Collective Matrix Completion

Mokhtar Z. Alaya

Joint work with:

Olga Klopp

ESSEC & CREST

MLMDA Seminar, Borelli Center

June, 2022

ALLIANCE SORBONNE UNIVERSITÉ

utc Recherche LMAC

FSMP Fondation Sciences Mathématiques de Paris

MATH INNOV

îledeFrance

Outline

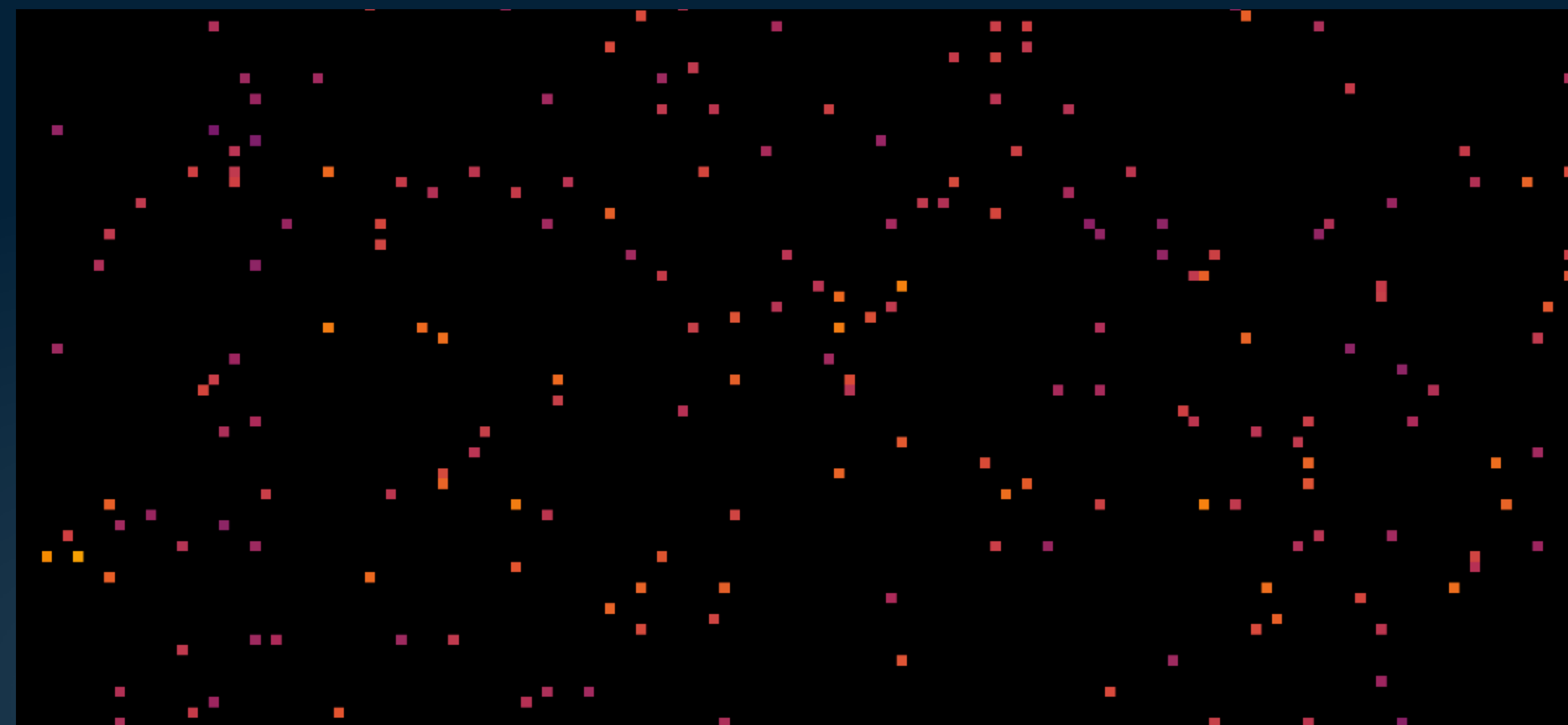
1. Overview of matrix completion

2. Collective matrix completion

3. Numerical experiments

Matrix completion is …

$$X =$$



- **Task:** given a partially observed data matrix X, predict the unobserved entries.

- Application to recommender systems, system identification, image processing, microarray data, etc.

Recommender systems, Netflix prize

- A popular example is the Netflix challenge (2006-2009).

$$X =$$

- Dataset: 480K users, 18K movies, 100M ratings.
- Only 1,1% of the matrix is filled!

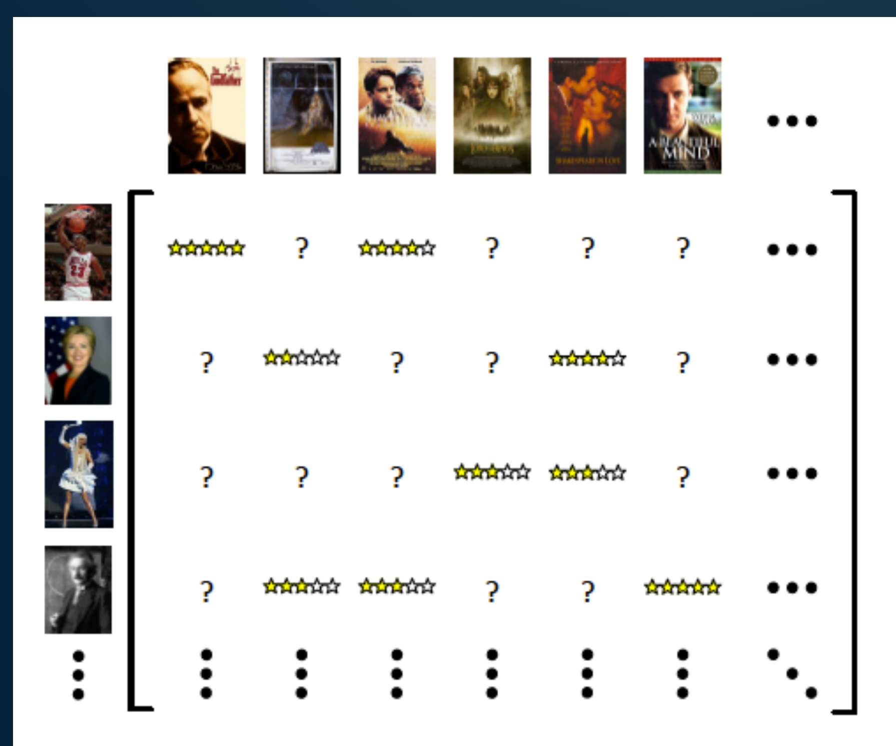
Some issues ...

- In general, we cannot infer missing ratings without any other information.

- This problem is under-determined, more unknown than observations (100M << 8.64M for Netflix).

- Low-rank assumption: fill matrix such that its rank is minimum.

→ A few factors explain most of the data.

$X =$  $\approx$ 

Low rank minimization

- Denote by Ω the set of entries of the matrix $\boldsymbol{X}$ that have been observed: we know the values $\boldsymbol{X}_{ij}$ for all $(i,j) \in \Omega$.

$$\underset{\boldsymbol{W}}{\text{minimize}}\, \text{rank}(\boldsymbol{W}) \quad \text{s. t. } W_{ij} = \underbrace{X_{ij}}_{\text{observed entries}}, \forall (i,j) \in \underbrace{\Omega}_{\text{sampling set}}.$$

- Or a slightly weaker version

$$\underset{\boldsymbol{W}}{\text{minimize}}\, \text{rank}(\boldsymbol{W}) \quad \text{s. t. } \sum_{(i,j)\in\Omega} (X_{ij} - W_{ij})^2 \le \delta.$$

- Or the regularization version

$$\underset{\boldsymbol{W}}{\text{minimize}} \sum_{(i,j)\in\Omega} (X_{ij} - W_{ij})^2 + \lambda\, \text{rank}(\boldsymbol{W}).$$

Low rank minimization

- Non-convex problem and combinatorially NP-hard!

$$\mathbf{rank}(\boldsymbol{X}) = \|\sigma(\boldsymbol{X})\|_0 = \sum_{i=1}^{\min\dim(\boldsymbol{X})} \underbrace{\mathbb{1}_{(\sigma_i(\boldsymbol{X})>0)}}_{i^{th}\ \text{largest singular value}}.$$

- Replace the ℓ_0 pseudo-norm by the ℓ_1-norm [Fazel ('02), Srebro et al ('05), Candès and Tao ('10), Negahban and Wainwright ('11), Davenport et al. ('14), Klopp ('14 and '15), ….].

Nuclear / trace / 1-Schatten norm:

$$\|\boldsymbol{X}\|_* = \|\sigma(\boldsymbol{X})\|_1 = \sum_{i=1}^{\min\dim(\boldsymbol{X})} \sigma_i(\boldsymbol{X}).$$

Nuclear norm minimization

- Hence temping to consider the nuclear norm minimization problem:

$$\underset{\boldsymbol{W}}{\text{minimize}} \|\boldsymbol{W}\|_* \quad \text{s. t.} \quad \sum_{(i,j)\in\Omega} (X_{ij} - W_{ij})^2 \le \delta.$$

- Or equivalently the regularization / Lagrangian formulation:

$$\underset{\boldsymbol{W}}{\text{minimize}} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - W_{ij})^2 + \lambda \|\boldsymbol{W}\|_*.$$

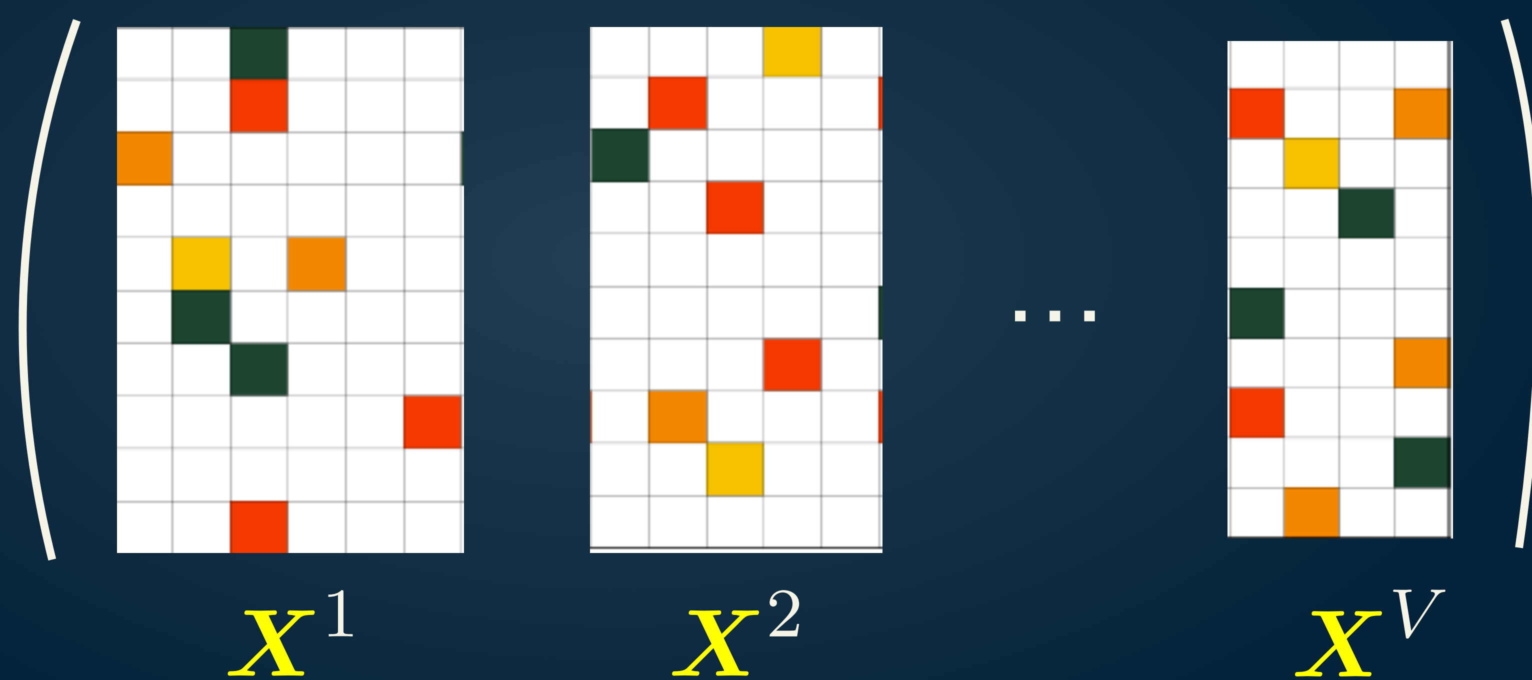
- This is convex problem.

Motivations of collective MC

- Data is often obtained from a collection of source matrices:

$$\mathcal{X} = (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^V)$$

$$\mathcal{X} =$$



- **Cold-Start problem**: in recommender systems, when a new user has no rating it is impossible to predict his ratings.

- Shared structure among the sources can be useful to get better predictions.

Collective MC: setup

- Each source view $\boldsymbol{X}^v \in \mathbb{R}^{d_u \times d_v}$ and $D = \sum_{v=1}^{V} d_v.$
- **Model**: let B_{ij}^v be independent Bernoulli random variables and

independent from X_{ij}^v with parameter π_{ij}^v. Setting:

$$\boldsymbol{Y}^v = \boldsymbol{B}^v \odot \boldsymbol{X}^v \quad \text{that is} \quad Y_{ij}^v = B_{ij}^v X_{ij}^v.$$

Collective MC: sampling scheme

- We consider **general sampling model** where we only assume that each entry is observed with a positive probability.

Assumption 1

There exists a positive constant $0 < p < 1$ such that

$$\min_{v \in [V]} \min_{(i,j) \in [d_u] \times [d_v]} \pi_{ij}^{v} \geq p.$$

[Klopp('15), Klopp et al. ('15), Lafond ('15), Cai and Zou ('16)]

Collective MC: sampling scheme

- Let $\pi^v_{i\bullet}$ (resp. $\pi^v_{\bullet j}$) the probability of sampling a coefficient from i-th row (resp. j-th column) of $\boldsymbol{X}^v$. Namely:

$$\pi^v_{i\bullet} = \sum_{j\in[d_v]} \pi^v_{ij} \quad \text{and} \quad \pi^v_{\bullet j} = \sum_{i\in[d_u]} \pi^v_{ij}. \text{ Let } \pi_{i\bullet} = \sum_{v\in[V]} \pi^v_{i\bullet}$$

Assumption 2

There exists a positive constant μ such that

$$\max_{v\in[V]} \max_{(i,j)\in[d_u]\times[d_v]} (\pi_{i\bullet}, \pi^v_{\bullet j}) \leq \mu.$$

[Klopp('15), Klopp et al. ('15), Lafond ('15), Cai and Zou ('16)]

Case 1: Exponential family noise

- We assume that the distribution of for each source $\boldsymbol{X}^v$ depends on the matrix of parameters $\boldsymbol{M}^v$ and satisfied a natural exponential family [Gunasekar et al. ('14); Cao and Xie ('16); Lafond ('15)]

$$X_{ij}^v | M_{ij}^v \sim f_{h^v, G^v}(X_{ij}^v | M_{ij}^v) = h^v(X_{ij}^v) \exp\left(X_{ij}^v M_{ij}^v - G^v(M_{ij}^v)\right).$$

Assumption 3

Assume that $G^v(\cdot)$ is twice differentiable and there exits two constants L_γ^2, U_γ^2

$$\sup_{\eta \in [-\gamma - \frac{1}{K}, \gamma + \frac{1}{K}]} (G^v)''(\eta) \le U_\gamma^2 \quad \text{and} \quad \inf_{\eta \in [-\gamma - \frac{1}{K}, \gamma + \frac{1}{K}]} (G^v)''(\eta) \ge L_\gamma^2$$

for some $K > 0$.

Exponential family noise: estimation procedure of $\mathcal{M} = (M^1, \ldots, M^V)$

- Given the observations $\mathcal{Y} = (\,\mathbf{Y}^1, \ldots, \,\mathbf{Y}^V)$, the normalized negative log-likelihood write as, for any $\mathcal{W} = (\mathbf{W}^1, \ldots, \mathbf{W}^V) \in \mathbb{R}^{d_u \times D}$,

$$\mathscr{L}_{\mathcal{Y}}(\mathcal{W}) = -\frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B^v_{ij}\big(\,Y^v_{ij} W^v_{ij} - G^v(W^v_{ij})\big)$$

- The nuclear norm penalized estimator $\widehat{\mathcal{M}}$ of $\mathcal{M}$ is defined as:

$$\widehat{\mathcal{M}} = (\widehat{M}^1, \ldots, \widehat{M}^V) = \underset{\mathcal{W} \in \mathscr{C}_\infty(\gamma)}{\operatorname{argmin}} \; \mathscr{L}_{\mathcal{Y}}(\mathcal{W}) + \lambda \|\mathcal{W}\|_*$$

where $\mathscr{C}_\infty(\gamma) = \big\{\mathcal{W} \in \mathbb{R}^{d_u \times D} : \|\mathcal{W}\|_\infty \leq \gamma\big\}$.

[Foygel et al. ('10), Salakhutdinov and Srebro ('10)]

Exponential family noise: theoretical guarantee

- Upper bound on the rescaled Frobenius estimation risk:

Theorem [A., Klopp 2019]

Assume that Assumptions 1, 2 and 3 hold and

$$\lambda = \mathcal{O}\left(\frac{(U_\gamma \vee K)\left(\sqrt{\mu} + (\log(d_u \vee D))^{3/2}\right)}{d_u D}\right).$$

Then, with probability exceeding $1 - 4/(d_u + D)$ one has,

$$\frac{1}{d_u D}\|\widehat{\mathcal{M}} - \mathcal{M}\|_F^2 \lesssim \frac{\operatorname{rank}(\mathcal{M})}{p^2 d_u D}\left(\gamma^2 + \frac{(U_\gamma \vee K)^2}{L_\gamma^4}\right)(\mu + \log^3(d_u \vee D))$$

$$\lesssim \frac{\operatorname{rank}(\mathcal{M})\mu}{p^2 d_u D}.$$

Exponential family noise: remarks

- For a close uniform sampling distribution, that is $c_1 p \le \pi^v_{ij} \le c_2 p$

$$\frac{1}{d_u D}\|\widehat{\mathcal{M}} - \mathcal{M}\|_F^2 \lesssim \frac{\operatorname{rank}(\mathcal{M})}{p(d_u \wedge D)}.$$

- Rate of convergence achieved by our estimator is faster compared to the penalization by the sum-nuclear-norm since

$$\operatorname{rank}(\mathcal{M}) \le \sum_{v=1}^{V} \operatorname{rank}(M^v).$$

- For small estimation error, one can choose $p \ge \operatorname{rank}(\mathcal{M})/(d_u \wedge D)$.

This implies $$n \gtrsim \operatorname{rank}(\mathcal{M})(d_u \vee D).$$

where $n = \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \pi^v_{ij}$ the expected number of observations

Case 2: Distribution-free-setting

- We do not assume any specific model for the observations.
- We consider the risk of estimating $\boldsymbol{X}^v$ with a loss function $\ell^v(\cdot,\cdot)$.

Assumption 4

For every v the loss function $\ell^v(y,\cdot)$ is ρ_v-Lipschitz in its second argument:

$$|\ell^v(y,x)-\ell^v(y,x')|\leq\rho_v|x-x'|.$$

Distribution-free-setting: estimation procedure

- For any matrix $\mathcal{Q} = (\boldsymbol{Q}^1, \ldots, \boldsymbol{Q}^V)$, we define the empirical risk as

$$R_{\mathcal{Y}}(\mathcal{Q}) = \frac{1}{d_u D} \sum_{v \in [V]} \sum_{(i,j) \in [d_u] \times [d_v]} B_{ij}^v \ell^v(Y_{ij}^v, Q_{ij}^v)$$

- We define the oracle as:

$$\overset{\star}{\mathcal{M}} = (\overset{\star}{M}^1, \ldots, \overset{\star}{M}^V) = \operatorname*{argmin}_{\mathcal{Q} \in \mathscr{C}_\infty(\gamma)} R(\mathcal{Q})$$

where $R(\mathcal{Q}) = \mathbb{E}[R_{\mathcal{Y}}(\mathcal{Q})]$.

- We consider excess risk $R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}})$.

Distribution-free-setting: estimation procedure

- For a tuning parameter $\Lambda > 0$ the nuclear norm penalized estimator reads as

$$\widehat{\mathcal{M}} \in \underset{\mathcal{Q} \in \mathscr{C}_\infty(\gamma)}{\operatorname{argmin}} \{ R_{\mathcal{Y}}(\mathcal{Q}) + \Lambda \|\mathcal{Q}\|_* \}$$

Assumption 5

There exists a constant $\varsigma > 0$ such that for every $\mathcal{Q} \in \mathscr{C}_\infty(\gamma)$, one has

$$R(\mathcal{Q}) - R(\overset{\star}{\mathcal{M}}) \geq \frac{\varsigma}{p d_u D} \|\mathcal{Q} - \overset{\star}{\mathcal{M}}\|_F^2$$

- Assumption 4 is called "Bernstein" condition [Mendelson (2008); Bartlett et al., (2004); Alquier et al., (2017); Elsener and van de Geer, (2018)].

Distribution-free-setting: theoretical gurantee

Theorem [A., Klopp 2019]

Assume that Assumptions 1, 2, 4 and 5 hold and set $\rho = \max_{v \in [V]} \rho_v$.

Let

$$\Lambda = \mathcal{O}\left(\frac{\rho\left(\sqrt{\mu} + \sqrt{\log(d_u \vee D)}\right)}{d_u D} \right).$$

Then, with probability exceeding $1 - 4/(d_u + D)$ one has,

$$R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}}) \lesssim \frac{\operatorname{rank}(\overset{\star}{\mathcal{M}})}{p} \frac{(\rho^2 + \rho^{3/2}\sqrt{\gamma/\varsigma})(\mu + \log(d_u \vee D))}{d_u D}$$

3. Numerical Experiments

Optimization of $\widehat{\mathcal{M}} = (\widehat{M}^1, \ldots, \widehat{M}^V) = \underset{\mathcal{W} \in \mathscr{C}_\infty(\gamma)}{\operatorname{argmin}} \mathscr{L}_{\mathcal{Y}}(\mathcal{W}) + \lambda \|\mathcal{W}\|_*$

- Proximal gradient (PG): [Beck and Teboulle ('09), Cai et al. ('09), Mazumder et al., ('10); Yao and Kwok ('15)]

- The PG generates a sequence of estimates

$$\mathcal{W}_{t+1} = \operatorname{prox}_{\frac{\lambda}{L}\|\cdot\|_*}(\mathcal{Z}_t), \text{ where } \mathcal{Z}_t = \mathcal{W}_t - \frac{1}{L}\nabla \mathscr{L}_{\mathcal{Y}}(\mathcal{W}_t)$$

- Assume a singular value decomposition $\mathcal{W} = \mathcal{U}\Sigma\mathcal{V}^\top$, then one has

$$\operatorname{prox}_{\frac{\lambda}{L}\|\cdot\|_*}(\mathcal{W}) = \mathrm{SVT}_{\lambda/L}(\mathcal{W}) = \mathcal{U}\operatorname{diag}((\sigma_1 - \lambda/L)_+, \ldots, (\sigma_r - \lambda/L)_+)\mathcal{V}^\top$$

[Cai et al. ('10)]

Power method to reduce complexity

- To compute $\mathcal{W}_{t+1}$ we need to perform an SVD of $\mathcal{Z}_t$ $\boxed{\mathcal{O}((d_u \wedge D) d_u D)}$
- We do not require to do a full SVD only a fewer k_t singular values of $\mathcal{Z}_t$ which are large than λ/L.
- As $\mathcal{W}_t$ converges to a low rank solution then k_t will be small during iterations.
- [Yao and Kwok ('15)] showed the following result:

$$\mathrm{SVT}_{\lambda/L}(\mathcal{Z}_t) = \mathcal{Q}\,\mathrm{SVT}_{\lambda/L}(\mathcal{Q}^\top \mathcal{Z}_t) \quad \boxed{\mathcal{O}(k_t d_u D)}$$

Algorithm 2: Power Method: `PowerMethod`$(\mathcal{Z}, \mathcal{R}, \epsilon)$

1. **input:** $\mathcal{Z} \in \mathbb{R}^{d_u \times D}$, initial $\mathcal{R} \in \mathbb{R}^{D \times k}$ for warm-start, tolerance δ;
2. **initialize** $\mathcal{W}_1 = \mathcal{Z}\mathcal{R}$;
3. **for** $t = 1, 2, \dots,$ **do**
4. $\quad \mathcal{Q}_{t+1} = \mathrm{QR}(\mathcal{W}_t)$;// QR denotes the QR factorization
5. $\quad \mathcal{W}_{t+1} = \mathcal{Z}(\mathcal{Z}^\top \mathcal{Q}_{t+1})$;
6. $\quad$ **if** $\|\mathcal{Q}_{t+1}\mathcal{Q}_{t+1}^\top - \mathcal{Q}_t \mathcal{Q}_t^\top\|_F \le \delta$ **then**
   $\quad\quad$ break;
7. **return** $\mathcal{Q}_{t+1}$.

[Halko et al ('11)]

23

Approximate SVT based on power method

Algorithm 3: Approximate SVT: $\texttt{Approx-SVT}(\mathcal{Z}, \mathcal{R}, \lambda, \delta)$

1. **input:** $\mathcal{Z} \in \mathbb{R}^{d_u \times D}$, $\mathcal{R} \in \mathbb{R}^{D \times k}$, thresholds λ and δ;
2. $\mathcal{Q} = \texttt{PowerMethod}(\mathcal{Z}, \mathcal{R}, \delta)$; // Approximate the top k_t left singular values
3. $[\mathcal{U}, \Sigma, \mathcal{V}] = \text{SVD}(\mathcal{Q}^\top \mathcal{Z})$;
4. $\mathcal{U} = \{u_i | \sigma_i > \lambda\}$; // Much smaller and less (exact) SVT performed on $\mathcal{Q}^\top \mathcal{Z}$
5. $\mathcal{V} = \{v_i | \sigma_i > \lambda\}$;
6. $\Sigma = \max(\Sigma - \lambda \mathcal{I}, \mathbf{0})$; // ($\mathcal{I}$ denotes the identity matrix)
7. **return** $\mathcal{QU}, \Sigma, \mathcal{V}$.

[Yao and Kwok ('15)]

Algorithm 4: PLAIS-Impute for Collective Matrix Completion $\mathcal{O}(1/T^2)$

1. **input:** observed collective matrix $\boldsymbol{\mathcal{Y}}$, parameter λ, decay parameter $\nu \in (0,1)$, tolerance ε;
2. $[\boldsymbol{\mathcal{U}}_0, \lambda_0, \boldsymbol{\mathcal{V}}_0] = \text{rank-1 SVD}(\boldsymbol{\mathcal{Y}})$;
3. **initialize** $c = 1$, $\delta_0 = \|\boldsymbol{\mathcal{Y}}\|_F$, $\boldsymbol{\mathcal{W}}_0 = \boldsymbol{\mathcal{W}}_1 = \lambda_0 \boldsymbol{\mathcal{U}}_0 \boldsymbol{\mathcal{V}}_0^\top$;
4. **for** $t = 1, \ldots, T$ **do**
5. $\quad \delta_t = \nu^t \delta_0$; // Regularization is dynamically reduced by continuation strategy
6. $\quad \lambda_t = \nu^t(\lambda_0 - \lambda) + \lambda$;
7. $\quad \theta_t = (c-1)/(c+2)$; //Acceleration (FISTA)
8. $\quad \boldsymbol{\mathcal{Q}}_t = (1+\theta_t)\boldsymbol{\mathcal{W}}_t - \theta_t \boldsymbol{\mathcal{W}}_{t-1}$;
9. $\quad \boldsymbol{\mathcal{Z}}_t = \nabla \mathscr{L}_{\boldsymbol{\mathcal{Y}}}(\boldsymbol{\mathcal{Q}}_t))$;
10. $\quad \boldsymbol{\mathcal{V}}_{t-1} = \boldsymbol{\mathcal{V}}_{t-1} - \boldsymbol{\mathcal{V}}_t(\boldsymbol{\mathcal{V}}_t^\top \boldsymbol{\mathcal{V}}_{t-1})$ //Warm-start
11. $\quad \boldsymbol{\mathcal{R}}_t = \text{QR}([\boldsymbol{\mathcal{V}}_t, \boldsymbol{\mathcal{V}}_{t-1}])$;
12. $\quad [\boldsymbol{\mathcal{U}}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \boldsymbol{\mathcal{V}}_{t+1}] = \texttt{Approx-SVT}(\boldsymbol{\mathcal{Z}}_t, \boldsymbol{\mathcal{R}}_t, \lambda_t, \delta_t)$; //Approximate SVT
13. $\quad$ **if** $\mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\mathcal{V}}_{t+1}^\top) > \mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_t\boldsymbol{\Sigma}_t\boldsymbol{\mathcal{V}}_t^\top)$ **then** //Restart the algorithm if the objective function increases
    $\quad\quad c = 1$;
14. $\quad$ **else**
    $\quad\quad c = c + 1$;
15. $\quad$ **if** $|\mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\mathcal{V}}_{t+1}^\top) - \mathscr{F}_\lambda(\boldsymbol{\mathcal{U}}_t\boldsymbol{\Sigma}_t\boldsymbol{\mathcal{V}}_t^\top)| \le \varepsilon$ **then**
    $\quad\quad$ break;
16. **return** $\boldsymbol{\mathcal{W}}_{T+1}$.

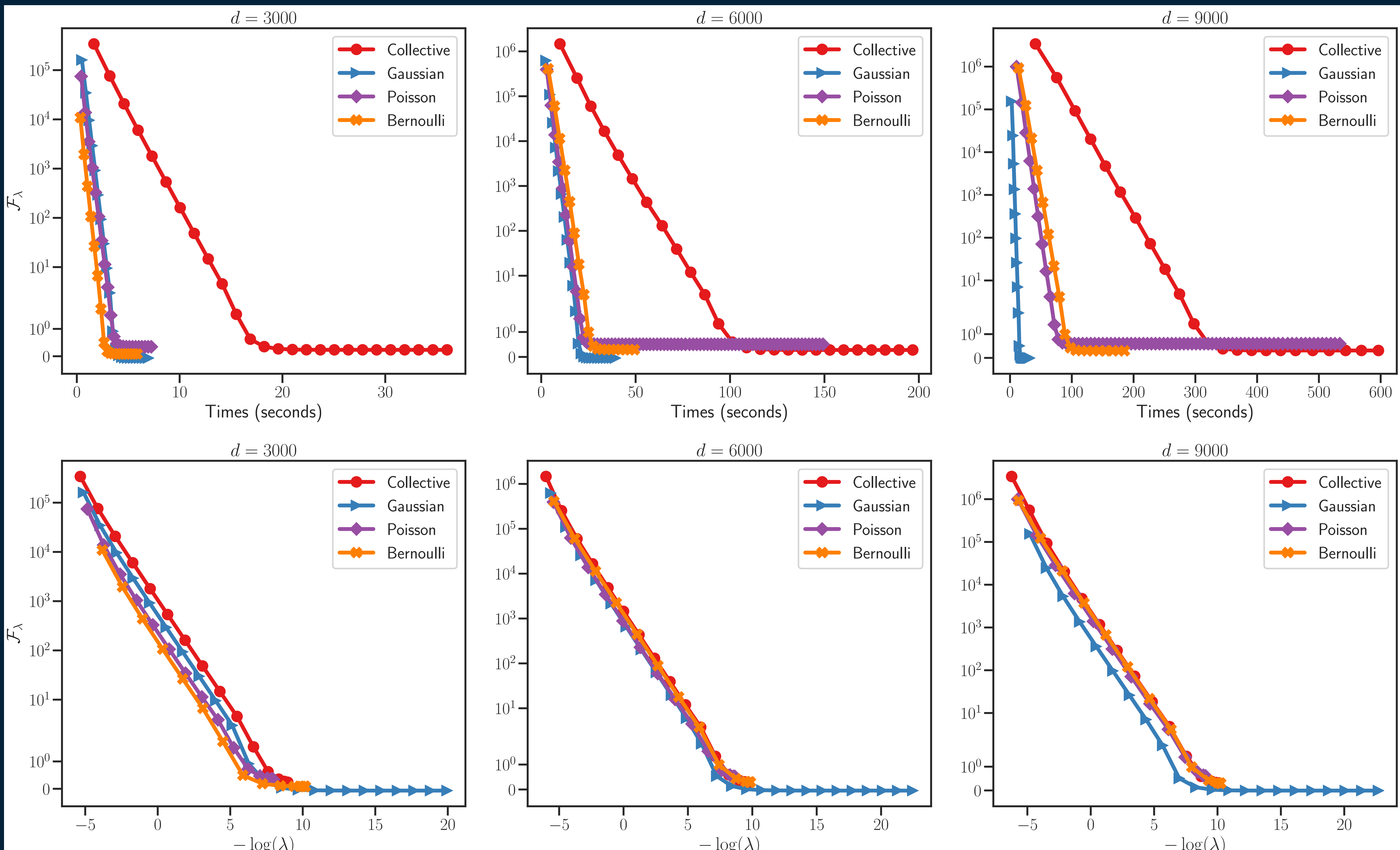
Experimental results on synthetic data

Each source matrix $\boldsymbol{M}^v$ is constructed as $\boldsymbol{M}^v = \boldsymbol{L}^v \boldsymbol{R}^{v\top}$ where $\boldsymbol{L}^v \in \mathbb{R}^{d \times r_v}$ and $\boldsymbol{R}^v \in \mathbb{R}^{d_v \times r_v}$

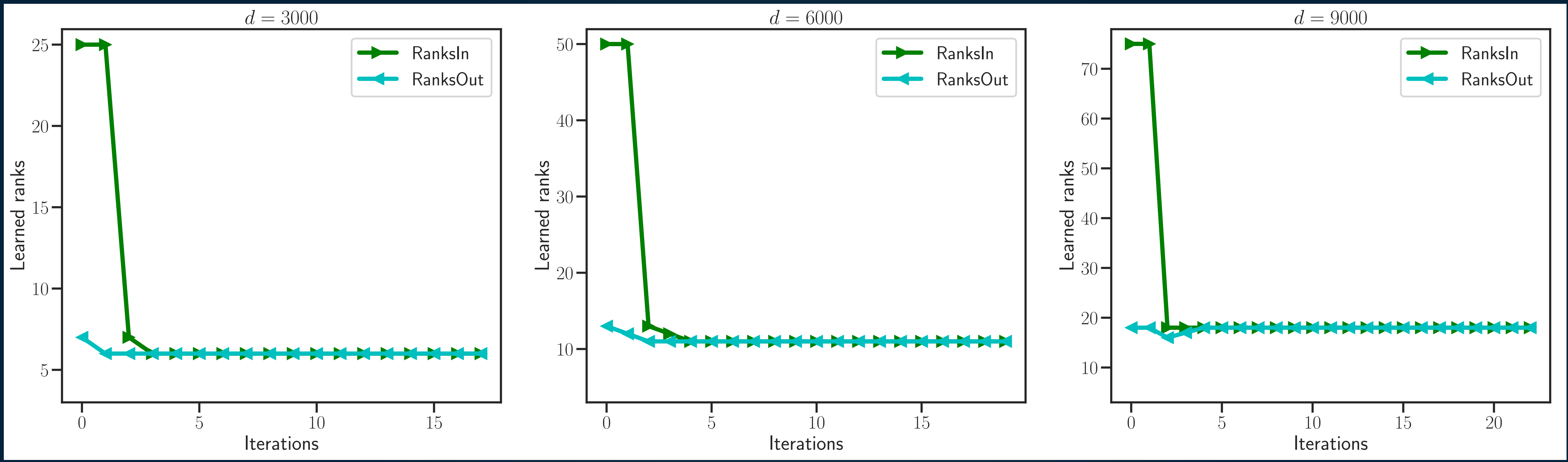
A fraction of the entries of $\boldsymbol{M}^v$ is removed uniformly at random with probability $p \in [0, 1]$.

		$i.i.d.\,\mathcal{N}(0.5, 1)$	$i.i.d.\,\mathcal{P}(0.5)$	$i.i.d.\,\mathcal{B}(0.5)$	
		$\boldsymbol{M}^1$ (*Gaussian*)	$\boldsymbol{M}^2$ (*Poisson*)	$\boldsymbol{M}^3$ (*Bernoulli*)	$\boldsymbol{\mathcal{M}}$ (*Collective*)
`exp.1`	*dimension*	3000×1000	3000×1000	3000×1000	3000×3000
	rank	5	5	5	*unknown*
`exp.2`	*dimension*	6000×2000	6000×2000	6000×2000	6000×6000
	rank	10	10	10	*unknown*
`exp.3`	*dimension*	9000×3000	9000×3000	9000×3000	9000×9000
	rank	15	15	15	*unknown*

Experimental results on synthetic data: convergence of the objective functions



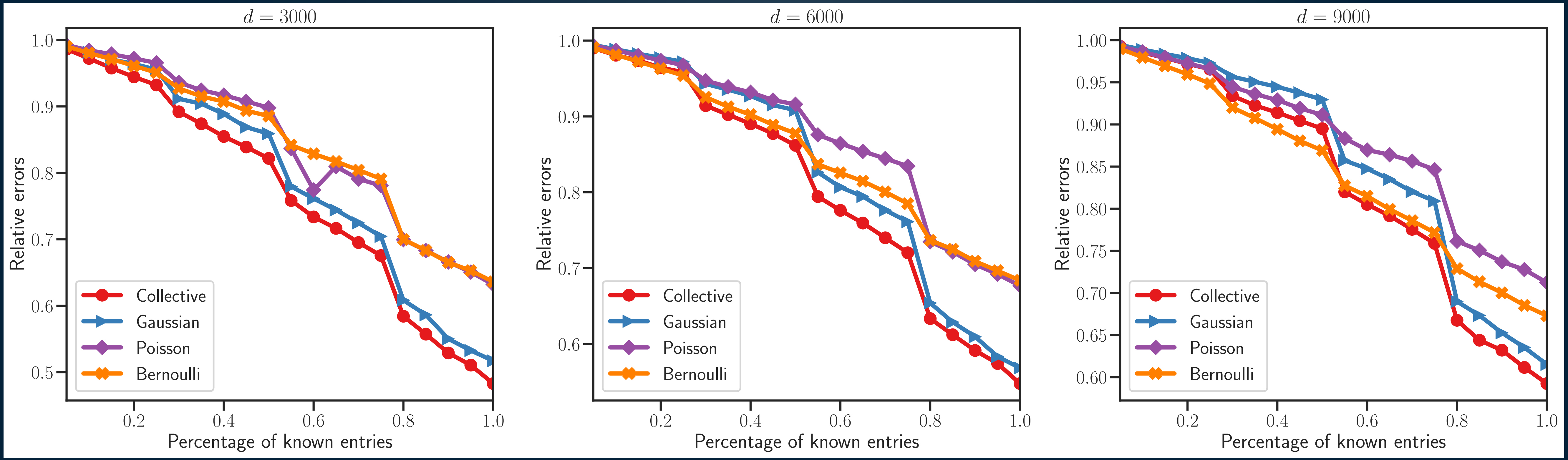
Experimental results on synthetic data: Learning ranks curve



Experimental results on synthetic data: evaluation of the estimator

- Our metric matrix completion is defined by the relative error, [Cai. et al. ('10); Davenport et al. ('14); Cai and Zhou ('13)],

$$\mathrm{RE}(\widehat{\boldsymbol{W}}, \boldsymbol{W}) = \frac{\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^o\|_F}{\|\boldsymbol{W}^o\|_F}$$

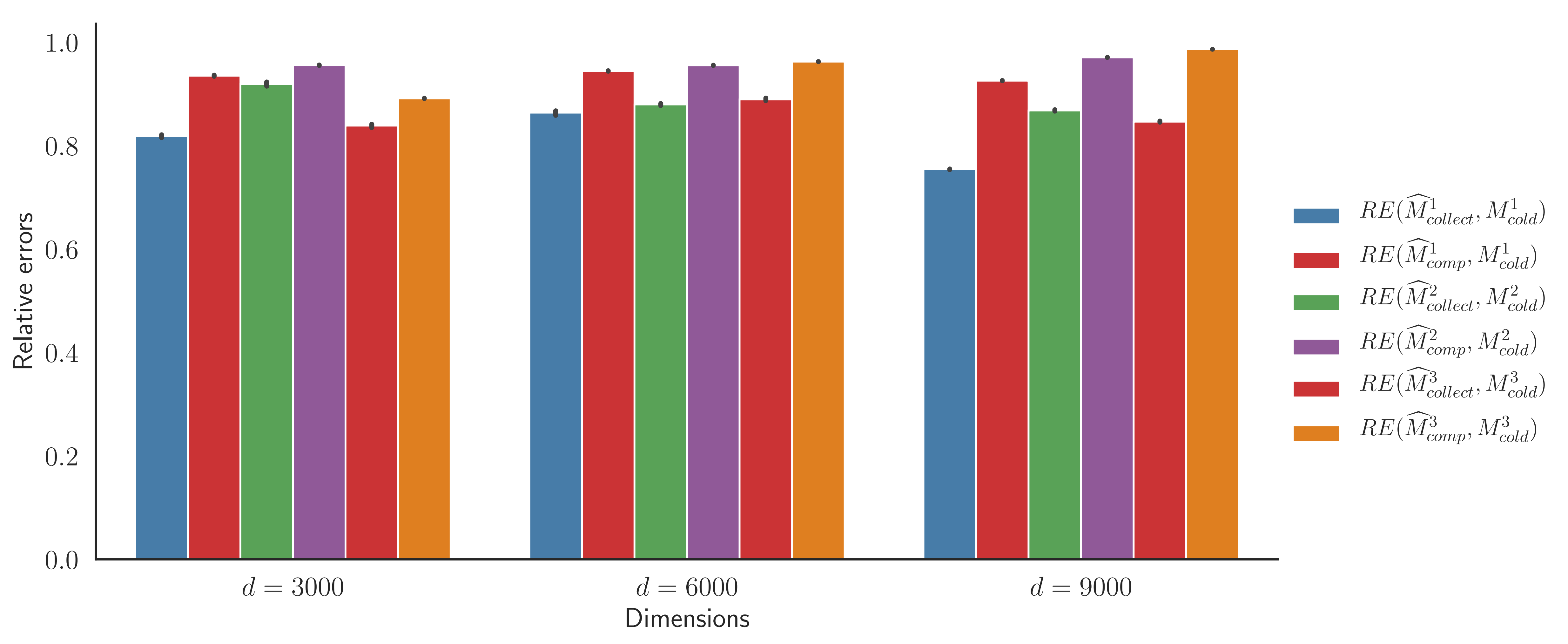


Performance on the synthetic data in terms of relative errors between the target and the estimator matrices

Experimental results on synthetic data: Cold-Start problem

- We construct the ``cold'' collective matrices: we extract vector of known entries of the chosen matrix and we set the first 1/5 fraction of its entries to be equal to zero.

$\mathcal{M}^1_{\text{cold}} = (M^1_{\text{cold}}, M^2, M^3)$, $\mathcal{M}^2_{\text{cold}} = (M^1, M^2_{\text{cold}}, M^3)$, and $\mathcal{M}^3_{\text{cold}} = (M^1, M^2, M^3_{\text{cold}})$.



Take home message

- Recovering a low-rank matrix when the data are collected from multiple and heterogeneous source matrices.

- Estimators are based on minimizing the sum of a goodness-of-fit term and the nuclear norm penalization of the whole collective matrix.

- Upper bounds on the prediction risk of the estimators.

- Empirical evidence of the efficiency of the collective matrix completion approach in the case of joint low-rank structure compared to estimate each source matrices separately.

References

Alaya, Mokhtar Z., and Olga Klopp. 2019. "Collective Matrix Completion." *Journal of Machine Learning Research* 20:1–43.

T. Cai and W. X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14(1):3619–3647, 2013.

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

M. Fazel. *Matrix Rank Minimization with Applications.* PhD thesis, Stanford University, 2002.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

O. Klopp. Matrix completion by singular value thresholding: Sharp bounds. *Electron. J. Statist.*, 9(2):2348–2369, 2015.

Q. Yao and J. T. Kwok. Accelerated inexact soft-impute for fast large-scale matrix completion. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 4002–4008. AAAI Press, 2015.

Thank you!

Distribution-free-setting: remarks

- In 1-bit matrix completion with logistic (resp. hinge) loss, the Bernstein assumption is satisfied with $\varsigma = 1/(4e^{2\gamma})$ (resp. $\varsigma = 2\tau$, such that $|\overset{\star}{M}{}^v_{ij} - 1/2| \geq \tau, \forall v \in [V], (i,j) \in [d_u] \times [d_v]$) [Alquier et al. (2017)].

- The excess risk with respect to these two losses under the uniform sampling is obtained without a logarithmic factor [Alquier et al. (2017)],

$$R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}}) \lesssim \frac{\mathrm{rank}(\overset{\star}{\mathcal{M}})}{p(d_u \wedge D)}.$$