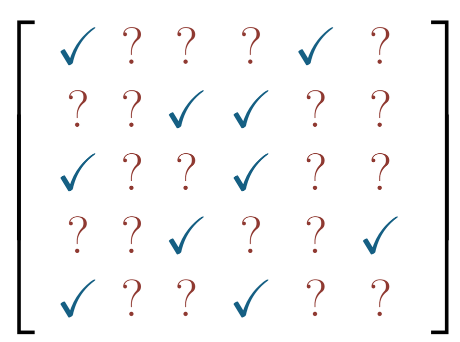
THE UNIVERSITY OF CHICAGO

# Conformalized Matrix Completion

Yu Gui Rina Foygel Barber Cong Ma

Department of Statistics, University of Chicago

## Matrix completion



- Econometrics: panel data prediction and inference
- Recommender system: collaborative filtering
- Bioinformatics: gene-disease association

## Point completion

- Noisy and incomplete observation

$$M_{ij} = M^*_{ij} + \text{noise}, \qquad (i,j) \in \mathcal{S} \subseteq [d_1] \times [d_2]$$

- Point estimation: estimate $\mathbf{M}^*$ (low-rank)
- Point prediction: predict stochastic entries $M_{ij}$ for $(i,j) \in \mathcal{S}^c$

## Success of model-based matrix completion

- Model assumptions: (1) low-rank matrix: $\mathrm{rank}(\mathbf{M}^*) = r = O(1)$, (2) random sampling: $\mathbb{P}((i,j) \in \mathcal{S}) = p$ independently, (3) random i.i.d. sub-Gaussian noise, (4) incoherent and well-conditioned
- Minimax rate $\|\widehat{\mathbf{M}} - \mathbf{M}^*\|_{\mathrm{F}} \asymp \sigma\sqrt{n/p}$ matches computational limit [1]
- Question: How can we quantify the uncertainty in completed entries?

## Model-based inference is feasible

- Asymptotically valid $(1-\alpha)$-confidence interval for $M_{ij}$ based on the asymptotic distribution $\widehat{M}_{ij} - M^*_{ij} \approx \mathcal{N}(0, \theta^2_{ij})$:

$$C(i,j) = \widehat{M}_{ij} \pm q_{1-\alpha/2}\sqrt{\widehat{\theta}^2_{ij} + \widehat{\sigma}^2}$$

- Validity is not guaranteed when the model is misspecified

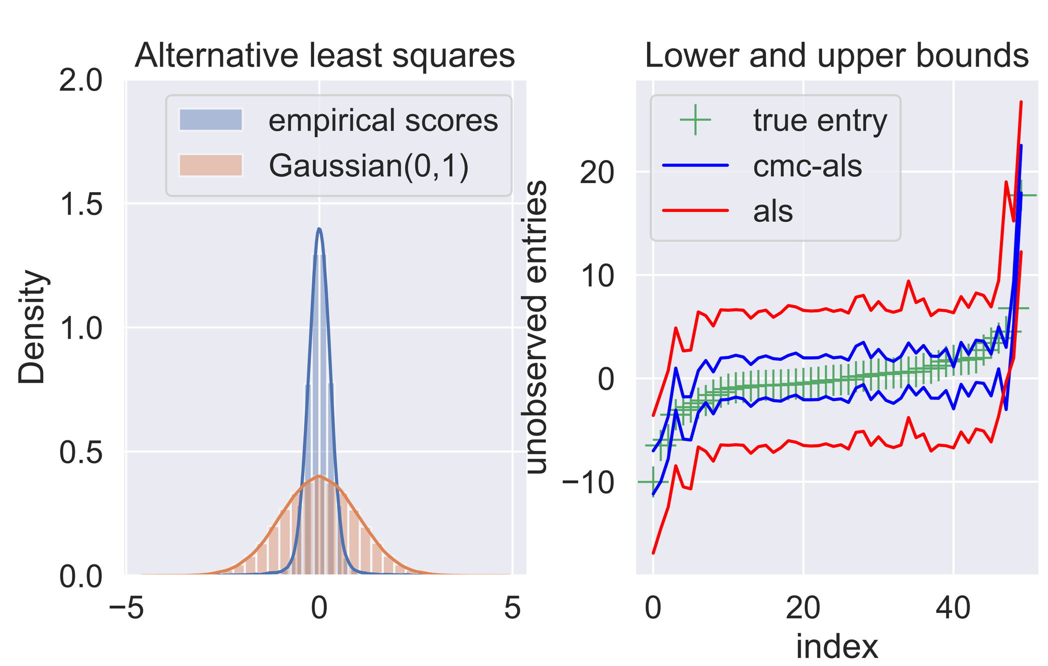



Figure 1. Violation of incoherence.

Question: Is distribution-free inference possible for matrix completion?

- Free of model assumptions on the underlying matrix $\mathbf{M}$
- Free of the choice of estimation algorithms

## Distribution-free uncertainty quantification via split conformal prediction

Heterogeneous sampling: each entry $(i,j)$ is observed with probability $p_{ij} > 0$ independently

Question: Why is matrix completion different from regression problem?

1. How to address the dependence between $\mathcal{S}_{\text{tr}}$ and $\mathcal{S}_{\text{cal}}$?
2. Since the "covariates" $(i,j)$ are sampled without replacement, can we still have a tractable form of weights?

## Conformalized matrix completion (cmc)

- **Input:** $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, $\mathcal{S} = \{(i,j) \in [d_1] \times [d_2] \mid M_{ij} \text{ is observed}\}$
- **Step 1:** For any pre-specified $\eta \in (0,1)$, draw $W_{ij} \sim \mathrm{Bern}(\eta)$
- **Step 2:** Data splitting $\mathcal{S} = \mathcal{S}_{\text{tr}} \cup \mathcal{S}_{\text{cal}}$:

$$\mathcal{S}_{\text{tr}} = \{(i,j) \in \mathcal{S} : W_{ij} = 1\}, \qquad \mathcal{S}_{\text{cal}} = \{(i,j) \in \mathcal{S} : W_{ij} = 0\}.$$

- **Step 3:** With the training set $\mathbf{M}_{\mathcal{S}_{\text{tr}}}$,
  1. Estimate $\widehat{\mathbf{M}} = (\widehat{M}_{ij})$ and $\widehat{\mathbf{P}} = (\widehat{p}_{ij})$
  2. Calculate a local uncertainty estimate $\widehat{\mathbf{s}} = (\widehat{s}_{ij})$
- **Step 4:** With the calibration set,
  1. Calculate the nonconformity scores $R_{ij} = \frac{|M_{ij} - \widehat{M}_{ij}|}{\widehat{s}_{ij}}$, $(i,j) \in \mathcal{S}_{\text{cal}}$
  2. Calculate the weights for each $(i,j) \in \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}$

$$\widehat{w}_{ij} = \frac{\widehat{h}_{ij}}{\sum_{(i',j') \in \mathcal{S}_{\text{cal}}} \widehat{h}_{i'j'} + \widehat{h}_{i_*j_*}}$$

  3. Calculate the quantile for each $(i_*, j_*)$:

$$\widehat{q}_{i_*j_*} = \mathrm{Quantile}_{1-\alpha}\left(\sum_{(i,j) \in \mathcal{S}_{\text{cal}}} \widehat{w}_{ij}\delta_{R_{ij}} + \widehat{w}_{i_*j_*}\delta_{\infty}\right)$$

- **Output:**

$$\widehat{C}(i_*, j_*) = \widehat{M}_{i_*j_*} \pm \widehat{q}_{i_*j_*}\widehat{s}_{i_*j_*}$$

## Weighted exchangeability conditioning on $\mathcal{S}_{\text{tr}}$

Lemma. If $(i_*, j_*) \mid \mathcal{S} \sim \mathrm{Unif}(\mathcal{S}^c)$, it holds that

$$\mathbb{P}\Big\{(i_*, j_*) = (i_k, j_k) \mid \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\} = \mathbb{S}^{(n_{\text{cal}}+1)}, \mathcal{S}_{\text{tr}}\Big\} = w_{i_kj_k}$$

where $\mathbb{S}^{(n_{\text{cal}}+1)} = \{(i_1, j_1), \ldots, (i_{n_{\text{cal}}+1}, j_{n_{\text{cal}}+1})\}$ is the unordered set and we define the weights

$$w_{i_kj_k} = \frac{h_{i_kj_k}}{\sum_{k'=1}^{n_{\text{cal}}+1} h_{i_{k'}j_{k'}}} \qquad \text{with odds ratio} \qquad h_{ij} = \frac{1-p_{ij}}{p_{ij}}.$$

## Theoretical guarantee

Define the average coverage rate over the unsampled set

$$\mathtt{AvgCov}(\widehat{C}; \mathbf{M}, \mathcal{S}) = \frac{1}{|\mathcal{S}^c|}\sum_{(i,j) \in \mathcal{S}^c} \mathbf{1}\Big\{M_{ij} \in \widehat{C}(i,j)\Big\}$$

### Theorem

Conformalized matrix completion (cmc) satisfies

$$\mathbb{E}\Big[\mathtt{AvgCov}(\widehat{C}; \mathbf{M}, \mathcal{S})\Big] \geq 1 - \alpha - \mathbb{E}[\Delta],$$

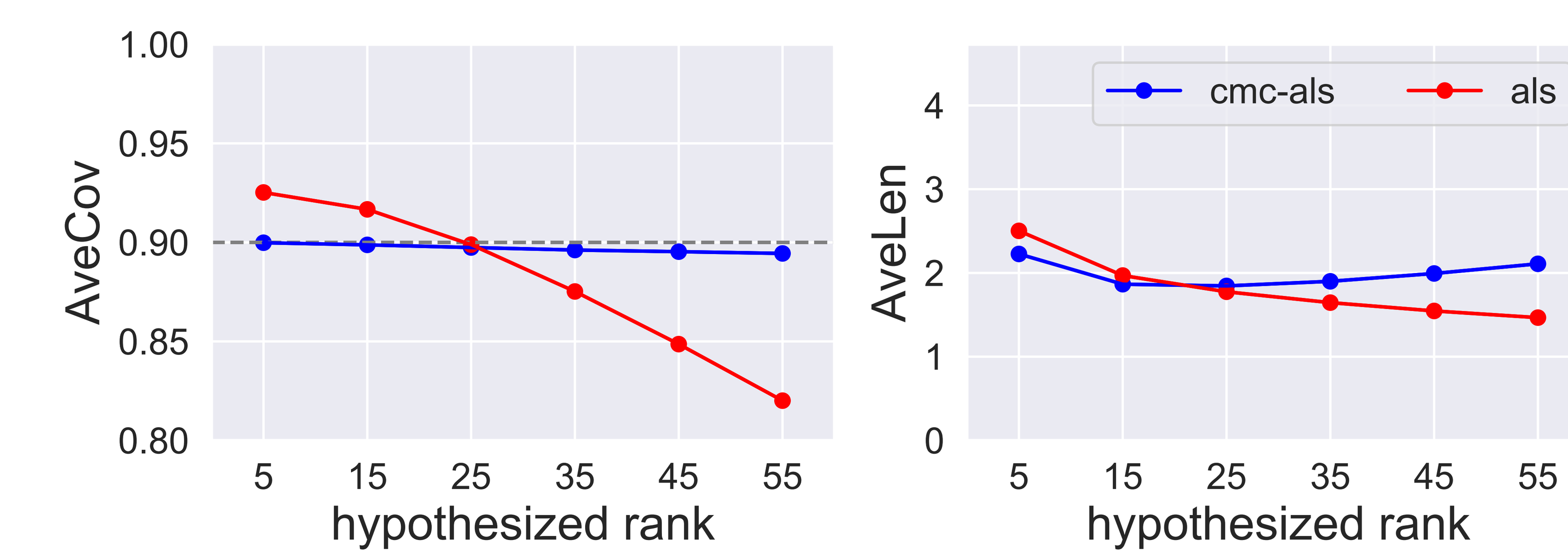
where $\Delta = \frac{1}{2}\sum_{(i,j) \in \mathcal{S}_{\text{cal}} \cup \{(i_*, j_*)\}} \Big|\widehat{w}_{ij} - w_{ij}\Big|$

**Coverage gap with common sampling models.**

- Uniform sampling $p_{ij} = p \implies \Delta = 0$
- Logistic missingness $-\log(h_{ij}) = u_i + v_j$ and $\boldsymbol{u}^\top \mathbf{1} = 0$. Maximum likelihood estimator yields $\mathbb{E}[\Delta] \lesssim \sqrt{\frac{\log(\max\{d_1,d_2\})}{\min\{d_1,d_2\}}}$
- Missingness with a general link function $-\log\left(h_{ij}\right) = \phi(A_{ij})$ and $\mathrm{rank}(\mathbf{A}) = k^*$. MLE yields $\mathbb{E}[\Delta] \lesssim \min\{d_1, d_2\}^{-1/4}$

## Numerical simulations with Rossmann sales dataset

- Heterogeneous missingness: $p_{ij} = 0.8$ for weekdays and $p_{ij} = 0.8/3$ for weekends; $p_{ij} = 0.8/3$ for 200 randomly sampled stores
- Working model: one-bit model with a logistic link function
- More simulation results can be found in Gui et al. [2]




## References

[1] Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

[2] Gui, Y., Barber, R. F., and Ma, C. (2023). Conformalized matrix completion. *arXiv preprint arXiv:2305.10637*.