n, j_n) using (8), with $(\mathcal{P}_U, \mathcal{P}_V) = (\hat{\mathcal{P}}_U, \hat{\mathcal{P}}_V)$.
- Update $\Omega \leftarrow \Omega \cup (i_n, j_n)$.

end for

the sequential procedure (8) can be extended to a *batch*-sequential procedure, which samples multiple indices simultaneously with large objective values. These heuristics enable MaxEnt to provide efficient information-theoretic sampling for matrices with dimensions on the order of thousands.

4. Numerical examples

4.1. Simulations

We first investigate the performance of MaxEnt in simulations. The true matrix \mathbf{X} is simulated from the SMG model with priors (2), with hyperparameters $\sigma^2 = 1$, $\eta^2 = 10^{-4}$, $\alpha_{\eta^2} = \alpha_{\sigma^2} = 9$, $\beta_{\eta^2} = 10^{-3}$, $\beta_{\sigma^2} = 10$. We consider matrices of sizes 30×30 and 60×60 , with true rank $R = 3$ and $R = 4$, respectively. We begin with $N_{\text{ini}} = m_1 = m_2$ initial samples, then observe $N_{\text{seq}} = 50$ and $N_{\text{seq}} = 100$ entries sequentially for the 30×30 and 60×60 cases. This is then replicated 10 times to measure error variability. Figure 2 shows the averaged errors and error quantiles for MaxEnt and uniform sampling. For initial sampling, MaxEnt yields reduced errors to uniform sampling, which shows the effectiveness of a *balanced* initial design. For sequential sampling, the improvement of MaxEnt over uniform sampling grows larger as more entries are sampled; near the end, the averaged errors from uniform sampling are noticeably higher than the 75% quantiles from MaxEnt. This shows the effectiveness of our integrated approach, in first (a) learning the underlying subspaces of \mathbf{X} via probabilistic Bayesian modeling, then (b) incorporating this learning to guide information-theoretic sampling.

4.2. Collaborative filtering

Next, we apply MaxEnt to two collaborative filtering datasets. The first, ‘Jester’, is collected from the Jester Online Joke Recommender System (Goldberg et al., 2001) – a test bank of 100 joke ratings from 500 users. The second, ‘MovieLens’, is a database of movie ratings for 300 users on 1,700 movies. For Jester, $N_{\text{ini}} = 500$ initial samples are

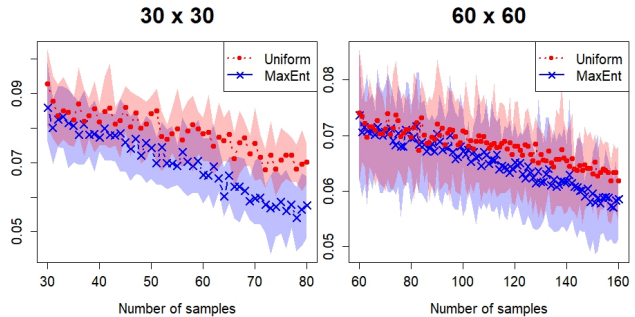


Figure 2. Avg. Frob. errors (line) and 25-th/75-th error quantiles (shaded) for the 30×30 and 60×60 simulated matrices.

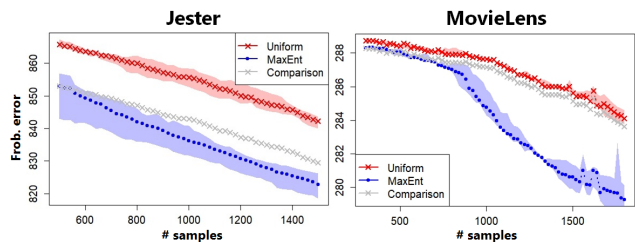


Figure 3. Avg. Frob. errors (line) and 25-th/75-th error quantiles (shaded) for Jester (left) and MovieLens (right).

collected, with $N_{\text{seq}} = 1,000$ entries observed sequentially; for MovieLens, $N_{\text{ini}} = 300$ and $N_{\text{seq}} = 1,500$. This is then replicated 10 times to measure error variability. The heuristics from Section 3.4 are used to speed up the sampling procedure.

Figure 3 shows the resulting errors using MaxEnt and uniform sampling. Two observations are of interest. First, MaxEnt yields lower initial errors to uniform sampling, which again demonstrates the importance of a balanced initial sample. Second, the gap between MaxEnt and uniform sampling grows larger as entries are observed sequentially, more so than from simulations. One reason is that high coherences are present in both datasets – there may be users who are overly critical, or jokes or movies which are particularly good. By (a) identifying these preference structures via Bayesian subspace learning, then (b) incorporating this into an information-theoretic sampling procedure, MaxEnt offers an effective way of learning the full rating database from partial observations.

5. Conclusion

In this paper, we propose a novel information-theoretic sampling method for noisy matrix completion. Using the Bayesian SMG (Yuchi et al., 2021) as a probabilistic model for the unknown low-rank matrix, we presented an initial and sequential sampling algorithm called MaxEnt, which makes use of subspace learning to guide an information-theoretic sampling of matrix \mathbf{X} . Simulations and applications demonstrate the effectiveness of MaxEnt over uniform sampling, and confirm insights developed in the paper.

References

- Bhargava, A., Ganti, R., and Nowak, R. Active positive semidefinite matrix completion: Algorithms, theory and applications. In *Artificial Intelligence and Statistics*, pp. 1349–1357, 2017.
- Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- Candès, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- Candès, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- Carson, W. R., Chen, M., Rodrigues, M. R., Calderbank, R., and Carin, L. Communications-inspired projection design with application to compressive sensing. *SIAM Journal on Imaging Sciences*, 5(4):1185–1212, 2012.
- Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- Chikuse, Y. *Statistics on Special Manifolds*. Springer Science & Business Media, 2012.
- Colbourn, C. J., Klove, T., and Ling, A. C. Permutation arrays for powerline communication and mutually orthogonal Latin squares. *IEEE Transactions on Information Theory*, 50(6):1289–1291, 2004.
- Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.
- Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.
- Fisher, R. A. *The Design of Experiments*. Oliver and Boyd, London, 1937.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*, volume 2. CRC Press, 2014.
- Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.
- Goldman, S. A. and Warmuth, M. K. Learning binary relations using weighted majority voting. *Machine Learning*, 20(3):245–271, 1995.
- Gupta, A. K. and Nagar, D. K. *Matrix Variate Distributions*. CRC Press, 1999.
- Hoff, P. D. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.
- Huczynska, S. Powerline communication and the 36 officers problem. *Philosophical Transactions of the Royal Society of London A*, 364(1849):3199–3214, 2006.
- Jacobson, M. T. and Matthews, P. Generating uniformly distributed random Latin squares. *Journal of Combinatorial Designs*, 4(6):405–437, 1996.
- Keedwell, A. D. and Dénes, J. *Latin Squares and Their Applications*. Elsevier, 2015.
- Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.
- Koltchinskii, V., Lounici, K., Tsybakov, A. B., et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
- Lan, C., Deng, Y., and Huan, J. A disagreement-based active matrix completion approach with provable guarantee. In *International Joint Conference on Neural Networks*, pp. 4082–4088. IEEE, 2016.
- MacKay, D. J. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.
- Mak, S. and Xie, Y. Maximum entropy low-rank matrix recovery. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):886–901, 2018.
- Nakamura, A. and Abe, N. Online learning of binary and n-ary relations over clustered domains. *Journal of Computer and System Sciences*, 65(2):224–256, 2002.
- Natarajan, N. and Dhillon, I. S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):60–68, 2014.
- Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.
- Palomar, D. P. and Verdú, S. Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, 2006.

- Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.
- Ruchansky, N., Crovella, M., and Terzi, E. Matrix completion with queries. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1034. ACM, 2015.
- Shen, J. On the singular values of Gaussian random matrices. *Linear Algebra and Its Applications*, 326(1-3):1–14, 2001.
- Shewry, M. C. and Wynn, H. P. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.
- Shlezinger, N., Dabora, R., and Eldar, Y. C. Measurement matrix design for phase retrieval based on mutual information. *IEEE Transactions on Signal Processing*, 66(2): 324–339, 2017.
- Sutherland, D. J., Póczos, B., and Schneider, J. Active learning and search on low-rank matrices. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pp. 212–220, 2013.
- Wang, L., Razi, A., Rodrigues, M., Calderbank, R., and Carin, L. Nonlinear information-theoretic compressive measurement design. In *International Conference on Machine Learning*, pp. 1161–1169, 2014.
- Yuchi, H. S., Mak, S., and Xie, Y. Bayesian uncertainty quantification for low-rank matrix completion. *arXiv preprint arXiv:2101.01299*, 2021.

A. Appendix

A.1. BayeSMG: Posterior sampling for UQ

In this section, we introduce the posterior sampling algorithm (Yuchi et al., 2021) for quantifying uncertainty on \mathbf{X} utilized in 3.4. For efficient sampling, we require a slight reparametrization of \mathbf{X} via its SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Define the *Stiefel manifold* $\mathcal{V}_{R,m}$, the space of $m \times R$ matrices with orthonormal columns (an R -frame in \mathbb{R}^m). By the SVD, the matrix of left and right singular vectors, \mathbf{U} and \mathbf{V} , must lie on \mathcal{V}_{R,m_1} and \mathcal{V}_{R,m_2} . Note that the span of an R -frame from the Stiefel manifold $\mathcal{V}_{R,m}$ corresponds to a unique R -plane from the Grassmann manifold $\mathcal{G}_{R,m-R}$, but an R -plane from $\mathcal{G}_{R,m-R}$ corresponds to infinitely many R -frames from $\mathcal{V}_{R,m}$. For our model in Section 2.2 with the priors configured as in (2), random matrix theory (Shen, 2001) then shows: (a) \mathbf{U} and \mathbf{V} are uniformly distributed on \mathcal{V}_{R,m_1} and \mathcal{V}_{R,m_2} , and (b) $\mathbf{D} = \text{diag}(\{d_k\}_{k=1}^R)$ follows the so-called *Quadrant Law* (QL; (Shen, 2001)). The uniform distributions on \mathcal{V}_{R,m_1} and \mathcal{V}_{R,m_2} are special cases of the *von Mises-Fisher* (MF) distributions $MF(m_1, R, \mathbf{0})$ and $MF(m_2, R, \mathbf{0})$; a random matrix $\mathbf{W} \sim MF(m, R, \mathbf{F})$ has density (Hoff, 2009):

$$[\mathbf{W}|R, \mathbf{F}] = \left[{}_0F_1\left(\frac{m}{2}; \frac{\mathbf{F}^T \mathbf{F}}{4}\right) \right]^{-1} \text{etr}(\mathbf{F}^T \mathbf{W}), \quad \mathbf{W} \in \mathcal{V}_{R,m}, \quad (9)$$

where ${}_0F_1(\cdot; \cdot)$ is the hypergeometric function. The singular values \mathbf{D} follow $QL(\mathbf{0}, \sigma^2)$, where $QL(\boldsymbol{\mu}, \delta^2)$ is the quadrant law with density:

$$[\mathbf{D}|\boldsymbol{\mu}, \delta^2] = \frac{\exp\left\{-\frac{1}{2\delta^2} \sum_{k=1}^R (d_k - \mu_k)^2\right\}}{Z_R(2\pi\delta^2)^{R/2}} \prod_{k,l=1; k < l}^R |d_k^2 - d_l^2|, \quad (10)$$

and Z_R is a normalization constant depending on R .

Algorithm 2 BayeSMG

Input: $\mathbf{Y}_\Omega, R, \alpha_{\eta^2}, \beta_{\eta^2}, \alpha_{\sigma^2}, \beta_{\sigma^2}$,
 Complete \mathbf{X}_0 from \mathbf{Y}_Ω via (3).
 Initialize $[\mathbf{U}_0, \mathbf{D}_0, \mathbf{V}_0] \leftarrow \text{svd}(\mathbf{X}_0)$, η_0^2 and σ_0^2 .
Gibbs sampler:
for $t = 1$ **to** T **do**
 $\mathbf{X}_t \leftarrow \mathbf{U}_{t-1} \mathbf{D}_{t-1} \mathbf{V}_{t-1}^T$.
 $\mathbf{Y}_{\Omega^c} \sim \mathcal{N}(\mathbf{X}_{\Omega^c}^P, \boldsymbol{\Sigma}_{\Omega^c}^P + \eta^2 \mathbf{I})$.
 $\mathbf{U}_t \sim MF(m_1, R, \mathbf{Y} \mathbf{V}_{t-1} \mathbf{D}_{t-1} / \eta_{t-1}^2)$.
 $\mathbf{V}_t \sim MF(m_2, R, \mathbf{Y}^T \mathbf{U}_t \mathbf{D}_{t-1} / \eta_{t-1}^2)$.
 $\mathbf{D}_t \sim QL(\boldsymbol{\mu}, \delta^2)$,
 where $\boldsymbol{\mu} = [\sigma_{t-1}^2 \mathbf{u}_{k,t}^T \mathbf{Y} \mathbf{v}_{k,t} / (\eta_{t-1}^2 + \sigma_{t-1}^2)]_{k=1}^R$
 $\delta^2 = \eta_{t-1}^2 \sigma_{t-1}^2 / (\eta_{t-1}^2 + \sigma_{t-1}^2)$.
 $\sigma_t^2 \sim IG(\alpha_{\sigma^2} + R/2, \beta_{\sigma^2} + \text{tr}(\mathbf{D}_t^2) / 2)$
 $\eta_t^2 \sim IG(\alpha_{\eta^2} + m_1 m_2 / 2, \beta_{\eta^2} + \|\mathbf{Y} - \mathbf{X}_t\|_F^2 / 2)$.
end for
 Return posterior samples $\{(\mathbf{X}_t, \mathbf{U}_t, \mathbf{V}_t)\}_{t=1}^T$.
