On Inferences from Completed Data

Jamie Haddock   Denali Molitor   Deanna Needell   Sneha Sambandam   Joy Song   Simon Sun

Abstract—Matrix completion has become an extremely important technique as data scientists are routinely faced with large, incomplete datasets on which they wish to perform statistical inferences. We investigate how error introduced via matrix completion affects statistical inference. Furthermore, we prove recovery error bounds which depend upon the matrix recovery error for several common statistical inferences. We consider matrix recovery via nuclear norm minimization and a variant, ℓ_1-regularized nuclear norm minimization for data with a structured sampling pattern. Finally, we run a series of numerical experiments on synthetic data and real patient surveys from MyLymeData, which illustrate the relationship between inference recovery error and matrix recovery error. These results indicate that exact matrix recovery is often not necessary to achieve small inference recovery error.

I. BACKGROUND AND MOTIVATION

Real-world data is often high-dimensional and incomplete; e.g., a survey may be incomplete because respondents may skip questions or as a consequence of the structure of the survey. In recent years, much work has been invested towards determining efficient and accurate methods for *data completion* [4], [9], [10], [11]. Often, however, data practitioners are not interested in any particular missing entry or the completed data itself, but in performing statistical inferences on the completed data set (e.g., entrywise mean, regression, classification) [1]. For this reason, we study how missing and artificially completed data introduces error into the recovery of statistical inferences.

In general, incomplete data can be modeled as a matrix with subsampled entries. Matrix completion results typically assume that the entries are sampled uniformly. Unfortunately, this is invalid in many practical situations. We consider various *sampling* strategies which select entries from a complete matrix to construct an incomplete matrix. Entries of a data matrix could be selected using *uniform sampling*; that is, each entry could be sampled with equal probability as in [3]. On the other hand, one could employ *structured sampling* and select entries with probabilities that depend upon their value as in [8]. The details of these two sampling methods are given in Section I-B. Such strategies can be used to model the ways that incomplete data appears in the real world. For instance, we consider a structured sampling strategy in which entries of smaller magnitude are sampled less often. This models the situation in which survey participants are more likely to skip questions that are not important to them (in which their answers may have smaller magnitude). There are numerous other works studying matrix completion techniques

under various sampling strategies; see e.g., [5] and references therein.

If the matrix to be recovered is low rank, one can accurately infer the missing entries of the data matrix under various conditions. Indeed, Candès and Recht show that if the observed sample of entries is uniformly distributed and sufficiently large then one can exactly recover the matrix via nuclear norm minimization [3]. There are many matrix completion approaches, however we focus on nuclear norm minimization (NNM) and ℓ_1-regularized nuclear norm minimization (ℓ_1-NNM), defined in Subsection I-A.

Data completion can also be helpful for data collection purposes; only partial information may be required for data completion to preserve the statistical properties of a dataset, allowing for reduction in the quantity of data that must be collected, stored, or transmitted. Returning to the survey example, one could ask respondents a small selection of questions from a larger set of candidate questions, predict their answers to the unasked questions using data completion, and apply inference methods to the recovered dataset. These applications are of particular interest to LymeDisease.org, an advocacy organization that collects survey data from Lyme patients through studies like MyLymeData [7]. The surveys used in MyLymeData branch, presenting different sets of questions to respondents based on their previous answers. Patients may also skip questions. The resulting data matrix, in which rows correspond to patients and columns correspond to questions, is highly incomplete. Another concern of LymeDisease.org is the length of the MyLymeData surveys, since overlong surveys can cause survey fatigue and lead patients to ignore questions or answer inaccurately. Developing sound inference methods for incomplete data would allow us to sample strategically and use data completion techniques to design shorter surveys that preserve high-level information about the respondents.

In this report, we study the effects of different sampling techniques on statistical inference. We derive provable error bounds for certain statistics and run numerical simulations on synthetic data as well as large-scale, incomplete survey data from MyLymeData with the goal of reducing the amount of data required from each survey respondent while preserving population-level insights.

A. Notation

We begin by establishing notation that will be used throughout the paper. Recall that $[n] = \{1, 2, ..., n\}$. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, we denote the (i, j) entry of $\mathbf{A}$ as A_{ij} and the ith row of

A as $\mathbf{a}_i$. The standard ℓ_q-norm on $\mathbb{R}^n$ is denoted $\|\cdot\|_q$ for $1 \le q \le \infty$. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\|\mathbf{A}\|_q$ is the entrywise matrix q-norm; i.e., the ℓ_q-norm of the vectorization of $\mathbf{A}$. The matrix nuclear norm is denoted $\|\mathbf{A}\|_* = \text{trace}(\sqrt{\mathbf{A}^*\mathbf{A}})$.

We consider two sampling strategies, uniform and structured sampling. For uniform sampling, the probability of sampling each entry is given by $p \in (0,1)$. We also investigate a structured sampling strategy in which the probability of sampling entries equal to zero is given by p_0, and the probability of sampling nonzero entries is given by p_1; we assume $p_0 < p_1$.

We denote the original complete matrix by $\mathbf{M}$, the set of observed indices from the original matrix by $\Omega \subset [m] \times [n]$, the unobserved indices by Ω^C, the observed matrix by $\mathbf{M}_\Omega$, the recovered matrix by $\widetilde{\mathbf{M}}$, and the fraction of entries which are observed as $\omega \in (0,1)$. We consider two recovery methods, nuclear norm minimization (NNM) and ℓ_1-regularized nuclear norm minimization (ℓ_1-NNM). The recovered matrix $\widetilde{\mathbf{M}}$ for NNM is defined as

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

The recovered matrix $\widetilde{\mathbf{M}}$ for ℓ_1-NNM is defined as

$$\underset{\mathbf{X} \in \mathbb{R}^{m \times n}}{\arg\min} \|\mathbf{X}\|_* + \alpha\|\mathbf{X}_{\Omega^C}\|_1 \text{ s.t. } M_{ij} = X_{ij} \text{ for all } (i,j) \in \Omega$$

for some regularization parameter $\alpha > 0$. The addition of the ℓ_1-regularization term in the objective of ℓ_1-NNM encourages unobserved entries of the recovered matrix to be near 0, which makes it a natural choice for recovery on an incomplete matrix generated by structured sampling [8].

The inferences we consider are basic statistics. The first inference is the *entrywise mean*, defined as $\bar{\lambda}(\mathbf{A}) := \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n A_{ij}$. We additionally consider the *row mean*, a row vector containing the mean value for each column or feature, which is defined as $\mu(\mathbf{A}) := \frac{1}{m}\sum_{i=1}^m \mathbf{a}_i$. In terms of our application to health survey data consisting of patient responses to questions regarding their symptoms, these inferences could be regarded as an average "wellness" score for a group of patients (entrywise mean) or the average responses of a group of patients (row mean).

B. Methodology

To perform our experiments, we begin with a complete matrix $\mathbf{M}$ either artificial or extracted from real data, which we take as the ground truth. We then use either the uniform or structured sampling strategies to obtain an incomplete observed matrix, $\mathbf{M}_\Omega$. The values of p and p_0, p_1 used for uniform and structured sampling respectively are noted in each experiment. We recover $\widetilde{\mathbf{M}}$ via either NNM or ℓ_1-NNM. For $\mathbf{M}_\Omega$ constructed via the uniform sampling strategy, we use NNM to recover $\widetilde{\mathbf{M}}$ while for $\mathbf{M}_\Omega$ constructed via the structured sampling strategy, we use ℓ_1-NNM to recover $\widetilde{\mathbf{M}}$. Here, we choose the regularization parameter α optimally from among $\{0.05, 0.1, 0.2, ..., 0.5\}$ to minimize the resulting

error $\|\mathbf{M} - \widetilde{\mathbf{M}}\|_F$. We use the alternating direction method of multipliers (ADMM) to solve both NNM and ℓ_1-NNM. The details of the implementation of ADMM for NNM can be found in [6]. To apply ADMM to solve ℓ_1-NNM, we rewrite the ℓ_1-NNM model as

$$\begin{aligned}\min \quad & \|\mathbf{X}\|_* + \alpha\|\mathbf{Y}\|_1 \\ s.t. \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \\ & P_{\Omega^C}(\mathbf{X}) = P_{\Omega^C}(\mathbf{Y}).\end{aligned}$$

The augmented Lagrangian function of this program is

$$\begin{aligned} L_t(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = & \|\mathbf{X}\|_* + \alpha\|\mathbf{Y}\|_1 + \text{Tr}[\mathbf{Z}_1^T(P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{M}))] \\ & + \text{Tr}[\mathbf{Z}_2^T(P_{\Omega^C}(\mathbf{X}) - P_{\Omega^C}(\mathbf{Y}))] \\ & + \frac{t}{2}[\|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{M})\|_F^2 + \|P_{\Omega^C}(\mathbf{X}) - P_{\Omega^C}(\mathbf{Y})\|_F^2] \end{aligned}$$

where $\mathbf{Z}_1 = P_\Omega(\mathbf{Z})$, $\mathbf{Z}_2 = P_{\Omega^C}(\mathbf{Z})$, and the iteration steps are

$$\begin{aligned} \mathbf{X}^{k+1} &= \arg\min_{\mathbf{X}} L_t(\mathbf{X}, \mathbf{Y}^k, \mathbf{Z}^k), \\ \mathbf{Y}^{k+1} &= \arg\min_{\mathbf{Y}} L_t(\mathbf{X}^{k+1}, \mathbf{Y}, \mathbf{Z}^k), \\ \mathbf{Z}_1^{k+1} &= \mathbf{Z}_1^k + t(P_\Omega(\mathbf{X}^{k+1}) - P_\Omega(\mathbf{M})), \\ \mathbf{Z}_2^{k+1} &= \mathbf{Z}_2^k + t(P_{\Omega^C}(\mathbf{X}^{k+1}) - P_{\Omega^C}(\mathbf{Y}^{k+1})). \end{aligned}$$

We consider the normalized matrix recovery error $E(\mathbf{M}, \widetilde{\mathbf{M}}) := \|\mathbf{M} - \widetilde{\mathbf{M}}\|_F / \|\mathbf{M}\|_F$ as an estimate of the error introduced by sampling and data completion. Finally, we compute inferences on the original matrix $\mathbf{M}$ and the recovered matrix $\widetilde{\mathbf{M}}$. We estimate the inference error between these two matrices via various measures. We define the *absolute error of the entrywise mean* as $E_{\bar{\lambda}}(\mathbf{M}, \widetilde{\mathbf{M}}) := |\bar{\lambda}(\mathbf{M}) - \bar{\lambda}(\widetilde{\mathbf{M}})|$ and the *normalized error of the row mean* as $E_\mu(\mathbf{M}, \widetilde{\mathbf{M}}) := \|\mu(\mathbf{M}) - \mu(\widetilde{\mathbf{M}})\|_2 / \|\mu(\mathbf{M})\|_2$.

We perform numerical experiments on both synthetic and real-world data. The real-world dataset consists of survey data from the MyLymeData patient study conducted by LymeDisease.org [7]. For experiments on synthetic data, we generate artificial matrices as follows. To guarantee a certain rank r, we generate $m \times n$ scalar matrices by multiplying two matrices whose sizes are $m \times r$ and $r \times n$. The entries of each pair of matrices we generate are uniformly distributed integers within the range $[0, C]$. For experiments on real data, we extract a complete portion of MyLymeData consisting of patient responses to questions regarding their symptoms and health history.

II. EXPERIMENTAL RESULTS

In Figures 1, 2, and 3, we plot experimentally collected matrix and inference recovery errors on synthetic matrices; the figures differ by the choice of zero sampling probability p_0 for the structured sampling strategy. We generate a 30×30 matrix with rank 5 as described in Subsection I-B. For various p and (p_0, p_1) sampling probabilities, we measure the resulting matrix recovery errors and inference recovery

errors. These results are averaged over 10 trials (each trial consists of a sample of observed entries) and plotted with the standard deviation of these errors. Errors are plotted versus the proportion of observed entries ω. We additionally record the optimal regularization parameter α which resulted in the smallest matrix recovery error for the given structured sampling proportion ω in the plots in the upper left of each figure.
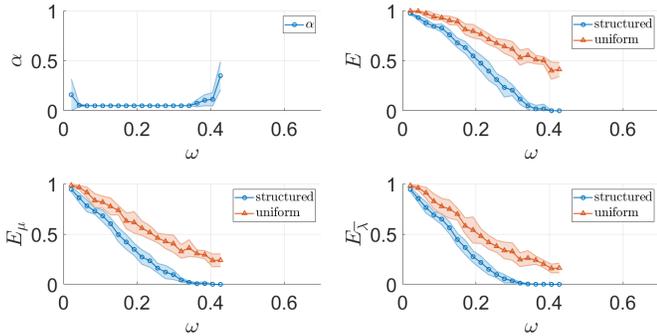

Fig. 1. Recovery errors for unif. sampling with NNM and structured sampling with $p_0 = 0$ (no entries equal to zero are sampled) and ℓ_1-NNM on synthetic data. Upper left: optimal regularization parameter α for observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.
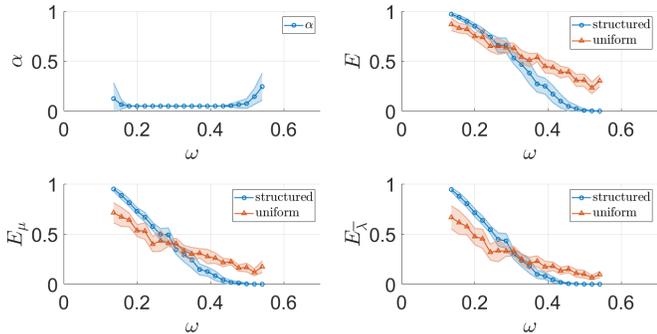

Fig. 2. Recovery errors for unif. sampling with NNM and structured sampling with $p_0 = 0.2$ and ℓ_1-NNM on synthetic data. Upper left: optimal regularization parameter α for observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.

In Figures 4, 5, and 6, we plot experimentally collected matrix and inference recovery errors on MyLymeData matrices; the figures differ by the choice of zero sampling probability p_0 for the structured sampling strategy. We select a complete matrix of size 30×16 by selecting the 16 questions (columns) every patient must answer and select the 30 patients with the most zero entries. For various p and (p_0, p_1) sampling probabilities, we measure the resulting matrix recovery errors and inference recovery errors. These results are averaged over 10 trials (each trial consists of a sample of observed entries) and plotted with the standard deviation of these errors. Errors
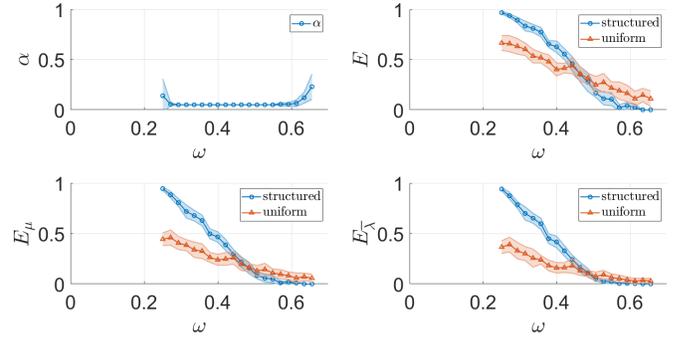

Fig. 3. Recovery errors for uniform sampling with NNM and structured sampling with $p_0 = 0.4$ and ℓ_1-NNM on synthetic data. Upper left: optimal regularization parameter α for observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.
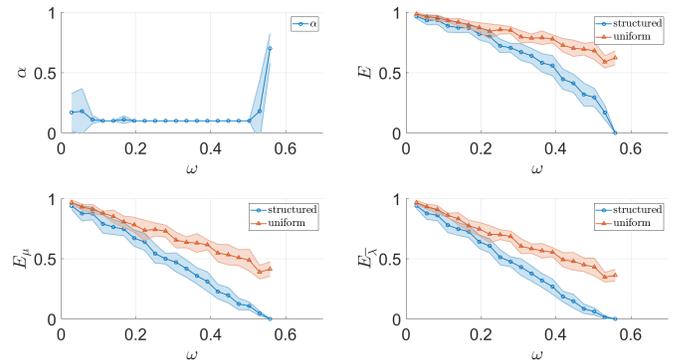

Fig. 4. Recovery errors for uniform sampling with NNM and structured sampling with $p_0 = 0$ with ℓ_1-NNM on MyLymeData. Upper left: optimal regularization parameter α for various observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.

are plotted versus the proportion of observed entries ω. We additionally record the optimal regularization parameter α which resulted in the smallest matrix recovery for the given structured sampling proportion ω error in the plots in the upper left of each figure.

Note that in Figures 1, 2, 3, 4, and 5, the optimal regularization parameter α is greater than zero for sufficiently large observation proportion ω. Furthermore, in Figures 1, 2, 3, 4, and 5, the ℓ_1-NNM recovered solution is exact for sufficiently large ω, and the ℓ_1-NNM recovery for the observations sampled via the structured strategy is more accurate than the NNM recovery for the observations sampled via the uniform strategy for larger proportion ω. Finally, the inference recoveries are often exact for smaller ω than is necessary for exact matrix recovery, as in Figure 1, 2, and 3.

III. THEORETICAL RESULTS

Given that the matrix recovery error has been studied closely in the literature [3], [2], we aim to bound the inference recovery error by a function of the matrix recovery error.
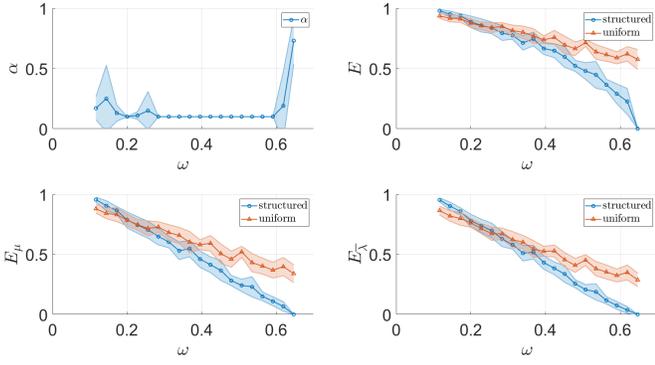
Fig. 5. Recovery errors for uniform sampling with NNM and structured sampling with $p_0 = 0.2$ with ℓ_1-NNM on MyLymeData. Upper left: optimal regularization parameter α for various observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.
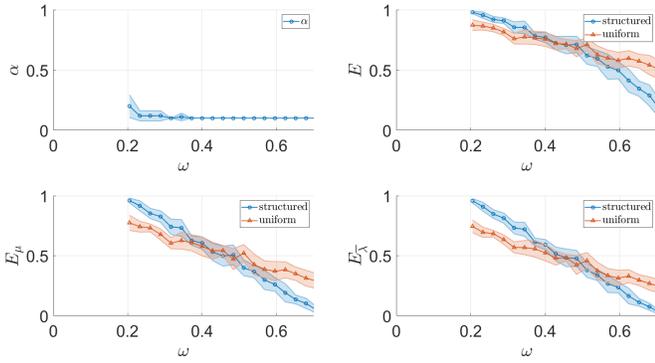

Fig. 6. Recovery errors for uniform sampling with NNM and structured sampling with $p_0 = 0.4$ with ℓ_1-NNM on MyLymeData. Upper left: optimal regularization parameter α for various observation proportions ω; upper right: normalized matrix recovery errors E; lower left: normalized error of the row mean E_μ; lower right: absolute error of the entrywise mean $E_{\bar{\lambda}}$.

Specifically, we establish bounds on the recovery error for the entrywise mean and row mean.

The first result bounds the recovery error of the entrywise mean $\bar{\lambda}$ and the row mean μ by a scalar multiple of the matrix recovery error. Recall that $\|\mathbf{A}\|_q$ denotes the standard ℓ_q vector-norm of the vectorization of the matrix $\mathbf{A}$.

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

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

and

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

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

Proof. First, note that $\bar{\lambda}$ and μ are linear operators, so it suffices to show that $|\bar{\lambda}(\mathbf{A})| \leq (mn)^{-1/q} \|\mathbf{A}\|_q$ and $\|\mu(\mathbf{A})\|_q \leq (n^{q-1}/m)^{1/q} \|\mathbf{A}\|_q$ for $\mathbf{A} \in \mathbb{R}^{m \times n}$. Next, note

that $|\bar{\lambda}(\mathbf{A})| \leq \|\mathbf{A}\|_1 / mn$. Applying Hölder's inequality, we have

$$|\bar{\lambda}(\mathbf{A})| \leq \frac{1}{mn} \|\mathbf{A}\|_1 \leq (mn)^{-\frac{1}{q}} \|\mathbf{A}\|_q$$

where $1/q$ assumes the value 0 if $q = \infty$.

Next, note that

$$\|\mu(\mathbf{A})\|_q^q = \frac{1}{m^q} \sum_{j=1}^{n} \left| \sum_{i=1}^{m} A_{ij} \right|^q \leq \frac{1}{m^q} \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |A_{ij}| \right)^q$$

$$= \frac{\|\mathbf{A}\|_1^q}{m^q} \leq \frac{n^{q-1} \|\mathbf{A}\|_q^q}{m}$$

where the last inequality follows from Hölder's inequality. $\square$

In Figure 7, we explore the bounds given in Theorem III.1. We generate 20 random scalar matrices of size 16×80 as described in Subsection I-B. For each matrix, we collect 20 uniform samples of the entries using the sampling probability p, then calculate the averages of the entrywise mean recovery error, the row mean recovery error, and the derived upper bounds based on the matrix recovery error for each sample. We perform this process for $p = 0, 0.01, ..., 1$.
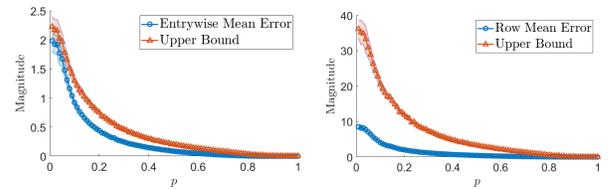

Fig. 7. The averages of the 400 sampled inference recovery errors and the derived upper bounds for uniform observation sampling probabilities from 0 to 1, with step size 0.01. Left: entrywise mean error; right: row mean error.

Finally, we present a simple analytic bound for NNM matrix recovery error. This result holds for any sampling strategy, even deterministic strategies. Note that this bound along with the previous result offers a bound on the inference recovery errors even if the matrix recovery is not exact.

Theorem III.2. *Let $\mathbf{M} \in \mathbb{R}^{m \times n}$, Ω, and $\widetilde{\mathbf{M}}$ be computed via NNM as described in Subsection I-A. Let $r = \mathrm{rank}(\mathbf{M})$ denote the rank of $\mathbf{M}$, and denote the singular values of $\mathbf{M}$ by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ in decreasing order. Then*

$$\|\mathbf{M} - \widetilde{\mathbf{M}}\|_F \leq 2\sqrt{r^2 \sigma_1^2 - \|\mathbf{M}_\Omega\|_F^2}. \tag{1}$$

Proof. Applying the Parallelogram Identity, we have

$$\|\mathbf{M} - \widetilde{\mathbf{M}}\|_F^2 = 2 \left(\|\mathbf{M}\|_F^2 + \|\widetilde{\mathbf{M}}\|_F^2 \right) - \|\mathbf{M} + \widetilde{\mathbf{M}}\|_F^2.$$

We bound each term of the right-hand side, beginning with the $\|\mathbf{M}\|_F^2$ term. By Hölder's Inequality, we have

$$\|\mathbf{M}\|_F^2 = \|(\sigma_1, \sigma_2, \ldots, \sigma_r)\|_2^2 \leq r^2 \sigma_1^2.$$

Next, we bound the $\|\widetilde{\mathbf{M}}\|_F^2$ term above. Since $\mathbf{M}$ is feasible for the nuclear norm minimization problem, note that $\|\widetilde{\mathbf{M}}\|_* \leq$

$\|\mathbf{M}\|_*$. Therefore, through repeated use of Hölder's Inequality, we calculate that

$$\|\widetilde{\mathbf{M}}\|_F^2 \leq \|\widetilde{\mathbf{M}}\|_*^2 \leq \|\mathbf{M}\|_*^2 = \|\,(\sigma_1, \sigma_2, \ldots, \sigma_r)\,\|_1^2 \leq r^2 \sigma_1^2.$$

Finally, note that

$$\|\mathbf{M} + \widetilde{\mathbf{M}}\|_F^2 \geq \|\mathbf{M}_\Omega + \widetilde{\mathbf{M}}_\Omega\|_F^2 = \sum_{(i,j)\in\Omega} (2M_{i,j})^2 = 4\|\mathbf{M}_\Omega\|_F^2.$$

$\square$

Note that this bound guarantees exact recovery when all entries of the matrix are observed and all singular values of the matrix are equal, but is likely not tight for many situations when exact recovery can be guaranteed by e.g., [3], [2].

IV. CONCLUSION

In this work, we explored how error introduced by data completion affects recovery of statistical inferences. Our numerical experiments demonstrate that simple inferences such as the entrywise mean or the row mean can be recovered accurately even when the matrix is not recovered exactly. 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.

Future directions include exploring more common statistical inferences, such as support vector machine models. Additionally, we hope to develop a better analytic bound on the matrix recovery error which generalizes the exact recovery results in the literature. Furthermore, we will explore theory for exact recovery via ℓ_1-NNM for matrices whose observations are sampled via the structured sampling strategy.

V. ACKNOWLEDGEMENTS

REFERENCES

[1] A. L. Bello. Imputation techniques in regression analysis: looking closely at their implementation. *Computational statistics & data analysis*, 20(1):45–57, 1995.

[2] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.

[3] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

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

[5] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–3034, 2015.

[6] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

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

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

[9] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[10] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[11] S. Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.