Polynomial Matrix Completion for Missing Data Imputation and Transductive Learning (Supplementary Material)

Jicong Fan, Yuqian Zhang, Madeleine Udell
Cornell University, Ithaca, NY 14853, USA
{jf577, yz2557, udell}@cornell.edu

Proof for Lemmas and Theorems

Proof for Theorem 1

Theorem 1. *Suppose $\boldsymbol{X}$ is given by Assumption 1 with $\|\boldsymbol{Z}\|_\infty \leq c_z$. Then for any q' obeying $\binom{d+q'}{q'} \leq \min\{l, n\}$, there exists a matrix $\boldsymbol{\Phi}$ with rank at most $\binom{d+q'}{q'}$, such that*

$$\|\phi(\boldsymbol{X}) - \boldsymbol{\Phi}\|_\infty \leq c_{q'}, \tag{1}$$

where $c_{q'} = \frac{c_z^{q'+1}}{(q'+1)!} \max_{\|\boldsymbol{z}\|_\infty \leq c_z} \|(\phi \circ f)^{(q'+1)}(\boldsymbol{z})\|_\infty$.

Proof. Denoting $\phi \circ f$ by h, we have $\phi(\boldsymbol{x}) = h(\boldsymbol{z}) = [h_1(\boldsymbol{z}), h_2(\boldsymbol{z}), \ldots, h_l(\boldsymbol{z})]^T$. h is analytic and hence has convergent Taylor expansion. The q' order Taylor expansion of $h_i(\boldsymbol{z})$ at $\boldsymbol{a}$ is given by

$$\begin{aligned} h_i(\boldsymbol{z}) &= h_i(a) + \sum_{j=1}^{d} \frac{\partial h_i^{(1)}(\boldsymbol{a})}{\partial z_j}(z_j - a_j) \\ &+ \frac{1}{2!}\sum_{j_1=1}^{d}\sum_{j_2=1}^{d} \frac{\partial h_i^{(2)}(\boldsymbol{a})}{\partial z_{j_1} z_{j_2}}(z_{j_1} - a_{j_1})(z_{j_2} - a_{j_2}) + \cdots \\ &+ \frac{1}{q'!}\sum_{j_1=1}^{d} \cdots \sum_{j_{q'}=1}^{d} \frac{\partial h_i^{(q')}(\boldsymbol{a})}{\partial z_{j_1} \cdots z_{j_{q'}}}(z_{j_1} - a_{j_1}) \cdots (z_{j_{q'}} - a_{j_{q'}}) \\ &+ c_{i,q'}, \end{aligned} \tag{2}$$

where $c_{i,q'}$ denotes the residual of Taylor expansion. In addition,

$$|c_{i,q'}| \leq \frac{\|\boldsymbol{z}\|_\infty^{q'+1}}{(q'+1)!} \max_{|z| \leq \|\boldsymbol{z}\|_\infty + \|\boldsymbol{a}\|_\infty} |(\phi \circ f)_{ij}^{(q'+1)}(z)|.$$

For convenience, we set $\boldsymbol{a} = 0$. Let $\tilde{\boldsymbol{z}}$ be the vector consisting of $\{z_1^{\mu_1} z_2^{\mu_2} \cdots z_d^{\mu_d}\}_{|\mu| \leq q'}$. Then $\tilde{\boldsymbol{z}} \in \mathbb{R}^{\tilde{d}'}$, where $\tilde{d}' = 1 + \sum_{i=1}^{q'} \binom{d+i-1}{i} = \binom{d+q'}{q'}$. Now (2) can be rewritten as

$$h_i(\boldsymbol{z}) = \boldsymbol{\theta}_i^T \tilde{\boldsymbol{z}} + c_{i,q'}, \tag{3}$$

where $\boldsymbol{\theta}_i \in \mathbb{R}^{\tilde{d}'}$ consists of the corresponding coefficients. Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l]^T \in \mathbb{R}^{l \times \tilde{d}'}$ and $\tilde{\boldsymbol{Z}} = [\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \ldots, \tilde{\boldsymbol{z}}_n] \in \mathbb{R}^{\tilde{d}' \times n}$. Then for all $1 \leq i \leq l$ and $1 \leq j \leq n$,

$$\left|[h(\boldsymbol{Z}) - \boldsymbol{\Theta}\tilde{\boldsymbol{Z}}]_{ij}\right| \leq c_{q'}, \tag{4}$$

where $c_{q'} = \frac{c_z^{q'+1}}{(q'+1)!} \max_{\|\boldsymbol{z}\|_\infty \leq c_z} \|(\phi \circ f)^{(q'+1)}(\boldsymbol{z})\|_\infty$.

Denote $\boldsymbol{\Phi} = \boldsymbol{\Theta}\tilde{\boldsymbol{Z}}$ and $\text{rank}(\boldsymbol{\Phi}) = \tilde{d}$. We have

$$\|\phi(\boldsymbol{X}) - \boldsymbol{\Phi}\|_\infty \leq c_{q'}. \tag{5}$$

Since $\boldsymbol{Z}$ is of full-rank, it is easy to show that $\tilde{\boldsymbol{Z}}$ is of full-rank. For a specific $\{f_i\}_{i=1}^m$, the rank of $\boldsymbol{\Theta}$ may be lower than $\tilde{d}'$ especially when some of the partial derivatives in (2) vanish or/and repeated. Therefore, $\tilde{d} \leq \tilde{d}' = \binom{d+q'}{q'}$. This finished the proof. □

Proof for Lemma 1

Lemma 1. *Suppose $\boldsymbol{X}$ is given by Assumption 2. Then*

$$\text{rank}(\boldsymbol{X}) \leq \min\{\tbinom{d+\alpha}{\alpha}, m, n\}$$

and

$$\text{rank}(\phi(\boldsymbol{X})) \leq \min\{\tbinom{d+\alpha q}{\alpha q}, \tbinom{m+q}{q}, n\}.$$

Proof. It is easy to have $\text{rank}(\boldsymbol{X}) \leq \min\{\binom{d+\alpha}{\alpha}, m, n\}$. In addition, $\phi(\boldsymbol{X}) \in \mathbb{R}^{l \times n}$, where $l = \binom{m+\bar{q}}{\alpha}$. As $\phi(\boldsymbol{x}) = \phi(f(\boldsymbol{z}))$ is a αq order polynomial mapping of $\boldsymbol{z}$, the number of polynomial features of $\boldsymbol{z}$ is at most $\binom{d+\alpha q}{\alpha q}$. Therefore $\text{rank}(\phi(\boldsymbol{X})) \leq \min\{\binom{d+\alpha q}{\alpha q}, \binom{m+q}{q}, n\}$. □

Proof for Lemma 2

Lemma 2. *For any $\boldsymbol{X}$ given by Assumption 1 and any positive integer α, there exists an $\tilde{\boldsymbol{X}}$ obeying Assumption 2, such that*

$$\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_\infty \leq c_\alpha, \tag{6}$$

where $c_\alpha = \frac{c_z^{\alpha+1}}{(\alpha+1)!} \max_{\|\boldsymbol{z}\|_\infty \leq c_z} \|f^{(\alpha+1)}(\boldsymbol{z})\|_\infty$.

Proof. Lemma 2 is a special case of Theorem 1. The proof can be obtained via letting $\phi(\boldsymbol{x}) = \boldsymbol{x}$ in the proof of Theorem 1. □

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Proof for Theorem 2

Theorem 2. *Let* $\text{mnp}_\phi(\boldsymbol{X})$ *be the minimum number of parameters required to determine* $\boldsymbol{X}$ *uniquely among all matrices in the set* $\{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m\times n} : \text{rank}(\phi(\boldsymbol{\mathcal{X}})) = \tilde{d}\}$. *Define* $\tilde{r} := \min\{o : \binom{o+q}{q} \geq \tilde{d}\}$. *Then*

$$\text{mnp}_\phi(\boldsymbol{X}) = (m - \tilde{r}) \times \tilde{d} + n\tilde{r}. \tag{7}$$

Proof. As the first-order feature in $\phi(\boldsymbol{X})$ is $\boldsymbol{X}$ itself and the rank of $\phi(\boldsymbol{X})$ is $\tilde{d}$, then there exists a $\boldsymbol{W} \in \mathbb{R}^{m\times\tilde{d}}$ such that

$$\boldsymbol{X} = \boldsymbol{W}\phi_{\tilde{d}}(\boldsymbol{X}), \tag{8}$$

where $\phi_{\tilde{d}}(\boldsymbol{X})$ denotes the corresponding $\tilde{d}$ rows of $\phi(\boldsymbol{X})$. Suppose $\boldsymbol{X}_o$ consists of o rows of $\boldsymbol{X}$ and $\boldsymbol{X}_{m-o}$ consists of the remainders. Let o be sufficiently large such that the number of rows of $\phi(\boldsymbol{X}_o)$ is no less than $\tilde{d}$:

$$\binom{o+q}{q} \geq \tilde{d}. \tag{9}$$

Then it follows from (8) that

$$\boldsymbol{X}_{m-o} = \boldsymbol{W}_{m-o}\phi(\boldsymbol{X}_o), \tag{10}$$

for some $\boldsymbol{W}_{m-o} \in \mathbb{R}^{(m-o)\times\tilde{d}}$. We see that $\boldsymbol{X}_{m-o}$ can be reconstructed from $\boldsymbol{X}_o$. In addition, since $\boldsymbol{X}_o = f_o(\boldsymbol{Z})$ (according to Assumption 2), we have

$$\boldsymbol{X}_{m-o} = \boldsymbol{W}_{m-o}(\phi(f_o(\boldsymbol{Z})) := \hat{f}_{m-o}(\boldsymbol{Z}). \tag{11}$$

Therefore, minimizing the rank of $\phi(\boldsymbol{X})$ implicitly approximates the unknown mapping f.

According to (10), we can count the minimum number of parameters required to determine $\boldsymbol{X}$ uniquely among all matrices in the set $\{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m\times n} : \text{rank}(\phi(\boldsymbol{\mathcal{X}})) = \tilde{d}\}$. First, we denote the minimum o satisfying (9) as

$$\tilde{r} := \min\{o : \binom{o+q}{q} \geq \tilde{d}\}, \tag{12}$$

and have

$$\left(q!\tilde{d}\right)^{1/q} - q \leq \tilde{r} \leq \left(q!\tilde{d}\right)^{1/q}, \tag{13}$$

which means we can compute $\tilde{r}$ via at most $q+1$ trials. Then $\boldsymbol{X}_{\tilde{r}}$ counts $n\tilde{r}$ parameters. Second, there are $(m-\tilde{r})\times\tilde{d}$ parameters for $\boldsymbol{W}_{m-\tilde{r}}$ to reconstruct $\boldsymbol{X}_{m-\tilde{r}}$ from $\boldsymbol{X}_{\tilde{r}}$ and ϕ has been given. Therefore, the minimum number of parameters we required is

$$\text{mnp}_\phi(\boldsymbol{X}) := (m - \tilde{r}) \times \tilde{d} + n\tilde{r}.$$

□

Proof for Lemma 3

Lemma 3. *For any* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$,

$$\|\phi(\boldsymbol{X})\|^p_{S_p|s} = \text{Tr}(\mathcal{K}(\boldsymbol{X})^{p/2}) - \max_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}_s, \boldsymbol{P}\in\mathbb{R}^{n\times s}} \text{Tr}\Big(\big(\boldsymbol{P}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{P}\big)^{p/2}\Big). \tag{14}$$

Proof. Denote $\tilde{\mathcal{K}}(\boldsymbol{X}) = \boldsymbol{P}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{P}$. Let λ_i be the i-th eigenvalue of $\mathcal{K}(\boldsymbol{X})$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Let $\tilde{\lambda}_i$ be the i-th eigenvalue of $\tilde{\mathcal{K}}(\boldsymbol{X})$ and $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n$. Then

$$\text{Tr}\Big(\big(\boldsymbol{P}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{P}\big)^{p/2}\Big) = \sum_{i=1}^{s} \tilde{\lambda}_i^{p/2}.$$

Using the Lemma 3.31(a) of (Horn and Johnson 1991), we have

$$\tilde{\lambda}_i \leq \lambda_i,$$

where $i = 1, 2, \ldots, s$. It follows that

$$\sum_{i=1}^{s} \tilde{\lambda}_i^{p/2} \leq \sum_{i=1}^{s} \lambda_i^{p/2}.$$

Therefore,

$$\max_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}_s} \text{Tr}\Big(\big(\boldsymbol{P}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{P}\big)^{p/2}\Big) = \sum_{i=1}^{s} \lambda_i^{p/2},$$

and

$$\begin{aligned}\|\phi(\boldsymbol{X})\|^p_{S_p|s} &= \text{Tr}(\mathcal{K}(\boldsymbol{X})^{p/2}) - \sum_{i=1}^{s} \lambda_i^{p/2} \\ &= \sum_{i=s+1}^{n} \lambda_i^{p/2} = \sum_{i=s+1}^{n} \sigma_i^p.\end{aligned}$$

This finished the proof. □

Proof for Lemma 4

Lemma 4. *For any* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$,

$$\|\phi(\boldsymbol{X})\|^p_{S_p|\boldsymbol{w}} = \min_{\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{Q}\boldsymbol{Q}^T=\boldsymbol{I}_n} \text{Tr}\Big(\big(\boldsymbol{W}^{1/p}\boldsymbol{Q}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{Q}\boldsymbol{W}^{1/p}\big)^{p/2}\Big).$$

Proof. Denote $\bar{\mathcal{K}}(\boldsymbol{X}) = \boldsymbol{Q}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{Q}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_n$ be the eigenvalues of $\mathcal{K}(\boldsymbol{X})$ and $\bar{\mathcal{K}}(\boldsymbol{X})$ respectively. Since $\boldsymbol{Q}$ is an orthogonal matrix, we have $\bar{\lambda}_i = \lambda_i$ for all $i = 1, 2, \ldots, n$. It follows that

$$\begin{aligned}&\text{Tr}\Big(\big(\boldsymbol{W}^{1/p}\boldsymbol{Q}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{Q}\boldsymbol{W}^{1/p}\big)^{p/2}\Big) \\ &\geq \text{Tr}\big(\boldsymbol{W}^{1/2}\bar{\mathcal{K}}(\boldsymbol{X})^{p/2}\boldsymbol{W}^{1/2}\big) \\ &= \text{Tr}\big(\boldsymbol{W}\bar{\mathcal{K}}(\boldsymbol{X})^{p/2}\big) \\ &\geq \sum_{i=1}^{n} w_i\bar{\lambda}_i^{p/2} = \sum_{i=1}^{n} w_i\lambda_i^{p/2} \\ &= \sum_{i=1}^{n} w_i\sigma_i^p = \|\phi(\boldsymbol{X})\|^p_{S_p|\boldsymbol{w}},\end{aligned} \tag{15}$$

where the first inequality holds due to the Araki–Lieb–Thirring inequality (Araki 1990) and the second inequality holds due to the product trace inequality in (Marshall, Olkin, and Arnold 2011) (pages 340-341). This finished the proof. □

Table 1: Datasets for classification

Dataset	# features	# samples	# classes	missing value
Mice protein	82	1080	8	1.68%
Shuttle	9	2175	5	10%
Dermatology	34	366	6	1%
Satimage	36	6435	7	10%

Table 2: Classification errors

Data set	θ	SVM	LRMC	LRMC+SVM	SRMC	VMC-2	VMC-3	NLMC	PMC-S	PMC-W
Mice protein	10%	8.96	6.33	0.8	2.96	0.54	0.46	0.44	0.41	**0.39**
	30%	21.41	10.44	2.59	5.65	0.61	0.48	0.44	0.44	**0.33**
	50%	32.5	18.78	5.26	13.19	1.24	0.87	0.81	0.71	**0.63**
	70%	46.7	34.02	19.2	29.8	8.28	5.15	4.06	**2.43**	2.63
Shuttle	10%	12.18	24.7	**2.48**	21.06	9.7	7.72	4.8	2.66	3.86
	30%	14.74	25.9	6.4	22.8	11.04	8.7	6.12	**4.32**	5.56
	50%	17.82	28.7	10.6	27.3	13.4	11.1	9.58	**8.02**	9.16
	70%	19.7	38.12	19.76	40.28	18.08	16.16	16.24	**14.8**	15.82
Dermatology	10%	4.48	4.54	3.28	4.21	3.17	3.12	3.08	**2.84**	**2.84**
	30%	6.83	6.56	5.96	6.01	5.41	5.3	5.14	5.03	**4.92**
	50%	13.07	9.95	8.83	9.08	8.64	8.74	8.31	8.16	**7.98**
	70%	26.67	21.09	20.87	21.31	19.67	20.08	19.13	**18.5**	18.87
Satimage	10%	39.6	23.34	14.38	20.62	17.32	15.5	14.7	**13.06**	14.24
	30%	41.46	23.58	16.22	22.52	17.86	16.12	15.2	**14.56**	14.94
	50%	44.2	24.24	16.96	24.1	18.44	16.18	15.84	**14.82**	15.18
	70%	53.58	27.24	22.62	26.5	21.92	18.86	17.86	**16.64**	17.3

More details about the experiments

Parameter setting

In our PMC-S and PMC-W, we use RBF kernel because it is often more effective than polynomial kernel. In VMC-2, VMC-3 (Ongie et al. 2017), NLMC (Fan and Chow 2018), PMC-S, and PMC-W, for the Schatten-p norm, we set $p = 1$ or 0.5 and report the better result. In VMC-2 and VMC-3, the parameter a of polynomial kernel is chosen from $\{1, 10\}$. In NLMC, PMC-S, and PMC-W, the hyper-parameter σ of RBF kernel is set as the average distance of pair-wise data points multiplied by 1 or 3. The hyper-parameters in LRMC, SRMC (Fan and Chow 2017) and LADMC (Ongie et al. 2018) (the iterative algorithm) are carefully determined to provide the best performance as possible.

Datasets with missing values for classification

As shown in Table 1, we consider four real datasets with missing values[1]. For each dataset, we randomly remove a fraction $(0.1, 0.3, 0.5, \text{or } 0.7)$ of the observed entries of the feature matrix and perform classification. In this study, we use a subset of Shuttle/Satimage dataset, which consists of 1000 randomly chosen samples. In addition, the proportion of training (labeled) data is 50%. Since the features of each dataset often have different scale, we normalize the features to have zero mean and unit variance. We use matrix completion to recover the missing entries and classify the data simultaneously. As a well-known baseline of classification technique, SVM with RBF kernel is compared, where the missing entries are replaced by zeros. Moreover, we also perform SVM on the data recovered by LRMC.

The classification results are shown in Table 2. Our PMC-S and PMC-W outperform other methods in almost all cases.

[1]https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression
https://sci2s.ugr.es/keel/missing.php

References

Araki, H. 1990. On an inequality of Lieb and Thirring. *Letters in Mathematical Physics* 19(2):167–170.

Fan, J., and Chow, T. W. 2017. Matrix completion by least-square, low-rank, and sparse self-representations. *Pattern Recognition* 71:290 – 305.

Fan, J., and Chow, T. W. 2018. Non-linear matrix completion. *Pattern Recognition* 77:378 – 394.

Horn, R. A., and Johnson, C. R. 1991. *Topics in Matrix Analysis*. Cambridge University Press.

Marshall, A. W.; Olkin, I.; and Arnold, B. C. 2011. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York.

Ongie, G.; Willett, R.; Nowak, R. D.; and Balzano, L. 2017. Algebraic variety models for high-rank matrix completion. In *Proceedings of the 34th International Conference on Machine Learning*, 2691–2700. Sydney, Australia: PMLR.

Ongie, G.; Balzano, L.; Pimentel-Alarcón, D.; Willett, R.; and Nowak, R. D. 2018. Tensor Methods for Nonlinear Matrix Completion. *ArXiv e-prints*.