On Inferences from Completed Data

Jamie Haddock

February 14, 2019

Computational and Applied Mathematics, UCLA

joint with 2019 UCLA REU group
(D. Molitor, D. Needell, S. Sambandam, J. Song, S. Sun)

Motivation

MyLymeData is a large collection of Lyme disease patient survey data collected by LymeDisease.org ($\sim$12,000 patients, 100s of questions)

Motivation

MyLymeData is a large collection of Lyme disease patient survey data collected by LymeDisease.org (∼12,000 patients, 100s of questions)

- data is highly incomplete due to branching structure of surveys and missing responses

Motivation

MyLymeData is a large collection of Lyme disease patient survey data collected by LymeDisease.org ($\sim$12,000 patients, 100s of questions)

- data is highly incomplete due to branching structure of surveys and missing responses
- research questions of interest do not require individual entries

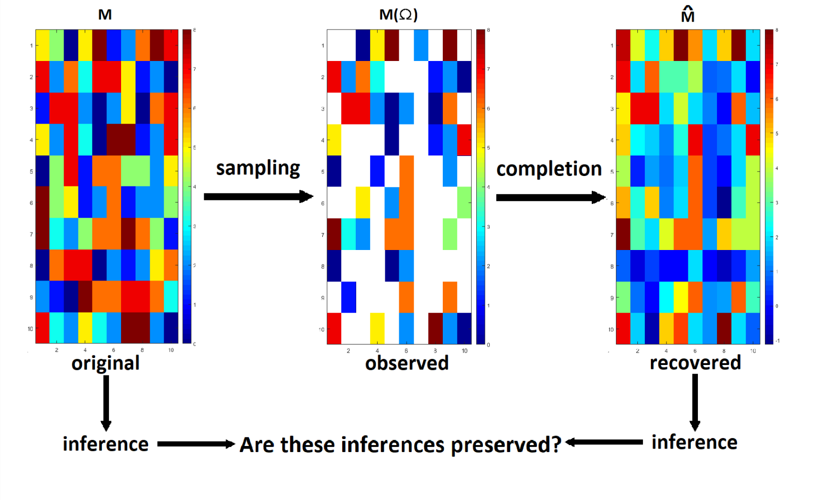
Motivation

MyLymeData is a large collection of Lyme disease patient survey data collected by LymeDisease.org ($\sim$12,000 patients, 100s of questions)

- data is highly incomplete due to branching structure of surveys and missing responses
- research questions of interest do not require individual entries

Question: Can we perform statistical inferences on imputed data?

Main Question



Sampling and Imputation Techniques

Uniform Sampling: Sample each entry with uniform probability p.

Sampling and Imputation Techniques

Uniform Sampling: Sample each entry with uniform probability p.

Structured Sampling: Sample zero and nonzero entries with p_0 and p_1.

Sampling and Imputation Techniques

Uniform Sampling: Sample each entry with uniform probability p.

Structured Sampling: Sample zero and nonzero entries with p_0 and p_1.

Nuclear Norm Minimization (NNM):

$$\begin{aligned} \min \quad & \|\mathbf{X}\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for all } (i,j) \in \Omega \end{aligned}$$

Sampling and Imputation Techniques

Uniform Sampling: Sample each entry with uniform probability p.

Structured Sampling: Sample zero and nonzero entries with p_0 and p_1.

Nuclear Norm Minimization (NNM):

$$\begin{aligned} \min \quad & \|\mathbf{X}\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for all } (i,j) \in \Omega \end{aligned}$$

ℓ_1-Regularized Nuclear Norm Minimization (ℓ_1-NNM):

$$\begin{aligned} \min \quad & \|\mathbf{X}\|_* + \alpha \|\mathbf{X}_{\Omega^c}\|_1 \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for all } (i,j) \in \Omega \end{aligned}$$

Simple Inferences

Entrywise Mean

$\overline{\lambda}(M)$: mean of the entries of M

- Entrywise mean error:

$$E_{\overline{\lambda}} = |\overline{\lambda}(\hat{\mathbf{M}}) - \overline{\lambda}(\mathbf{M})|.$$

Row Mean

$\mu(M)$: average row of M

- Normalized row mean error:

$$E_{\mu} = \frac{\|\mu(\hat{\mathbf{M}}) - \mu(\mathbf{M})\|_2}{\|\mu(\mathbf{M})\|_2}.$$

▷ original matrix, $\mathbf{M}$

▷ recovered matrix, $\hat{\mathbf{M}}$

Experimental Design - Synthetic Data

- 30×30 rank 5 matrix generated as product of sparse matrices with nonzero entries sampled uniformly from $[0, 1]$

Experimental Design - Synthetic Data

- 30×30 rank 5 matrix generated as product of sparse matrices with nonzero entries sampled uniformly from $[0, 1]$
- each trial consists of sampling, completion, and inference on original and completed matrices

Experimental Design - Synthetic Data

- 30×30 rank 5 matrix generated as product of sparse matrices with nonzero entries sampled uniformly from $[0, 1]$
- each trial consists of sampling, completion, and inference on original and completed matrices
  - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively

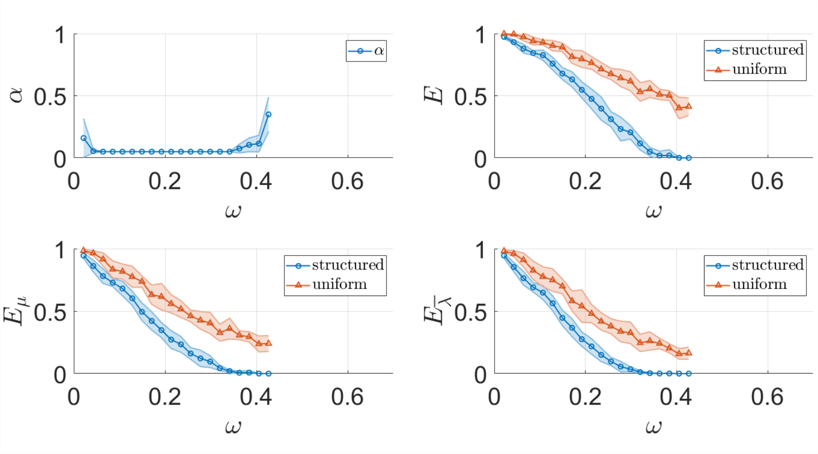
Experimental Design - Synthetic Data

- ▷ 30×30 rank 5 matrix generated as product of sparse matrices with nonzero entries sampled uniformly from $[0, 1]$
- ▷ each trial consists of sampling, completion, and inference on original and completed matrices
  - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively
  - ℓ_1 regularization parameter α is chosen in $\{0.05, 0.1, 0.2, \ldots, 0.5\}$ to minimize matrix recovery error

Experimental Design - Synthetic Data

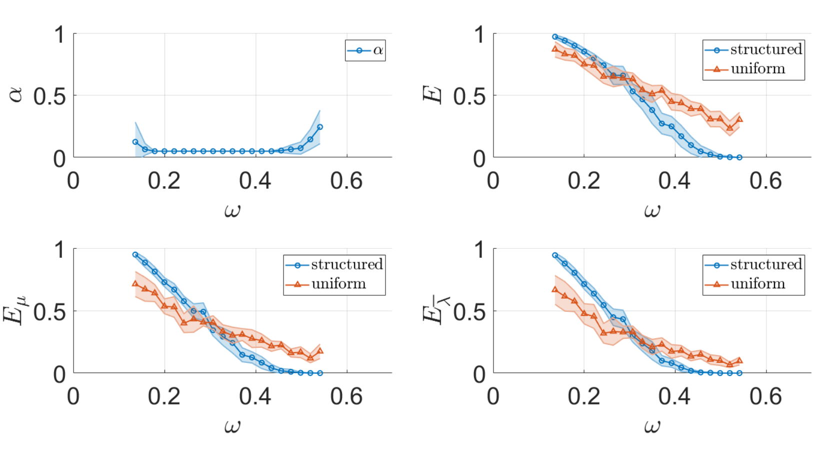
- 30×30 rank 5 matrix generated as product of sparse matrices with nonzero entries sampled uniformly from $[0, 1]$
- each trial consists of sampling, completion, and inference on original and completed matrices
  - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively
  - ℓ_1 regularization parameter α is chosen in $\{0.05, 0.1, 0.2, \ldots, 0.5\}$ to minimize matrix recovery error
- matrix recovery error and inference errors averaged over 10 trials

Synthetic Data



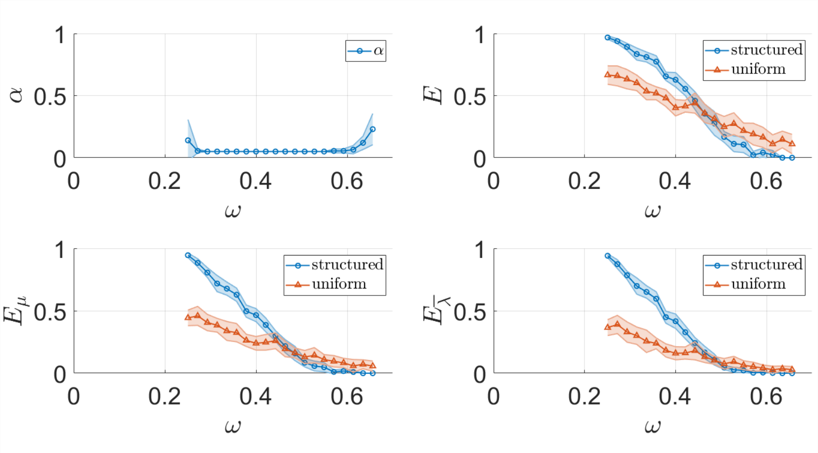
- ▷ $p_0 = 0$
- ▷ ω is proportion of entries sampled

Synthetic Data



- $p_0 = 0.2$
- ω is proportion of entries sampled

Synthetic Data



- $p_0 = 0.4$
- ω is proportion of entries sampled

Experimental Design - MyLymeData

- complete 30×16 submatrix of MyLymeData

Experimental Design - MyLymeData

- complete 30×16 submatrix of MyLymeData
- each trial consists of sampling, completion, and inference on original and completed matrices

Experimental Design - MyLymeData

- complete 30×16 submatrix of MyLymeData
- each trial consists of sampling, completion, and inference on original and completed matrices
    - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively

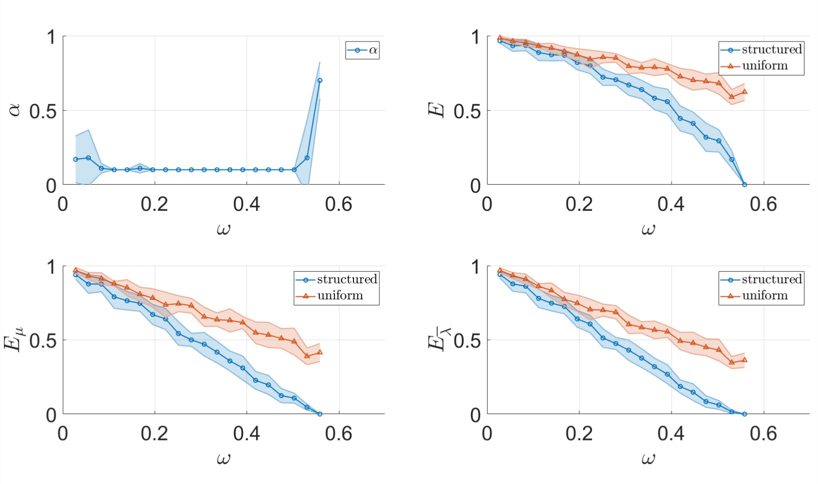
Experimental Design - MyLymeData

- complete 30×16 submatrix of MyLymeData
- each trial consists of sampling, completion, and inference on original and completed matrices
  - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively
  - ℓ_1 regularization parameter α is chosen in $\{0.05, 0.1, 0.2, \ldots, 0.5\}$ to minimize matrix recovery error

Experimental Design - MyLymeData

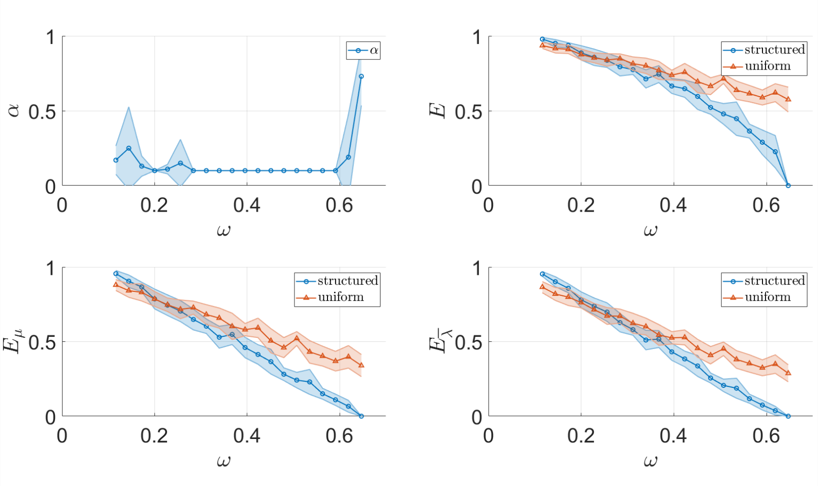
- $\triangleright$ complete 30×16 submatrix of MyLymeData
- $\triangleright$ each trial consists of sampling, completion, and inference on original and completed matrices
  - matrix is sampled via uniform sampling and structured sampling (with listed p_0), and completed with NNM and ℓ_1-NNM respectively
  - ℓ_1 regularization parameter α is chosen in $\{0.05, 0.1, 0.2, \ldots, 0.5\}$ to minimize matrix recovery error
- $\triangleright$ matrix recovery error and inference errors averaged over 10 trials

MyLyme Data



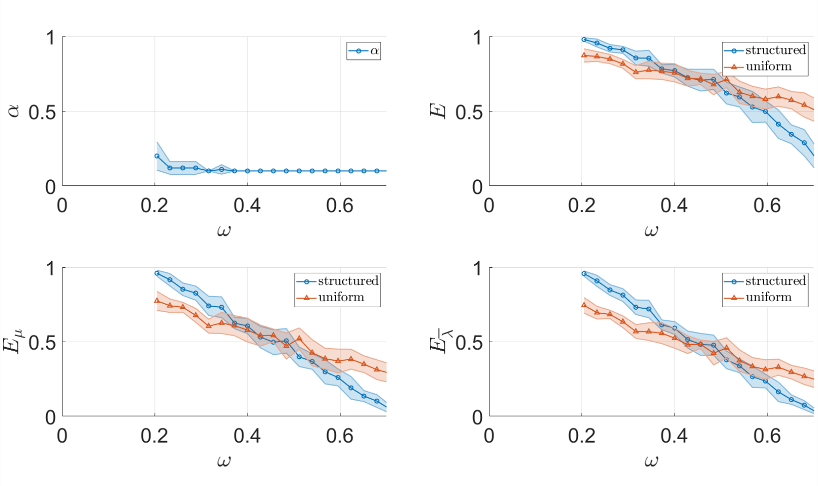
- $p_0 = 0$
- ω is proportion of entries sampled

MyLyme Data



- $p_0 = 0.2$
- ω is proportion of entries sampled

MyLyme Data



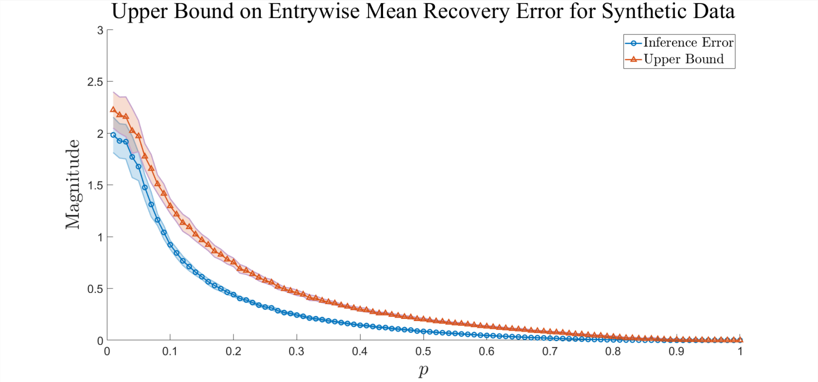
- ▷ $p_0 = 0.4$
- ▷ ω is proportion of entries sampled

Preliminary Error Bounds

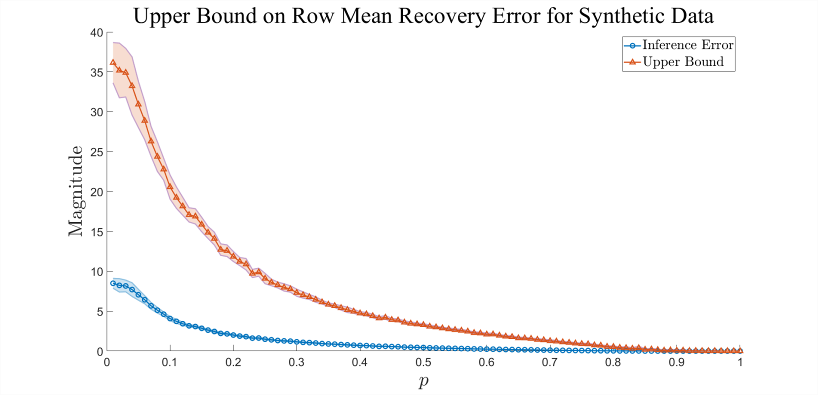
Inference	Error Bound
Entrywise Mean	$\lvert\overline{\lambda}(\mathbf{M}) - \overline{\lambda}(\hat{\mathbf{M}})\rvert \leq (mn)^{-\frac{1}{q}} \lVert\mathbf{M} - \hat{\mathbf{M}}\rVert_q$
Row Mean	$\lVert\mu(\mathbf{M}) - \mu(\hat{\mathbf{M}})\rVert_q \leq \left(\frac{n^{q-1}}{m}\right)^{\frac{1}{q}} \lVert\mathbf{M} - \hat{\mathbf{M}}\rVert_q$

- $\mathbf{M} \in \mathbb{R}^{m \times n}$
- recovered matrix, $\hat{\mathbf{M}}$

Entrywise Mean Simulation



Row Mean Simulation



Conclusions and Future Directions

- inference errors can be smaller than the associated matrix recovery errors

Conclusions and Future Directions

- inference errors can be smaller than the associated matrix recovery errors

- structured sampling and ℓ_1-NNM often results in better matrix and inference recovery than uniform sampling and NNM

Conclusions and Future Directions

- inference errors can be smaller than the associated matrix recovery errors
- structured sampling and ℓ_1-NNM often results in better matrix and inference recovery than uniform sampling and NNM
- develop exact recovery guarantees for ℓ_1-NNM on matrices with observed entries selected via structured sampling

References and Acknowledgements

[Candès and Recht, 2009] Emmanuel J. Candès and Benjamin Recht (2009)
Exact Matrix Completion via Convex Optimization
Foundations of Computational Mathematics 9, 771 – 772.

[Molitor and Needell, 2018] Denali Molitor and Deanna Needell (2018)
Matrix Completion for Structured Observations
arXiv preprint arXiv:1801.09657

[Eldèn, 2007] Lars Eldèn
Matrix Methods in Data Mining and Pattern Recognition, 69
Society for Industrial and Applied Mathematics, Philadelphia, 2007

Thank you to Professor Andrea Bertozzi, Dr. Anna Ma, Lorraine Johnson (LDo CEO), and the patients who contributed to the MyLymeData database!

Thanks!

Questions?