On Inferences from Completed Data

J. Haddock⋆, D. Molitor⋆, D. Needell⋆, S. Sambandam⋆, J. Song†, S. Sun‡

⋆ University of California, Los Angeles † Tsinghua University ‡ Peking University

UCLA

Objectives

We study the effects of sampling and data completion techniques on simple statistical inferences. We compare results of inferences on complete data and data that has been subsampled and then completed.

Sampling Techniques

Uniform sampling: Sample each entry with equal probability $p \in (0,1)$.

Structured sampling: Sample entries equal to zero with probability p_0, and nonzero entries with probability p_1, where $p_0 < p_1$.

The sampled entries are denoted Ω and the subsampled matrix is denoted M_Ω.

Completion Techniques

Nuclear-norm minimization:

$$\underset{X \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \|X\|_* \quad \text{s.t. } M_{ij} = X_{ij} \text{ for all } (i,j) \in \Omega. \tag{NNM}$$

ℓ_1-Regularized NNM:

$$\underset{X \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \|X\|_* + \alpha \|X_{\Omega^C}\|_1 \quad \text{s.t. } M_{ij} = X_{ij} \text{ for all } (i,j) \in \Omega \tag{ℓ_1-NNM}$$

with regularization parameter $\alpha > 0$.

The ℓ_1-regularization in the objective of ℓ_1-NNM encourages unobserved entries of the recovered matrix to be near 0 [1].

Inference Techniques

Entrywise mean:

$$\bar{\lambda}(A) := \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}$$

Row mean:

$$\mu(A) := \frac{1}{m} \sum_{i=1}^{m} \vec{a}_i$$

In our application to health survey data, these inferences could be an average "wellness" score for a patient group (entrywise mean) or the average responses of a patient group (row mean).

Error Measurements

Norm. matrix recovery error:

$$E(M, \bar{M}) := \|M - \bar{M}\|_F / \|M\|_F$$

Abs. entrywise mean error:

$$E_{\bar{\lambda}}(M, \bar{M}) := |\bar{\lambda}(M) - \bar{\lambda}(\bar{M})|$$

Norm. row mean error:

$$E_\mu(M, \bar{M}) := \frac{\|\mu(M) - \mu(\bar{M})\|_2}{\|\mu(M)\|_2}$$

The matrix recovery error measures the error introduced by sampling and data completion, while the inference errors measure the error introduced into the inference by these processes.

Methodology

(1) Begin with complete matrix M either artificial or extracted from real data. (We take this as the ground truth.)

(2) Use either uniform or structured sampling strategy to obtain an incomplete observed matrix, M_Ω. (The values of p and p_0, p_1 used for sampling are noted in each experiment.)

(3) Recover $\bar{M}$ via either NNM or ℓ_1-NNM.
- ▷ For M_Ω sampled uniformly, we recover $\bar{M}$ via NNM.
- ▷ For M_Ω sampled via structured sampling, we recover $\bar{M}$ via ℓ_1-NNM.
  - ◦ We choose the regularization parameter α optimally from among $\{0.05, 0.1, 0.2, ..., 0.5\}$ to minimize the resulting error $\|M - \bar{M}\|_F$.

(4) Compute either the entrywise mean or the row mean.

(5) Plot recovery and inference errors. (Plotted results are averaged over 10 trials.)

Experimental Results

In the figures below, we plot matrix and inference recovery errors on a 30×30 rank 5 synthetic matrix and a complete 30×16 submatrix of the MyLymeData health survey data; the figures differ by the choice of zero sampling probability p_0 for the structured sampling strategy and the data type. Errors are plotted versus the proportion of observed entries ω. In each of the groups of four plots below, we plot optimal regularization parameter α in the upper left; normalized matrix recovery errors E in upper right; normalized row mean errors E_μ in lower left; absolute entrywise mean errors $E_{\bar{\lambda}}$ in lower right.

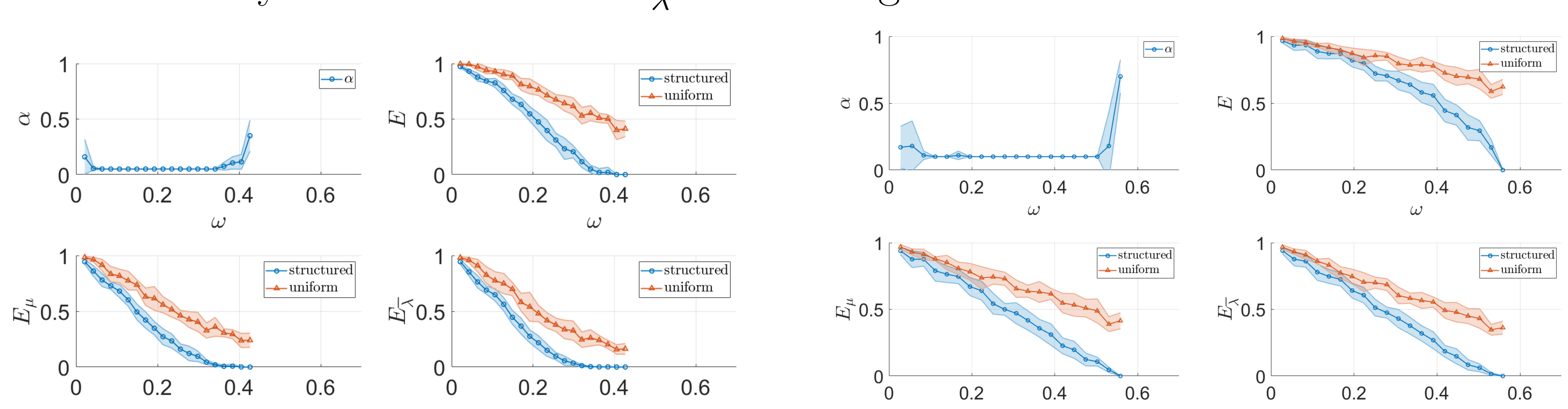

Figure 1: Recovery errors for unif. sampling with NNM and structured sampling with $p_0 = 0$ (no entries equal to zero are sampled) and ℓ_1-NNM on left: synthetic data; right: MyLymeData.

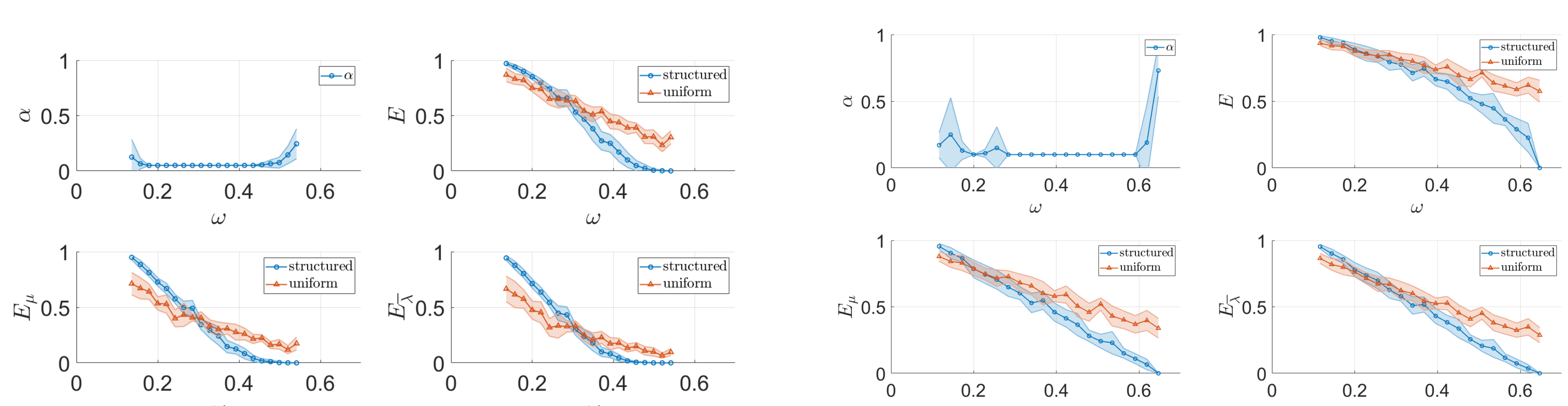

Figure 2: Recovery errors for unif. sampling with NNM and structured sampling with $p_0 = 0.2$ and ℓ_1-NNM on left: synthetic data; right: MyLymeData.

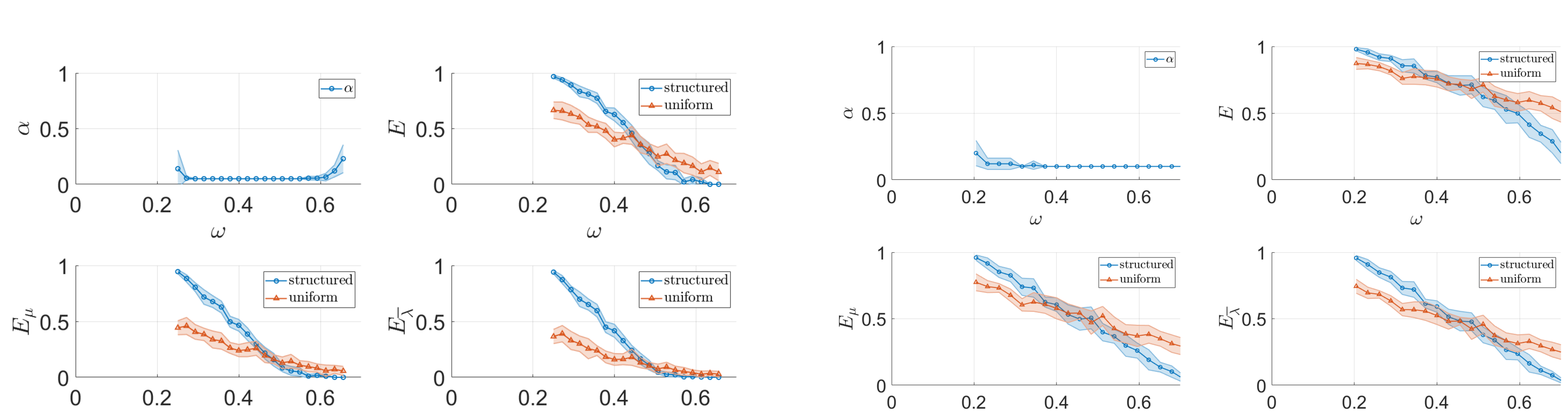

Figure 3: Recovery errors for uniform sampling with NNM and structured sampling with $p_0 = 0.4$ and ℓ_1-NNM on left: synthetic data; right: MyLymeData.

Acknowledgements

DN, JH, and DM are grateful to and were partially supported by NSF CAREER DMS #1348721 and NSF BIGDATA DMS #1740325. This work is based upon work completed at the UCLA CAM REU during Summer 2018 which was funded by NSF DMS #1659676. The authors would like to thank CEO Lorraine Johnson, LymeDisease.org, and the patients who participated in the MyLymeData survey. Additionally, they thank Dr. Anna Ma for her assistance with this data, and Prof. Andrea Bertozzi and the UCLA Applied and Computational Math REU program for their support.

Theoretical Results

Together our two theoretical results offer a bound on the inference recovery errors even if the matrix recovery is not exact.

Theorem 1: Let $\bar{\lambda}$ and μ be the entrywise and row mean operators respectively. Then

$$|\bar{\lambda}(M) - \bar{\lambda}(\bar{M})| \leq (mn)^{-\frac{1}{q}} \|M - \bar{M}\|_q$$

and

$$\|\mu(M) - \mu(\bar{M})\|_q \leq \left(\frac{n^{q-1}}{m}\right)^{\frac{1}{q}} \|M - \bar{M}\|_q$$

for all $M, \bar{M} \in \mathbb{R}^{m\times n}$ and $1 \leq q \leq \infty$.

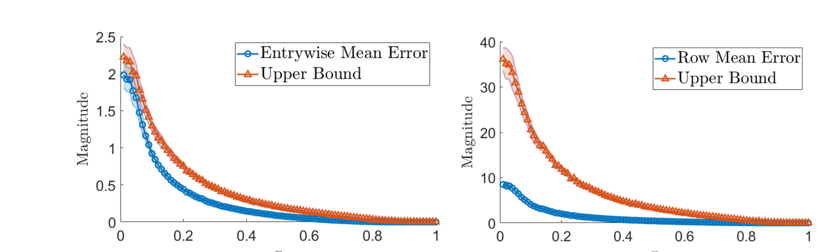

Figure 4: Averages of 400 sampled inference recovery errors and the derived upper bounds for uniform observation sampling probabilities from 0 to 1. Left: entrywise mean error; right: row mean error.

Theorem 2: Let $M \in \mathbb{R}^{m\times n}$, Ω, and $\bar{M}$ be computed via NNM. Let $r = \operatorname{rank}(M)$ denote the rank of M, and denote the singular values of M by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ in decreasing order. Then

$$\|M - \bar{M}\|_F \leq 2\sqrt{r^2\sigma_1^2 - \|M_\Omega\|_F^2}.$$

Conclusion

Our numerical experiments demonstrate that simple inferences such as the entrywise mean or the row mean can be recovered accurately even when errors are introduced by the matrix recovery. We prove bounds on the inference recovery error in terms of the matrix recovery error for the entrywise mean and the row mean. Additionally, we prove an analytical bound on the matrix recovery error which applies even when the matrix cannot be recovered exactly.

References

[1] D. Molitor and D. Needell. Matrix completion for structured observations. *arXiv preprint arXiv:1801.09657*, 2018.

[2] LymeDisease.org. Lymedisease.org, 2018. https://www.lymedisease.org, Last accessed on 2018-08-17.

Contact Information

Web: www.math.ucla.edu/~jhaddock/
Email: jhaddock@math.ucla.edu