Polynomial Matrix Completion for Missing Data Imputation and Transductive Learning

Jicong Fan Yuqian Zhang Madeleine Udell

Cornell University

Matrix completion

0.19	?	1.18	-0.77	?
?	0.85	?	-1.05	0.09
0.17	?	-0.21	-0.40	?
?	0.57	0.51	?	0.88
-0.37	-0.99	?	0.98	-0.63

→

0.19	1.33	1.18	-0.77	2.06
0.43	0.85	-0.49	-1.05	0.09
0.17	0.32	-0.21	-0.40	0.01
0.08	0.57	0.51	-0.33	0.88
-0.37	-0.99	0.03	0.98	-0.63

- Collaborative filtering (recommendation system)
- Classification (especially on incomplete data)
- Image inpainting
- Localization in wireless sensor networks

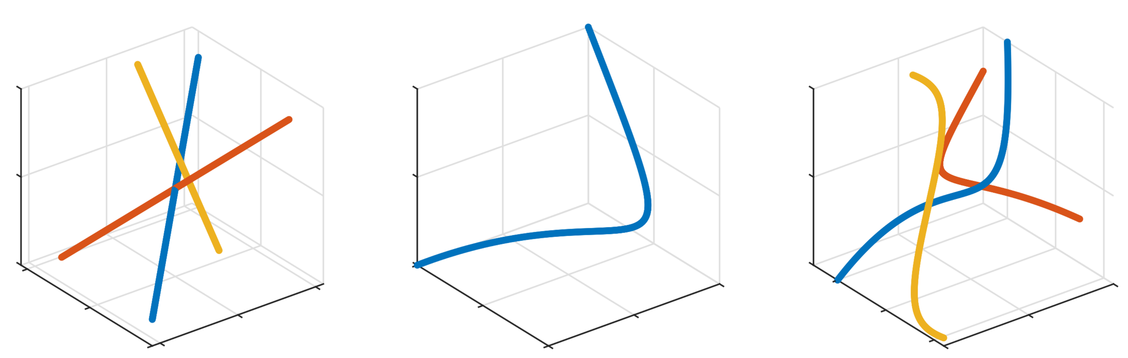
Conventional assumption: the decision matrix is low-rank

Limitation: not effective in recovering high-rank matrices

High-rank matrix with low-dimensional latent structure

Data generative models

- Union of low-dimensional subspaces
- Nonlinear manifold
- Union of nonlinear manifolds



Closely related problems

- Subspace clustering
- Manifold learning
- Nonlinear classification

Vulnerable to missing data

Analytic and polynomial generative model

Assumption 1 Suppose $d \ll m \ll n$, $\boldsymbol{Z} \in \mathbb{R}^{d\times n}$ is full-rank, $\|\boldsymbol{Z}\|_\infty \leqslant c_z$, and $f: \mathbb{R}^d \to \mathbb{R}^m$ is analytic. For $j = 1, 2, \ldots, n$, let $\boldsymbol{x}_j = f(\boldsymbol{z}_j)$ and form $\boldsymbol{X} \in \mathbb{R}^{m\times n}$.

Assumption 2 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is given by Assumption 1, in which f consists of polynomials of order at most α.

Example $f: \mathbb{R}^2 \to \mathbb{R}^6$, $\alpha = 3$, $\boldsymbol{x} = [z_1, z_2, z_1^2, z_1 z_2, z_2^3, z_1 z_2^2]^T$.

Rank property in the polynomial feature space

Polynomial (q-order) feature map: $\phi: \mathbb{R}^m \to \mathbb{R}^l$, $l = \binom{m+q}{q}$.

Example Suppose $m = 2$ and $q = 2$. Then $\phi(\boldsymbol{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]^T$.

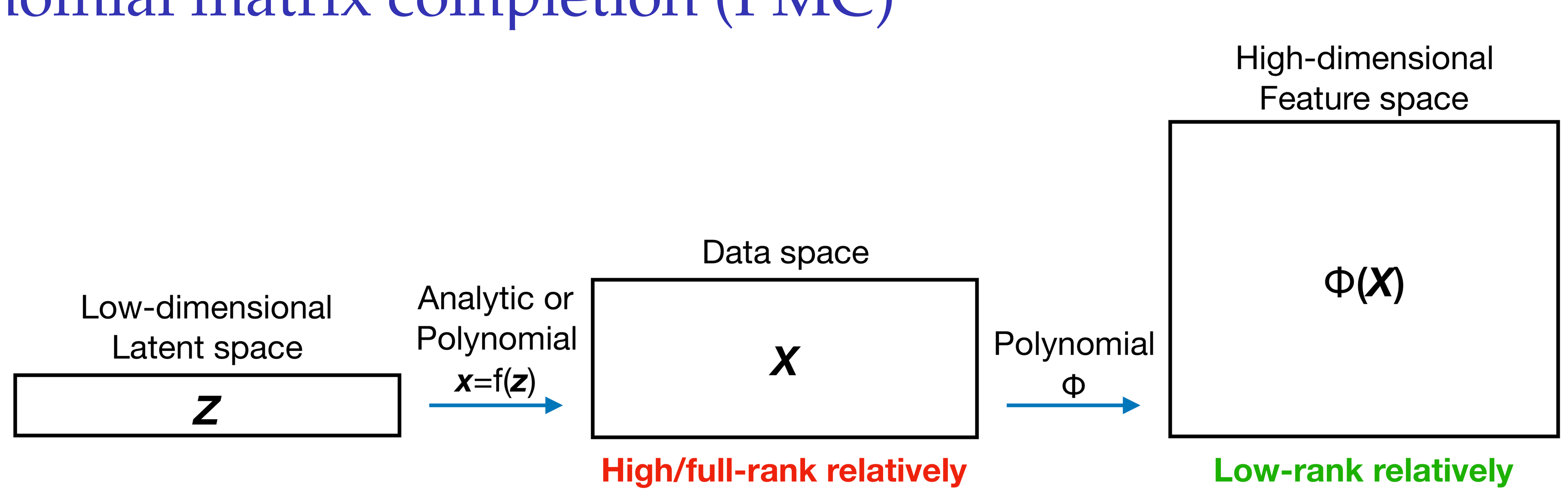
Theorem 1(informal) Suppose $\boldsymbol{X}$ is given by Assumption 1, n, q are large enough, and f is smooth enough. Then $\phi(\boldsymbol{X})$ is approximately low-rank.

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

$$\operatorname{rank}(\boldsymbol{X}) \leqslant \min\{\binom{d+\alpha}{\alpha}, m, n\} \quad \text{and} \quad \operatorname{rank}(\phi(\boldsymbol{X})) \leqslant \min\{\binom{d+\alpha q}{\alpha q}, \binom{m+q}{q}, n\}.$$

Example Let $m = 10$, $d = 2$, $n = 50$, $\alpha = 3$, and $q = 2$. We have $\boldsymbol{X} \in \mathbb{R}^{10\times 50}$, $\phi(\boldsymbol{X}) \in \mathbb{R}^{66\times 50}$, $\operatorname{rank}(\boldsymbol{X}) = 10$, and $\operatorname{rank}(\phi(\boldsymbol{X})) = 28$.

Polynomial matrix completion (PMC)



$$\text{minimize}_{\hat{\boldsymbol{X}}} \ \mathcal{R}(\phi(\hat{\boldsymbol{X}})), \quad \text{subject to } \mathcal{P}_\Omega(\hat{\boldsymbol{X}}) = \mathcal{P}_\Omega(\boldsymbol{X}).$$

Approximately rank minimization in feature space

Let $0 < p \leqslant 1$ and $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n$.

- Sum of p-th powers of singular values:

$$\mathcal{R}_1(\phi(\hat{\boldsymbol{X}})) := \|\phi(\hat{\boldsymbol{X}})\|_{S_p}^p = \sum_{i=1}^n \sigma_i^p(\phi(\hat{\boldsymbol{X}})) = \operatorname{Tr}\left(\mathcal{K}(\hat{\boldsymbol{X}})^{p/2}\right)$$

- - - - VMC(Ongie et al. 2017) and NLMC(Fan and Chow 2018)

- Sum of p-th powers of smaller singular values:

$$\mathcal{R}_2(\phi(\hat{\boldsymbol{X}})) := \|\phi(\hat{\boldsymbol{X}})\|_{S_p|s}^p = \sum_{i=s+1}^n \sigma_i^p(\phi(\hat{\boldsymbol{X}})), \quad s \leqslant \operatorname{rank}(\phi(\hat{\boldsymbol{X}}))$$
$$= \operatorname{Tr}(\mathcal{K}(\boldsymbol{X})^{p/2}) - \max_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}_s, \boldsymbol{P}\in\mathbb{R}^{n\times s}} \operatorname{Tr}\left(\left(\boldsymbol{P}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{P}\right)^{p/2}\right)$$

- - - - **PMC-S**

- Weighted sum of p-th powers of singular values:

$$\mathcal{R}_3(\phi(\hat{\boldsymbol{X}})) := \|\phi(\hat{\boldsymbol{X}})\|_{S_p|\boldsymbol{w}}^p = \sum_{i=1}^n w_i \sigma_i^p(\phi(\hat{\boldsymbol{X}})), \quad w_1 \leqslant w_2 \leqslant \cdots \leqslant w_n$$
$$= \min_{\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{Q}\boldsymbol{Q}^T=\boldsymbol{I}_n} \operatorname{Tr}\left(\left(\boldsymbol{W}^{1/p}\boldsymbol{Q}^T\mathcal{K}(\boldsymbol{X})\boldsymbol{Q}\boldsymbol{W}^{1/p}\right)^{p/2}\right)$$

- - - - - **PMC-W**

$[\mathcal{K}(\hat{\boldsymbol{X}})]_{ij} = \phi(\hat{\boldsymbol{x}}_i)^T\phi(\hat{\boldsymbol{x}}_j) = k(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$

– Polynomial kernel: $k^{\text{Poly}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + c)^q$

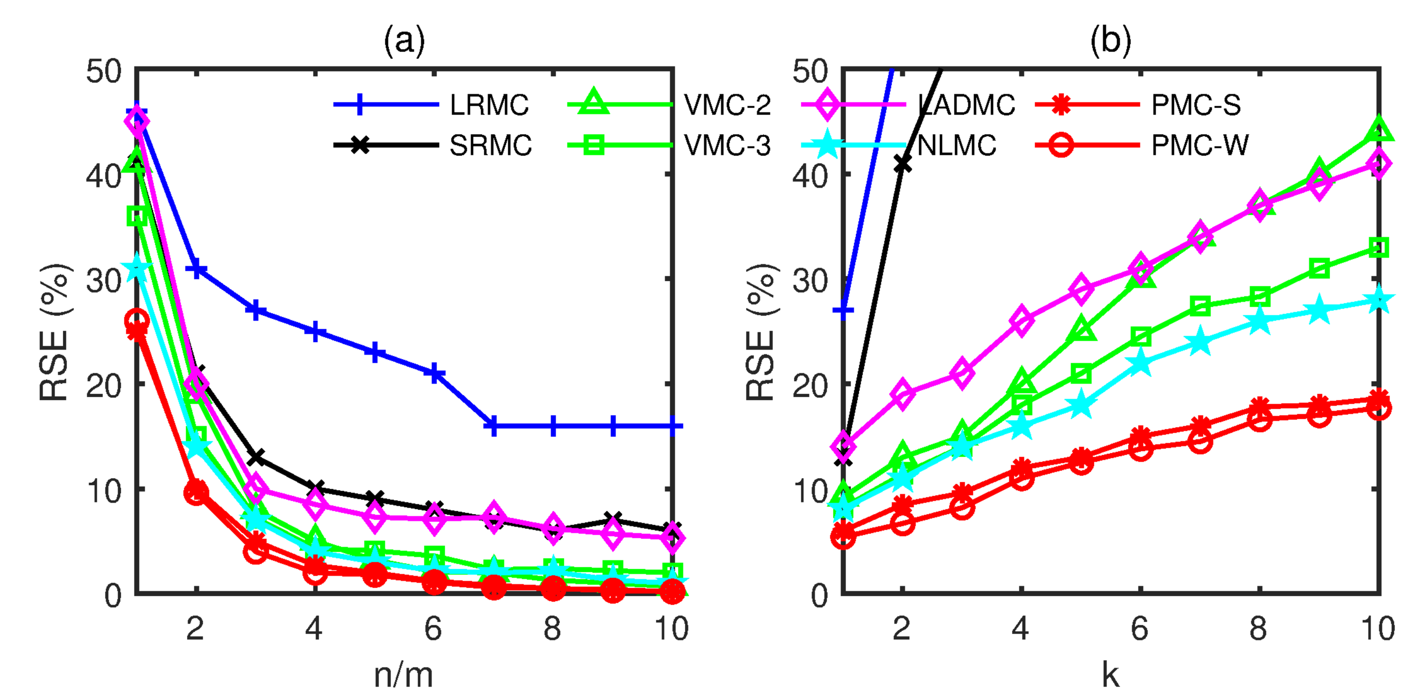
– Gaussian RBF kernel: $k^{\text{RBF}}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\sigma^2)\right)$

Optimization: Adam with adaptive step size

Recovery performance on synthetic data

(a) Single-manifold data ($\boldsymbol{X} \in \mathbb{R}^{m\times n}$): $\boldsymbol{X} = [\boldsymbol{X}_1] = \boldsymbol{A}_1\boldsymbol{Z}_1 + \frac{1}{2}(\boldsymbol{B}_1\boldsymbol{Z}_1^{\odot 2} + \boldsymbol{C}_1\boldsymbol{Z}_1^{\odot 3} + \boldsymbol{D}_1\boldsymbol{Z}_1^{\odot 4})$

(b) Multiple-manifold data: $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k] \in \mathbb{R}^{m\times kn}$



Proportion of missing entries: 50%

LRMC(Candès and Recht 2009)

SRMC(Fan and Chow 2017)

VMC(Ongie et al. 2017)

NLMC(Fan and Chow 2018)

LADMC(Ongie et al. 2018)

$\text{RSE} := \sqrt{\sum_{(i,j)\in\bar{\Omega}}(\boldsymbol{X}_{ij} - \hat{\boldsymbol{X}}_{ij})^2 / \sum_{(i,j)\in\bar{\Omega}} \boldsymbol{X}_{ij}^2}$

Transductive learning

Given l labeled data, classify u unlabeled data

- $\boldsymbol{x} \in \mathbb{R}^m$: feature vector $\boldsymbol{y} \in \mathbb{R}^k$: label vector
- $\boldsymbol{x}$ and $\boldsymbol{y}$ may have missing values

Step 1: Form a feature-label matrix

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_l & \boldsymbol{x}_{l+1} & \boldsymbol{x}_{l+2} & \ldots & \boldsymbol{x}_{l+u} \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \ldots & \boldsymbol{y}_l & ? & ? & \ldots & ? \end{bmatrix} \in \mathbb{R}^{(m+k)\times(l+u)}$$

Step 2: Recover the missing entries of $\boldsymbol{Z}$ for classification

Classification error (%) on incomplete data(θ: proportion of missing values):

Data set	θ	SVM	LRMC	LRMC+SVM	VMC -3	NLMC	**PMC -S**	**PMC -W**
Mice protein	10%	8.96	6.33	0.8	0.46	0.44	0.41	**0.39**
	50%	32.5	18.78	5.26	0.87	0.81	0.71	**0.63**
Shuttle	10%	12.18	24.7	**2.48**	7.72	4.8	2.66	3.86
	50%	17.82	28.7	10.6	11.1	9.58	**8.02**	9.16
Dermatology	10%	4.48	4.54	3.28	3.12	3.08	**2.84**	**2.84**
	50%	13.07	9.95	8.83	8.74	8.31	8.16	**7.98**
Satimage	10%	39.6	23.34	14.38	15.5	14.7	**13.06**	14.24
	50%	44.2	24.24	16.96	16.18	15.84	**14.82**	15.18