Collective Matrix Completion

Mokhtar Z. Alaya
Modal'X, Paris Nanterre University
joint work with Olga Klopp (ESSEC Business School)

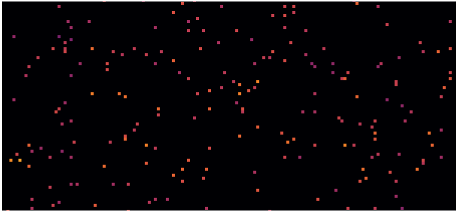
Université Paris Nanterre

MATH INNOV

îledeFrance

4th International Society for NonParametric Statistics
Salerno, Italy, June 13th 2018

Matrix completion is ...



- **Task:** given a partially observed data matrix $\boldsymbol{X}$, predict the unobserved entries
- Large matrices: $\#$ rows, $\#$ columns $\approx 10^5, 10^6$.
- Very under-determined (often only 1-2% observed)
- Application to recommender systems, system identification, image processing, microarray data, etc.

Motivations: recommendation systems, Netflix prize

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

- 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.
- **Low-rank assumption**: fill matrix such that rank is minimum. $\rightarrow$ A few factors explain most of the data.

Completion via rank minimization

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

- Non-convex problem and combinatorially NP-hard!!

Convex formulation of the rank minimization problem

$$\operatorname{rank}(\boldsymbol{X}) = \sum_{i=1}^{\min \dim(\boldsymbol{X})} \mathbb{1}_{(\sigma_i(\boldsymbol{X})>0)} = \|\sigma(\boldsymbol{X})\|_0.$$

Replace ℓ_0 by ℓ_1 [Fazel (2002), Srebro et al. (2005); Candes and Tao (2010); Recht et al. (2010); Negahban and Wainwright (2011); Klopp (2014)]:

$$\text{Nuclear norm: } \|\boldsymbol{X}\|_* = \sum_{i=1}^{\min \dim(\boldsymbol{X})} (\sigma_i(\boldsymbol{X})).$$

Hence temping to consider

Nuclear norm minimization:

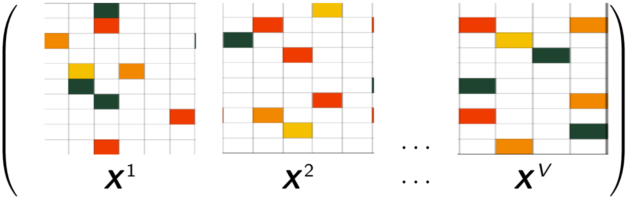
$$\text{minimize}_{\boldsymbol{W}} \|\boldsymbol{W}\|_* \text{ s. t. } W_{ij} = \underbrace{X_{ij}}_{\text{observed entries}}, \ (i,j) \in \underbrace{\Omega}_{\text{sampling set}}.$$

This is a convex problem !

Collective matrix completion: motivations

- Data is often obtained from a collection of matrices $\mathcal{X} = (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^V)$.

$\mathcal{X} =$



- It may be beneficial to leverage all the available user data by various sources.
- **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 matrix completion: model setup

- Each source view $\boldsymbol{X}^v \in \mathbb{R}^{d_u \times d_v}$ and $D = \sum_{v=1}^{V} d_v$.
- We assume that the distribution of for each source $\boldsymbol{X}^v$ depends on the matrix of parameters $\boldsymbol{M}^v$.
- **Model:** let B_{ij}^v be independent Bernoulli random variables and independent from X_{ij}^v, with parameter π_{ij}^v.

$$Y_{ij}^v = B_{ij}^v X_{ij}^v.$$

- We can think of the B_{ij}^v as masked variables.
- $\pi_{ij}^v =$ probability to observe the (i, j)-th entry of the v-th source.

- We consider **general sampling model** where we only assume:

Assumption 1: There exists a positive constant $0 < p < 1$ s.t. $\min_{v \in [V]} \min_{(i,j) \in [d_u] \times [d_v]} \pi_{ij}^v \geq p$.

[Klopp (2015); Klopp et al. (2015)]

- $\pi_{i\bullet}^v = \sum_{j=1}^{d_v} \pi_{ij}^v$ the probability to observe an element from the i-th row of $\boldsymbol{X}^v$.
- $\pi_{\bullet j}^v = \sum_{i=1}^{d_u} \pi_{ij}^v$ the probability to observe an element from the j-th column of $\boldsymbol{X}^v$.

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

Exponential family noise

- Heterogeneous sources: (ratings), (counting: number of clicks) (binomial: like/dislike)
- General framework: natural exponential family:

$$X_{ij}^v | M_{ij}^v \sim h^v(X_{ij}^v) \exp\big(X_{ij}^v M_{ij}^v - G^v(M_{ij}^v)\big).$$

[Gunasekar et al. (2014); Cao and Xie (2016); Lafond (2015)]

- Many distributions belong to the exponential family: Gaussian, binomial, Poisson, exponential, etc.

Assumption 2: - The distribution of X_{ij}^v has sub-exponential tail.
- Strong convexity of the log-partition function G^v.

- Given observations $\mathcal{Y} = (\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^V)$, we write the negative log-likelihood as

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

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

$$\widehat{\mathcal{M}} = (\widehat{\boldsymbol{M}}^1, \dots, \widehat{\boldsymbol{M}}^V) = \operatorname*{argmin}_{\|\mathcal{W}\|_\infty \le \gamma} \mathscr{L}_{\mathcal{Y}}(\mathcal{W}) + \lambda \|\mathcal{W}\|_*,$$

- $\lambda > 0$ is a positive regularization parameter that balances the trade-off between model fit and privileging a low-rank solution.

Exponential family noise: theoretical guarantee

Upper bound of Frobenius estimation risk norm: the rate of convergence has the following dominant term:

Theorem [A., Klopp 2018]

Assume that Assumptions 1 and 2 hold and

$$\lambda \approx \frac{\sqrt{\mu} + (\log(d_u \vee D))^{3/2}}{d_u D}.$$

Then, with high probability, one has

$$\frac{1}{d_u D}\|\widehat{\mathcal{M}} - \mathcal{M}\|_F^2 \lesssim \frac{\mathrm{rank}(\mathcal{M})(\mu + (\log(d_u \vee D))^{3/2})}{p^2 d_u D}.$$

Exponential family noise: theoretical guarantee

- **Uniform sampling:** If $c_1/(d_u dv) \leq \pi_{ij}^v \leq c_2/(d_u d_v)$, then

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

- We denote $n = \sum_{v\in[V]} \sum_{(i,j)\in[d_u]\times[d_v]} \pi_{ij}^v$, the expected number of observations.
- **Sample complexity:**

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

Example: 1-bit matrix completion

- **1-bit matrix completion:** $\mathcal{Y} \in \{+1, -1\}$ with probability $f(\mathcal{M})$ for some link-function f [Davenport et al. (2014); Klopp et al. (2015); Alquier et al. (2017)]
- Klopp et al. (2015) obtained the rate $\operatorname{rank}(\mathcal{M})(d_u \vee D)\log(d_u \vee D)/n$ as the upper bound and $\operatorname{rank}(\mathcal{M})(d_u \vee D)/n$ as the lower bound for 1-bit matrix completion.

Corollary[A., Klopp 2018]

$$\frac{1}{d_u D}\|\widehat{\mathcal{M}} - \mathcal{M}\|_F^2 \lesssim \frac{\operatorname{rank}(\mathcal{M})(d_u \vee D)}{n},$$

- **Answer** the important theoretical question: what is the exact minimax rate of convergence for 1-bit matrix completion which was previously known up to a logarithmic factor.

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 ℓ^v,
- We focus on non-negative loss functions that are Lipschitz:

Assumption 3: We assume that the loss function $\ell^v(y, \cdot)$ is ρ_v-Lipschitz in its second argument: $\ell^v(y, x) - \ell^v(y, x')| \leq \rho_v |x - x'|$.

- Examples: hinge loss with $\ell^v(y, y') = \max(0, 1 - yy')$, logistic loss with $\ell^v(y, y') = \log(1 + \exp(-yy'))$, etc.

Distribution-free setting: estimation procedure

- Goodness-of-fit term:

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

- We define the oracle as:

$$\overset{\star}{\mathcal{M}} = (\overset{\star}{M}{}^1, \ldots, \overset{\star}{M}{}^V) = \underset{\|\mathcal{W}\|_\infty \leq \gamma}{\operatorname{argmin}} \; R(\mathcal{W}),$$

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

- For a tuning parameter $\Lambda > 0$, the nuclear norm penalized estimator $\widehat{\mathcal{M}}$ is defined as

$$\widehat{\mathcal{M}} \in \underset{\|\mathcal{W}\|_\infty \leq \gamma}{\operatorname{argmin}} \left\{ R_{\mathcal{Y}}(\mathcal{W}) + \Lambda \|\mathcal{W}\|_* \right\}.$$

- We denote by $\|\mathcal{W}\|_{\Pi,F}^2 = \sum_v \sum_{(i,j)} \pi_{ij}^v (W_{ij}^v)^2$.

Assumption 4: Assume that for every $\mathcal{W}$ with $\|\mathcal{W}\|_\infty \leq \gamma$, one has $R(\mathcal{W}) - R(\overset{\star}{\mathcal{M}}) \gtrsim \frac{1}{d_u D}\|\mathcal{W} - \overset{\star}{\mathcal{M}}\|_{\Pi,F}^2$.

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

Theorem [A. Klopp 2018]

Let Assumptions 1, 3, and 4 hold and $\Lambda \approx (\sqrt{\mu} + \sqrt{\log(d_u \vee D)})/(d_u D)$. Then, with probability , one has

$$R(\widehat{\mathcal{M}}) - R(\overset{\star}{\mathcal{M}}) \lesssim \frac{\mu + \log(d_u \vee D)}{p d_u D}$$

Take Home Message

- First theoretical guarantees on the case of noisy collective MC.
- Collective approach provides faster rate of convergences in the case of joint low-rank structure.
- Exact minimax optimal rate of convergence for 1-bit matrix completion which was known upto a logarithmic factor.
- On going work: algorithmic study with numerical experiments.

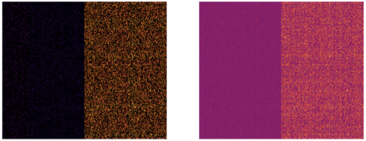
Thank You.

Toy Illustration using cvxpy package

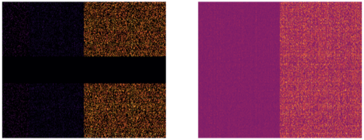
V	d_u	d_1	d_2	d_3	$M^1(\text{rank} = 5)$	$M^2(\text{rank} = 10)$	$M^3(\text{rank} = 15)$
3	500	100	200	300	$\mathcal{N}(-2, 0.5)$	$\mathcal{N}(1, 0.5)$	$\mathcal{N}(2, 0.5)$

	M^1	M^2	M^3	$\boldsymbol{\mathcal{M}}$	$\boldsymbol{\mathcal{M}}_{\text{cold}}$
% observations	10%	20%	30%	23.29%	18.69%

	CMC	SNN	Cold-Start	$\widehat{M^1}$	$\widehat{M^2}$	$\widehat{M^3}$
RMSE	0.223	0.224	0.220	0.198	0.194	0.311



observed + fitted collective matrix.



observed + fitted cold collective matrix.

cvxpy [Diamond and S. Boyd (2016)]

Alquier, P., V. Cottet, and G. Lecué (2017). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *arXiv:1702.01402*.

Bartlett, P. L., M. I. Jordan, and J. D. Mcauliffe (2004). Large margin classifiers: Convex loss, low noise, and convergence rates. In S. Thrun, L. K. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16*, pp. 1173–1180. MIT Press.

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

Cao, Y. and Y. Xie (2016, March). Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing 64*(6), 1609–1620.

Davenport, M. A., Y. Plan, E. van den Berg, and M. Wootters (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA 3*(3), 189.

Elsener, A. and S. van de Geer (2018). Robust low-rank matrix estimation. *To appear in The Annals of Statistics, arXiv preprint arXiv:1603.09071*.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. Ph. D. thesis, Stanford University.

Gunasekar, S., P. Ravikumar, and J. Ghosh (2014). Exponential family matrix completion under structural constraints. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1917–II–1925. JMLR.org.

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

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

Klopp, O., J. Lafond, E. Moulines, and J. Salmon (2015). Adaptive multinomial matrix completion. *Electron. J. Statist. 9*(2), 2950–2975.

Lafond, J. (2015, 03–06 Jul). Low rank matrix completion with exponential family noise. In P. Grünwald, E. Hazan, and S. Kale

(Eds.), *Proceedings of The 28th Conference on Learning Theory*, Volume 40 of *Proceedings of Machine Learning Research*, Paris, France, pp. 1224–1243. PMLR.

Mendelson, S. (2008). Obtaining fast error rates in nonconvex situations. *Journal of Complexity 24*(3), 380 – 397.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist. 39*(2), 1069–1097.

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev. 52*(3), 471–501.

Srebro, N., J. Rennie, and T. S. Jaakkola (2005). Maximum-margin matrix factorization.