Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula

Yuxuan Zhao Madeleine Udell
Cornell University

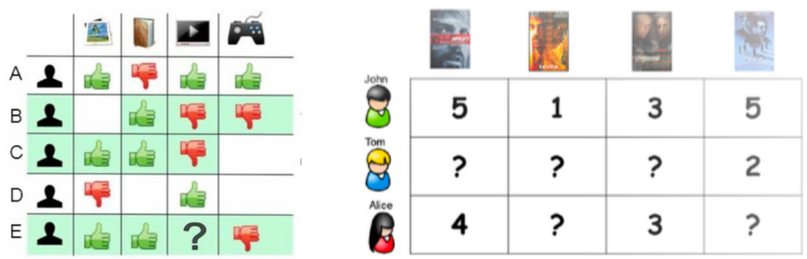
The Paper in 1 Minute

Motivating Questions

- How to impute missing values, unaffected by marginal distributions?
- How to quantify the uncertainty of a single imputation?

Our Contribution

- A new probabilistic method to impute real-valued and ordinal data.
- Confidence intervals for real-valued data.
- Probability lower bound on correct imputation for ordinal data.
- A measure "reliability" for selecting imputed entries with smaller error.



Our Model: PPCA + Gaussian Copula

Given matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, each row $\mathbf{x^i} = \mathbf{g}(\mathbf{z^i}) = \mathbf{g}(\mathbf{W}\mathbf{t}^i + \epsilon^i) \in \mathbb{R}^p$,

- $\mathbf{t}^i \in \mathbb{R}^k$ i.i.d. from $\mathcal{N}(0, \mathrm{I}_k)$ with $k < p$.
- ϵ^i i.i.d. from $\mathcal{N}(0, \sigma^2\mathrm{I}_p)$ and independent from $\mathbf{t}^i$.
- Elementwise monotone $\mathbf{g}(\mathbf{z}^i) := (g_1(z_1^i, \ldots, z_p^i))$ for $\mathbf{z}^i = \mathbf{W}\mathbf{t}^i$.
- $\mathbf{W}\mathbf{W}^\top + \sigma^2\mathrm{I}_p$ has unit diagonals.

Our Imputation: Row-wise Conditional Mean Imputation

For a row $\mathbf{x} \sim \mathrm{LRGC}(\mathbf{W}, \sigma^2, \mathbf{g})$ with observed $\mathbf{x}_\mathcal{O}$ and missing entries $\mathbf{x}_\mathcal{M}$,

$$\begin{aligned}\text{Imputation: } \mathbf{x}_\mathcal{M} &= \mathbf{g}_\mathcal{M}\left(\mathrm{E}[\mathbf{z}_\mathcal{M}|\mathbf{x}_\mathcal{O}]\right)\\ &= \mathbf{g}_\mathcal{M}\left(\mathbf{W}_\mathcal{M}(\sigma^2\mathrm{I}_k + \mathbf{W}_\mathcal{O}^\top\mathbf{W}_\mathcal{O})^{-1}\mathbf{W}_\mathcal{O}^\top\mathrm{E}[\mathbf{z}_\mathcal{O}|\mathbf{x}_\mathcal{O}]\right)\end{aligned}$$

- In practice, replace model parameters with their estimates.
- Estimate $\mathbf{g}$ by matching normal quantiles to observed quantiles in $\mathbf{X}$.
- Estimate $\mathbf{W}, \sigma^2$ using EM algorithm with closed form update.

How Accurate Is Our Imputation?

Real valued Data

If x_j is missing,

$$g_j\left(\mathrm{E}\left[z_j \mid \mathbf{x}_\mathcal{O}\right] - z^\star \mathrm{Var}\left[z_j \mid \mathbf{x}_\mathcal{O}\right]\right) < x_j < g_j\left(\mathrm{E}\left[z_j \mid \mathbf{x}_\mathcal{O}\right] + z^\star \mathrm{Var}\left[z_j \mid \mathbf{x}_\mathcal{O}\right]\right).$$

- $\alpha \in (0,1)$ and $z^\star = \Phi^{-1}(1-\frac{\alpha}{2})$.

Ordinal Data

If x_j is missing, the LRGC imputation $\hat{x}_j$ satisfies:

$$\Pr\left(\hat{x}_j = x_j\right) \geq 1 - \mathrm{Var}\left[z_j \mid \mathbf{x}_\mathcal{O}\right]/d_j^2, \text{ where } d_j = \mathrm{dist}(\mathrm{E}\left[z_j \mid \mathbf{x}_\mathcal{O}\right], \mathbf{S}_j).$$

- $\mathbf{S}_j$ is the set of points that cut normal z_j into ordinal x_j.

Which LRGC Imputed Entries Are Most Reliable?

Real valued Data

$$\text{reliability at missing } (i,j): \frac{||\mathrm{P}_{\Omega^c/(i,j)}\left(D_\alpha\right)||_F}{||\mathrm{P}_{\Omega^c/(i,j)}(\hat{\mathbf{X}})||_F}.$$

- Ω stores observed locations. D_α stores the confidence interval length at missing entries. P_A projects on the set A: it sets entries not in A as 0.
- An imputed entry is more reliable if evaluation removing it is worse.

Ordinal Data

$$\text{reliability at missing } (i,j): 1 - \mathrm{Var}\left[z_j^i \mid \mathbf{x}_{\mathcal{O}_i}^i\right]/d_{ij}^2$$

- An imputed entry is more reliable if it has larger probability to be correct.

Results: Confidence Intervals

Table 1: 95% Confidence intervals on synthetic data: monotonically transform noisy low rank Gaussian matrix.

Identical Transformation	LRGC	PPCA	LRMC	MI-PCA
Empirical coverage rate	0.927(.002)	0.940(.001)	0.878(.006)	0.933(.002)
Interval length	1.273(.004)	1.264(.004)	1.129(.015)	1.267(.004)
Run time in seconds	6.9(1)	3.4(1)	2.7(0)	190(15)
Cubic Transformation	LRGC	PPCA	LRMC	MI-PCA
Empirical coverage rate	0.927(.002)	0.943(.002)	0.925(.004)	0.948(.002)
Interval length	3.614(.068)	9.086(.248)	6.546(.191)	9.307(.249)
Run time in seconds	7.2(1)	0.4(0)	3(1)	220(30)

Takeaway:

- Marginal transformation can distort many imputation methods, but not ours!
- Our reliability predicts imputation accuracy well, while MI sample variance cannot!

Results: Select Reliable Imputed Entries

- Evaluate the imputation error on the subset of $m\%$ entries for which method's associated uncertainty metric indicates highest reliability.
- For multiple imputation (MI), lower sample variance indicates higher reliability.