Received 9 October 2020; revised 23 February 2021; accepted 18 March 2021. Date of publication 29 March 2021; date of current version 28 April 2021. The review of this article was arranged by Associate Editor Prof. Laura Balzano.

Digital Object Identifier 10.1109/OJSP.2021.3069373

Optimal Recovery of Missing Values for Non-Negative Matrix Factorization

REBECCA CHEN DEAN AND LAV R. VARSHNEY (Senior Member, IEEE)

Coordinated Science Laboratory and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

CORRESPONDING AUTHOR: REBECCA DEAN (e-mail: rebtchen@gmail.com)

This work was presented in part at the 2019 IEEE Data Science Workshop [1], and was supported in part by Air Force STTR under Grant FA8650-16-M-1819, and in part by Grant no. 2018-182794 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation.

ABSTRACT Missing values imputation is often evaluated on some similarity measure between actual and imputed data. However, it may be more meaningful to evaluate downstream algorithm performance after imputation than the imputation itself. We describe a straightforward unsupervised imputation algorithm, a minimax approach based on optimal recovery, and derive probabilistic error bounds on downstream non-negative matrix factorization (NMF). Under certain geometric conditions, we prove upper bounds on NMF relative error, which is the first bound of this type for missing values. We also give probabilistic bounds for the same geometric assumptions. Experiments on image data and biological data show that this theoretically-grounded technique performs as well as or better than other imputation techniques that account for local structure. We also comment on imputation fairness.

INDEX TERMS Clustering, error bound, missing values, non-negative matrix factorization.

I. INTRODUCTION

Performance of missing values imputation is typically measured by how similar imputed data is to original data, but Tuikkala *et al.* argue “the success of preprocessing methods should ideally be evaluated also in other terms, for example, based on clustering results and their biological interpretation, that are of more practical importance for the biologist” [2]. Several groups have evaluated downstream impacts of different imputation methods on clustering, regression, and classification [3]–[6]. Missing values for non-negative matrix factorization (NMF) have been studied for the application of stock price prediction, but previous approaches lack theoretical guarantees [7].

In the first presentation of this work [1], we extended Liu and Tan’s NMF *worst-case* error bound [8] to account for simple minimax imputation, which experimentally showed competitive performance with more complicated imputation techniques. Additionally, in this long version, we find several *probabilistic* error bounds which better characterize experimental results and serve as useful benchmarks for algorithms. We make no statistical assumptions on the missingness pattern for the worst-case bound [9], except that there is at least one fully observed data point per cluster. We assume samples are missing completely at random (MCAR) for our probabilistic bounds. Such theoretical bounds on downstream algorithms after imputation have not been previously found. Finally, we introduce new discussions on how our minimax approach aligns with certain notions of *fairness*.

Data often exhibits local structure, e.g., different groups of cells follow different gene expression patterns. Information about local structure can be used to improve imputation. We introduce a new imputation method based on *optimal recovery*, an approximation-theoretic approach for estimating linear functionals of a signal [10]–[12] previously applied in signal and image interpolation [13]–[15], to perform matrix imputation of clustered data. (Note that *optimal recovery* is simply the name of the minimax optimization method.) Theoretically characterizing optimal recovery for missing value imputation requires a new geometric analysis technique. Previous work on missing values take a statistical approach rather than a geometric one.

Our contributions include:

- A simple imputation algorithm that performs as well as or better than other imputation methods, as demonstrated on hyperspectral remote sensing data and biological data;

- A worst-case upper bound on the relative error of downstream analysis by NMF. This is the first such error bound for settings with missing values; and
- A probabilistic bound on NMF error after imputation that is predictive of algorithmic performance in typical settings.

The remainder of the paper is organized as follows. In Section II, we give background on missing data mechanisms, imputation algorithms, and NMF. In Section III, we introduce optimal recovery and apply it to NMF. In Section IV, we present an algorithm for optimal recovery imputation of clustered data and give a deterministic upper bound on algorithm performance. In Section V we give a probabilistic bound on the performance of our algorithm. In Section VI we give experimental results for both synthetic and real data, and we conclude in Section VII.

II. BACKGROUND

In this section, we describe the relationships between missingness patterns and the underlying data, which are referred to as *missing data mechanisms*. We then discuss prior work on imputation algorithms, and we present NMF, which is commonly performed after imputation in biological settings.

A. MISSINGNESS MECHANISMS

Rubin originally described three mechanisms that may account for missing values in data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [16]. When data is MCAR, the missing data is a random subset of all data, and the missing and observed values have similar distributions [17]. When data is MAR, the distribution of missing data is dependent on the observed data. For example, in medical records, patients with normal blood pressure levels are more likely to have missing values for glucose levels than patients with high blood pressure. When data is MNAR, the distribution of missing data is dependent on the unobserved (missing) data. For example, people with very high incomes may be less likely to report their incomes.

It is important to understand the missingness mechanism when analyzing data. When data is MCAR, the statistics (e.g. mean, variance, covariance) of the complete cases (data points with no missing observations) will represent the statistics of the entire dataset, but the sample size will be much smaller [18]. If data is MAR or MNAR, the complete cases may be a biased representation of the dataset. Although some research has been done on MNAR imputation, this is generally a difficult problem, and most imputation methods assume the MAR or MCAR model.

B. BASIC IMPUTATION ALGORITHMS

Imputation is often necessary before specific downstream analysis, such as clustering or manifold-finding for classification. Two main categories of imputation are single imputation, in which missing values are imputed once, and multiple imputation, in which missing values are imputed multiple times with some built-in randomness.

One of the simplest single imputation techniques is *mean imputation.* Missing values of each variable are imputed with the mean value of that variable. In *regression imputation*, a variable of interest is regressed on the other variables using the complete cases. Imputation puts points with missing values directly on the regression line. Bayesian imputation approaches also exist, including *Bayesian PCA* [19] and *maximum likelihood imputation* [20]. Bayesian methods are theoretically sound and assume that data samples are generated from some underlying joint distribution. In practice, these methods require numerical algorithms such as the Markov chain Monte Carlo method, which may be prohibitively time-consuming for large datasets. Another prevalent approach to data imputation uses matrix completion methods under the assumption that data is low-rank [21], [22].

Multiple imputation attempts to preserve the variance/covariance matrix of the data. Several imputations are randomly generated, resulting in multiple complete datasets. One popular algorithm for multiple imputation is *multiple imputation by chained equations* (MICE) [23]. While MICE does not have the theoretical backing that maximum likelihood imputation has, MICE is flexible and can accommodate known interactions and independencies of real-world datasets [24].

C. IMPUTATION WITH CLUSTERED DATA

When the underlying data is clustered, a data point should be imputed based on its cluster membership. Local imputation approaches outperform global ones when there is local structure in data. Global approaches generally perform some form of regression or mean matching across all samples [9], [23], whereas local approaches group subsets of similar samples. Popular imputation algorithms that utilize local structure include k-nearest neighbors (kNN), local least squares (LLSimpute), and bicluster Bayesian component analysis (biBPCA) [25]–[27]. The kNN imputation method finds the k closest neighbors of a sample with missing values (measured by some distance function) and fills in the missing values using an average of its neighbors. LLSimpute uses a multiple regression model to impute the missing values from k nearest neighbors. Rather than regressing on *all* variables, biBPCA performs linear regression using biclusters of a lower-dimensional space, i.e. coherent clusters consisting of correlated variables under correlated experimental conditions. Delalleau *et al.* develop an algorithm to train Gaussian mixtures with missing data using expectation-maximization (EM) [28]. By itself, MICE does not address clusters, but cluster-specific (group-wise) regression can be performed [29].

Interestingly, Tuikkala *et al.* found that even when *imputation accuracy* varies across methods, *clustering accuracy* after imputation does not vary that much [2]. They show that after imputing with BPCA, LLS, and kNN, clustering results are similar. Chiu *et al.* find that LLS-like algorithms performed better than kNN-like algorithms in terms of downstream clustering accuracy [3]. De Souto *et al.* evaluate whether the effect

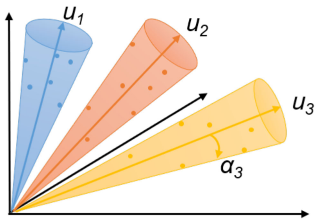

FIGURE 1. Clustered data separable by NMF.

of different imputation methods on clustering and classification are statistically significant [4]. They remove all genes with more than 10% missing values and after imputation, they find that simple methods such as mean and median imputation perform as well as weighted kNN and BPCA. However, while downstream analysis after imputation has been studied empirically, it has never been characterized theoretically. Works that have given theoretical error bounds for data with missing values do not consider error after imputation (e.g. linear regression error of data with missing values [5]).

D. NON-NEGATIVE MATRIX FACTORIZATION (NMF)

Matrix factorization is commonly used for clustering and dimensionality reduction in computational biology, imaging, and other fields. NMF is particularly favored by engineers and biologists because non-negativity constraints preclude negative values that are difficult to interpret in biological processes [30], [31]. A recent tutorial article highlighted the interpretability and identifiability (or model uniqueness) of NMF, both of which are valuable for practical applications [32]. Furthermore, experiments demonstrate that the latent factors are intuitive given the data [32]. In biology, NMF of gene count matrices can discover cell groups and lower-dimensional manifolds (latent factors) describing gene count ratios for different cell types. Due to channel noise, incomplete survey data, or biological limitations, however, data matrices are usually incomplete and matrix imputation is often necessary before further analysis [31]. In particular, Stein-O'Brien et al. argue that "newer MF algorithms that model missing data are essential for [single-cell RNA sequence] data" [33].

Donoho and Stodden interpret NMF as the problem of finding cones in the positive orthant which contain clouds of data points [34], see also [35] for earlier work. Tan and Févotte consider NMF with missing values, but they replace missing with zeros instead of performing imputation, and they do not provide error bounds [7]. In addition, their NMF algorithm requires hyperparameter tuning and makes some probabilistic assumptions on the data.

Liu and Tan show that a rank-one NMF gives a good estimation of near-separable data and provide an upper bound on the relative reconstruction error [8]. Using the cone representation previously described, a rank-one NMF can be used to characterize data clustered in non-overlapping cones, as shown in Fig. 1. By assuming that separation between cones is substantial (described more precisely in Eq. 6), Liu and Tan derive a deterministic upper bound on NMF error, as well as a probabilistic upper bound. Given that gene and protein expression data is often linearly separable on some manifold- or high-dimensional space [36], the separability assumption is valid for real data. Taking their assumptions and original theorems (described in Section III.B), we extend their work to characterize NMF error when data contains missing values that are imputed. Using these existing bounds as a starting point, we bound performance of downstream analysis of imputation for the first time. Our proof is based on the geometry of NMF and is parameter-free.

III. OPTIMAL RECOVERY

In this section, we introduce our approach to imputation based on approximation-theoretic ideas. Suppose we are given an unknown signal v that lies in some signal class C_k. The optimal recovery estimate $\hat{v}^*$ minimizes the maximum error between an imputed vector $\hat{v}$ and all signals in the feasible signal class. Given well-clustered non-negative data $\mathbf{V}$, we impute missing samples in $\mathbf{V}$ so the maximum error is minimized over feasible clusters, regardless of the missingness pattern.

A. APPLICATION TO CLUSTERED DATA

Let $\mathbf{V} \in \mathbb{R}_+^{F\times N}$ be a matrix of N sample points with F observations (N points in F-dimensional Euclidean space). Suppose the N data points lie in K disjoint clusters C_k (where $k = 1, 2, \ldots, K$), and that these clusters are compact, convex spaces (e.g., the convex hull of the points belonging to C_k).

Now suppose there are missing values in $\mathbf{V}$. Let $\Omega \in \{0, 1\}^{F\times N}$ be a matrix of indicators with $\Omega_{ij} = 1$ if v_{ij} is observed and 0 otherwise. We make no assumptions on the missingness pattern, such as MCAR or MAR because we take a geometric approach rather than a statistical one. We define the projection operator of a matrix $\mathbf{Y}$ onto an index set Ω by

$$[P_\Omega(\mathbf{Y})]_{ij} = \begin{cases} \mathbf{Y}_{ij} & \text{if} \quad \Omega_{ij} = 1 \\ 0 & \text{if} \quad \Omega_{ij} = 0 \end{cases}.$$

We use the subscripted vector $(\cdot)_{fo}$ to denote fully observed data points (columns), or data points with no missing values, and we use the subscripted vector $(\cdot)_{po}$ to denote partially observed data points, or data with missing entries. We use a subscripted matrix $(\cdot)_{fo}$ or $(\cdot)_{po}$ to denote the set of all fully observed or partially observed data columns in the matrix.

We can impute a partially observed vector v_{po} by observing where its observed samples intersect with the clusters $C_1, \ldots, C_k$. Let the *missing values plane* be the restriction set over $\mathbb{R}^F$ that satisfies the constraints on the observed values of v_{po}. Let $\hat{v}_{po}$ be possible imputations of v_{po}. For each v_{po}, we call this intersection the *feasible set* W:

$$W = \bigcup_{k\in[K]} W_k, \tag{1}$$

where

$$W_k = \{\hat{v}_{po} \in C_k : P_\Omega(\hat{v}_{po}) = P_\Omega(v_{po})\}. \tag{2}$$

Fig. 2 illustrates the feasible set (a circle) when $F = 2$ there are two missing samples, and when the cluster (the

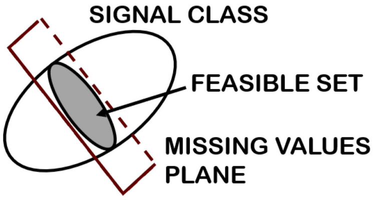

FIGURE 2. Feasible set of estimators.

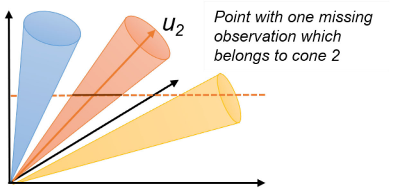

FIGURE 3. Feasible set of estimators.

signal class) is an ellipsoid. If the vector had only one missing sample, the feasible set would be a line segment (Fig. 3).

Since W cannot be empty, there must be at least one cluster for which W_k is non-empty. The optimal recovery estimator $\hat{v}^*_{po}$ minimizes the maximum error over the feasible set of estimates:

$$\hat{v}^*_{po} = \arg\min_{\hat{v}_{po}\in C_k} \max_{v\in C_k} \|\hat{v}_{po} - v\|, \tag{3}$$

where $\|\cdot\|$ denotes some norm or error function. If we use the ∞-norm, $\hat{v}^*_{po}$ is the Chebyshev center of the feasible set.

Feasible clusters are those for which W_k is not empty, and W_k are disjoint. If there are multiple non-empty W_k, we can find (3) over the cluster for which the corresponding W_k covers the largest volume: $k = \arg\max_k |W_k|$.

B. NON-NEGATIVE MATRIX FACTORIZATION

In this subsection, we describe the conical interpretation of NMF, keeping the notation and formulation used by Liu and Tan [8]. Let $\mathbf{V} \in \mathbb{R}^{F\times N}_{+}$ be a matrix of N sample points with F non-negative observations. Suppose the columns in $\mathbf{V}$ are generated from K clusters. There exist $\mathbf{W} \in \mathbb{R}^{F\times K}_{+}$ and $\mathbf{H} \in \mathbb{R}^{K\times N}_{+}$ such that $\mathbf{V} = \mathbf{WH}$. This is the NMF of $\mathbf{V}$ [37].

Suppose the N data points originate from K cones. We define a circular cone $C(u, \alpha)$ by a direction vector u and an angle $\alpha \in (0, 2\pi)$:

$$C(u, \alpha) := \left\{ x \in \mathbb{R}^F \backslash\{0\} : \frac{x \cdot u}{\|x\|_2} \geq \cos\alpha \right\}, \tag{4}$$

or equivalently,

$$C(u, \alpha) := \left\{ x \in \mathbb{R}^F \backslash\{0\} : (x \cdot u)^2 - (x \cdot x)\cos^2(\alpha) \geq 0 \right\}. \tag{5}$$

In a three-dimensional space, this conical hull is sometimes called an ice cream cone [32]. Since it is possible that a circular cone extends outside the non-negative orthant, we truncate the circular cones to be in the non-negative orthant P so that we have $C(u, \alpha) \cap P$. We can consider u_k to be the dictionary entry corresponding to C_k and all x's belonging to C_k as noisy versions of u_k. We call the angle between cones $\beta_{ij} := \arccos(u_i \cdot u_j)$. Assume the columns of $\mathbf{V}$ are in K well-separated cones, that is,

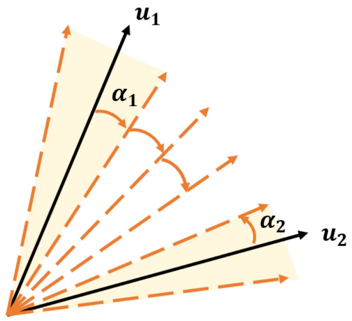

FIGURE 4. Geometric assumption for greedy clustering.

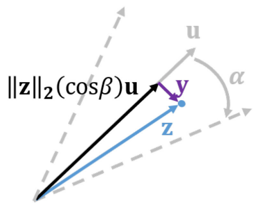

FIGURE 5. Decomposition of vectors in a circular cone.

$$\min_{i,j\in[K], i\neq j} \beta_{ij} > \max_{i,j\in[K], i\neq j} \alpha_i + 3\alpha_j. \tag{6}$$

This implies that the distance between any two points originating from the same cluster is less than the distance between any two points in different clusters, which is a common assumption used to guarantee clustering performance [38], [39] (see Fig. 4). We can then partition $\mathbf{V}$ into k sets, denoted $\mathbf{V}_k := \{\mathbf{v}_n \in C_k \cap P\}$, and rewrite $\mathbf{V}_k$ as the sum of a rank-one matrix $\mathbf{A}_k$ (parallel to u_k) and a perturbation matrix $\mathbf{E}_k$ (orthogonal to u_k). For any vector $\mathbf{z} \in \mathbf{V}_k$, $\mathbf{z} = \|\mathbf{z}\|_2(\cos\beta)\mathbf{u}_k + \mathbf{y}$, where $\|\mathbf{y}\|_2 = \|\mathbf{z}\|_2(\sin\beta) \leq \|\mathbf{z}\|_2(\sin\alpha_k)$. $\mathbf{E}_k$ is composed of the orthogonal part $\mathbf{y}$ of each vector. Liu and Tan use this rank-one approximation to find the error bound in Eq. 7 (see Fig. 5).

If $\mathbf{V}$ contains missing values, we can use the optimal recovery estimator to impute $\mathbf{V}$. Assuming the columns in $\mathbf{V}$ come from K circular cones defined as (4), there is a pair of factor matrices $\mathbf{W}^* \in \mathbb{R}^{F\times K}_{+}$, $\mathbf{H}^* \in \mathbb{R}^{K\times N}_{+}$, such that

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{V}\|_F} \leq \max_{k\in[K]}\{\sin\alpha_k\}. \tag{7}$$

Since the error is bounded by $\sin\alpha_k$, we choose our optimal recovery estimator to minimize α_k. Since $\cos^2(\alpha)$ is monotonically decreasing for $\alpha \in (0, \pi/2)$, this is equivalent to maximizing the left side of the inequality in (5):

$$\hat{v}^*_{po} = \arg\max_{\hat{v}_{po}\in C_k}\{(\hat{v}_{po} \cdot u_k)^2 - (\hat{v}_{po} \cdot \hat{v}_{po})\cos^2(\alpha_k)\}. \tag{8}$$

We can solve (8) analytically using the Lagrangian with known values of v_{po} as equality constraints. We can also solve (8) numerically using projected gradient descent.

Algorithm 1: Greedy Clustering with Missing Values.

```
Data: Data matrix
      $\mathbf{V} \in \mathbb{R}_+^{F\times N},\ K \in \mathbb{N},\ \Omega \in \{0,1\}^{F\times N}$
Result: Cone indices $J \in \{0,1,...,K\}^N$;
        $\alpha \in (0,\pi/2)^K$; $u \in \mathbb{R}_+^{F\times K}$
Partition columns in $\mathbf{V}$ into subsets $\mathbf{V}_{fo}$ and $\mathbf{V}_{po}$,
  where $\mathbf{V}_{fo}$ contains data columns for which
  $\sum_i r_{ij} = F$, and $\mathbf{V}_{po}$ contains remaining columns.;
Normalize $\mathbf{V}_{fo}$ so that all columns have unit $\ell_2$-norm.
  Let $\mathbf{V}'_{fo}$ be the normalized matrix ;
Cluster items in $\mathbf{V}'_{fo}$ using greedy clustering [8,
  Alg. 1] to obtain cluster indices $J$ and run Alg. 3 on
  $\mathbf{V}'_{fo}$ to get $u_1, ..., u_k$ from $W^*$. ;
for $v_{po} \in \mathbf{V}_{po}$ do
    Let $\Omega_j$ correspond to observed entries of $\mathbf{v}_{po}$. Find
    $k = \arg\max_{j\in[K]} \cos^{-1}\left(\frac{P_\Omega(u_j)\cdot P_\Omega(v_{po})}{\|P_\Omega(u_j)\|\|P_\Omega(v_{po})\|}\right)$. If
    this condition is maximized by more than one $k$,
    choose one at random. Add the index of $v_{po}$ to
    $J_k$. ;
end
for $k \in [K]$ do
    $\alpha_k = \max_{v_{po}} \cos^{-1}\left(\frac{P_\Omega(v_{po})\cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\|\|P_\Omega(u_k)\|}\right)$ ;
end
Return cone indices $J$, $u$, $\alpha$ ;
```

Algorithm 2: Rank-1 NMF with Missing Values.

```
Data: Partially observed data
      $\mathbf{V} \in \mathbb{R}_+^{F\times N},\ \Omega \in \{0,1\}^{F\times N},\ K \in \mathbb{N}$
Result: $\hat{\mathbf{W}}^* \in \mathbb{R}_+^{F\times K}$ and $\hat{\mathbf{H}}^* \in \mathbb{R}_+^{K\times N}$
Cluster data using Alg. 1 ;
Impute data using (3) ;
Perform rank-1 NMF on imputed data using an
  SVD-based algorithm [8, Alg. 2] ;
```

Generally, u_k is not known beforehand, but we can find u_k given W_k. Given an ellipse in $\mathbb{R}^3$, we reconstruct its cone by drawing lines from its limit points to the origin. Then it is straightforward to find the center of the cone. (Note that while this volume minimization problem is NP-hard, there are efficient and accurate algorithms when certain assumptions are met, which have been used with NMF [32].) Liu and Tan propose the following optimization problem (in the absence of missing values) over the optimal size angle and basis vector for each cluster. We write the data points in each cluster as $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_M] \in \mathbb{R}_+^{F\times M}$ where $M \in \mathbb{N}_+$:

$$\begin{aligned} &\text{minimize}_{(0,\pi/2)} && \alpha \\ &\text{subject to} && \mathbf{x}_m^T\mathbf{u} \geq \cos\alpha,\ m \in [M], \\ & && \mathbf{u} \geq 0,\ \|\mathbf{u}\|_2 = 1,\ \alpha \geq 0. \end{aligned} \tag{9}$$

Of course, we also do not know C_k or W_k, so we use a clustering algorithm to find the vectors belonging to each C_k (see Sec. IV).

IV. ALGORITHM AND ERROR BOUND

Now we consider clustering and NMF with missing values. If the geometric assumption (6) holds, a greedy clustering algorithm [8, Alg. 1] returns the correct clustering of fully observed data. Here we show that a greedy algorithm also guarantees correct clustering of partially observed data under certain conditions.

Lemma 1 (Greedy clustering with missing values): Let Ω indicate the missing values of v_{po}. Let α_k be the defining angle of C_k and $P_\Omega(\alpha_k)$ be the defining angle of the cone resulting from projecting C_k onto the missing value plane from Ω. If, for exactly one k,

$$\arccos\left(\frac{P_\Omega(v_{po})\cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\|\,\|P_\Omega(u_k)\|}\right) \leq P_\Omega(\alpha_k) \tag{10}$$

then v_{po} originated from the corresponding C_k. If α_k are identical for all k, Alg. 1 will cluster v_{po} correctly.

Proof: The result follows directly. ■

Now consider feasibility of imputing data points using the $\hat{\alpha}$ and $\hat{u}$ from Alg. 2. Clearly, the missing values plane for each point intersects the original corresponding cone defined by the true u and α of the cone. We know the $\hat{u}$ fall somewhere within the original cones, but if the $\hat{\alpha}$ are too small, the new cones may not intersect with the missing values plane.

Lemma 2 (Feasibility of imputation algorithm): The estimator in (3) is able to find an imputation within the feasible set given $\alpha_1, \ldots, \alpha_K$ and $u_1, \ldots, u_k$ returned by Alg. 1.

Proof: Let vector v_{po} be a partially observed version of $v_{fo} \in \mathbf{V}$. We define the angle between v_{po} and cluster center u_k in the F-dimensional space:

$$\gamma_k = \arccos\left(\frac{P_\Omega(v_{po})\cdot u_k}{\|P_\Omega(v_{po})\|\,\|u_k\|}\right), \tag{11}$$

and between v_{po} and the projected cluster center in the projected $(F - f)$-dimensional space:

$$\hat{\gamma}_k = \arccos\left(\frac{P_\Omega(v_{po})\cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\|\,\|P_\Omega(u_k)\|}\right), \tag{12}$$

where Ω is the observed values indicator corresponding to v_{po}. Then $\gamma_k \leq \hat{\gamma}_k$ since $P_\Omega(v_{po})\cdot u_k = P_\Omega(v_{po})\cdot P_\Omega(u_k)$ and $\|u_k\| \geq \|P_\Omega(u_k)\|$. Thus $\hat{\gamma}_k$ is large enough that an imputation on the missing values plane is feasible for each v_{po}. Since $\alpha_k = \max \gamma_k$, all partially observed points labeled as belonging to C_k can be imputed. ■

We extend bound (7) on the relative NMF error to missing values (Alg. 3). Note that the original bound allows for overlapping cones and does not assume (6) holds. It only requires all points be within α_k of u_k, which means vectors in the normalized perturbation matrix $\mathbf{E}_k$ (illustrated as $\mathbf{y}$ in Fig. 3) are upper-bounded by $\sin\alpha_k$. In other words, If the missing entries of each $\mathbf{v}_{po}$ are imputed using Alg. 1, then the perturbation from the original u_k, which we denote $\hat{\mathbf{E}}_k$, will be at most $2\mathbf{E}_k$. We can prove this using a worst-case scenario.

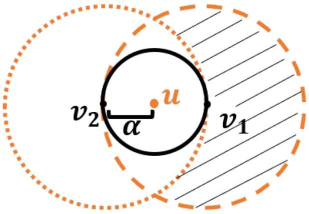

FIGURE 6. Geometric proof of relative NMF error bound.

Theorem 1 (Rank-1 NMF with missing values): Suppose $\mathbf{V}$ is drawn from K cones and missing values are introduced to get $\mathbf{V}_{po}$. If Alg. 2 correctly clusters data points and Alg. 1 is used to perform imputation, then

$$\frac{\|\mathbf{V}-\mathbf{W}_{po}^{*}\mathbf{H}_{po}^{*}\|_F}{\|\mathbf{V}\|_F} \le \max_{k\in[K]}\{\sin 2\alpha_k\}, \tag{13}$$

where $\mathbf{W}_{po}^{*}$ and $\mathbf{H}_{po}^{*}$ are found by Alg. 3.

Proof: Suppose there are two points v_1 and v_2 in a cone, as indicated by the solid circle in Fig. 6. Then u will be at an angle α from both v_1 and v_2. Now suppose v_2 contains missing values. Then the new v_1 will be the only vector in the cone, $\hat{v}_2$ is imputed using (8), where $\hat{u} = v_1$, and $\hat{v}_2$ is at an angular distance $\sin 2\alpha$ from $\hat{u}$. (One can check that if there are more than two points in the cone, this distance cannot increase.) A worst-case imputation places $\hat{v}_2$ at an angle 2α away from v_1 (suppose the optimizer places $\hat{v}_2$ at an angle greater than 2α from v_1, but this is a contradiction since then v_2 would be a better estimate than the optimum). The dashed circle in Fig. 6 represents points at an angle 2α from v_1. Any $\hat{v}_2$ outside the dotted circle is at an angle greater than 2α from v_2. So the shaded region indicates when the error may be greater than $\sin 2\alpha$. But the missing values of v_2 allow for "movement" only along the axes. Since the intersection of a hyperplane with a cone is a finite-dimensional ellipsoid [40], [41], which is compact [42], v_2 cannot "travel" via imputation to the shaded region without crossing a feasible region less than 2α from $\hat{u}$. Hence the theorem holds. ■

V. PROBABILISTIC ERROR

We now make some probabilistic assumptions on our data and missingness patterns to calculate the expected maximum error of optimal recovery imputation. First, consider a cone C in an F-dimensional space defined by u and α. Let us ignore the length of the vectors in C and preserve only the angles of the vectors from u. We can then represent vectors of an F-dimensional cone as points in an $(F-1)$-dimensional ball. For example, a 3-dimensional cone can be represented as points in a circle, as in Fig. 7.

Let there be N points $\{x_1, \ldots, x_N\} \in \mathbb{R}^F$, drawn uniformly at random from K F-dimensional balls, labeled $B_1, \ldots, B_K$. Let $d(x_i, x_j)$ be the Euclidean distance between x_i and x_j. We assume there is at least one data point in each ball, and that the distance between any two points in a ball B_k is less than the distance between any point in B_k and a point not in B_k. That is, for any $i, j \in [N], i \neq j$,

$$\max_{i,j\in B_k} d(x_i, x_j) < \min_{i\in B_k, j\notin B_k} d(x_i, x_j) \qquad \text{for all } k = 1, \ldots, K. \tag{14}$$

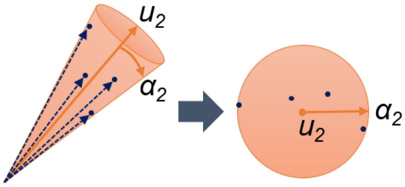

FIGURE 7. Geometric proof of relative NMF error bound.

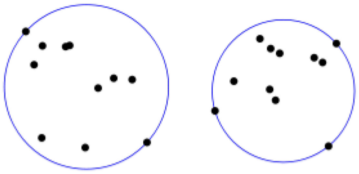

FIGURE 8. Minimum covering sphere in two dimensions.

This is equivalent to the geometric assumption in (6), and we can correctly cluster any points drawn from such balls using Alg. 1. After obtaining the clusters, we can compute the minimum covering sphere (MCS) on the points in each cluster [43] (Fig. 8). This gives us K balls with N_k points in each ball.

Now suppose that we have partially observed entries in our data. Let the missingness of a point be a Bernoulli random variable with parameter γ. That is, x is fully observed with probability γ and partially observed with probability $1-\gamma$. There is now some uncertainty about the position of partially observed data points, so we will find the MCS for only the fully observed points. This is analogous to step 3 in Algorithm 2. By calculating the expected change in the radius of the MCS, we can calculate the expected change in its corresponding cone.

Theorem 2 (Probabilistic bound on NMF error): Given the setting described above, and assuming that the N points are drawn uniformly at random from the K balls, then after imputing with Alg. 1, we can tighten the bound in (13) to

$$\mathbb{E}\left[\frac{\|\mathbf{V}-\mathbf{W}_{po}^{*}\mathbf{H}_{po}^{*}\|_F}{\|\mathbf{V}\|_F}\right] \le \max_{k\in[K]}\{\sin \alpha_k\}. \tag{15}$$

Proof: If the N points are drawn uniformly at random from the K balls, then $\mathbb{E}[N_k] = N/k$, and the expected number of fully observed and partially observed points in each cluster is

$$\mathbb{E}[|X_{k,fo}|] = \gamma N_k \qquad \text{and} \qquad \mathbb{E}[|X_{k,po}|] = (1-\gamma)N_k. \tag{16}$$

Clearly, the volume of the MCS can only decrease as $|X_{k,fo}|$ decreases. Let R_{max} be the radius of MCS if there were no missing values, and let $\hat{R}$ be the radius of the MCS of only the fully observed points. Then $\hat{R} < R_{max}$ only if any $x \in X_{po}$ originally lay on the surface of $\text{MCS}_{k,fo}$. Suppose the points are randomly distributed along the radius of the F-ball and we

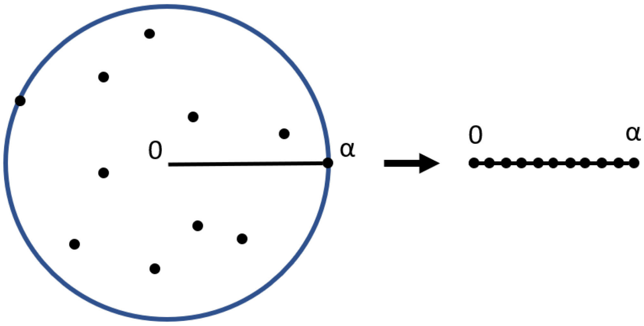

FIGURE 9. Assumption that points are uniformly random on the radius.

pick points to be partially observed uniformly at random. Let

$$N_{po} = \lceil (1-\gamma)N \rceil. \tag{17}$$

Assume x_i are i.i.d. and uniformly distributed (without loss of generality) on [0,1]. This matches the assumption in the probabilistic analysis in [8] that the angles are drawn uniformly at random on $[0, \alpha]$ (see Fig. 9). Assuming a continuous distribution, almost surely no two points have exactly the same radius, and the probability of picking the ℓ outermost points is

$$\mathbb{P}(\ell) = \frac{\binom{N-\ell}{N_{po}-\ell}}{\binom{N}{N_{po}}}, \text{ where } \ell = 0, 1, \ldots, N_{po}. \tag{18}$$

This gives us

$$\mathbb{E}[\ell] = \sum_{\ell=1}^{N_{po}} \ell \cdot \mathbb{P}[\ell] \tag{19}$$

$$= \sum_{\ell=1}^{N_{po}} \ell \cdot \frac{\binom{N-\ell}{N_{po}-\ell}}{\binom{N}{N_{po}}} \tag{20}$$

$$= \frac{1}{\binom{N}{N_{po}}} \sum_{\ell=1}^{N_{po}} \ell \cdot \binom{N-\ell}{N_{po}-\ell} \tag{21}$$

$$= \frac{\binom{N-1}{N_{po}-1} N(N+1)}{\binom{N}{N_{po}}(N-N_{po}+1)(N-N_{po}+2)}, \tag{22}$$

where N_{po} is dependent on γ, as defined in (17).

The radius of the resulting MCS is dependent on the distribution of points along the radius. We can determine $\hat{R}$ using order statistics. If we assume uniform distribution between 0 and 1, and order the points $x_1, \ldots, x_n$ so that x_1 is closest to the center of the sphere and x_n is farthest, the radius of the nth point, R_n, is given by the beta distribution

$$R_n \sim B(n, 1), \tag{23}$$

and

$$\mathbb{E}[R_n] = \frac{n}{n+1}. \tag{24}$$

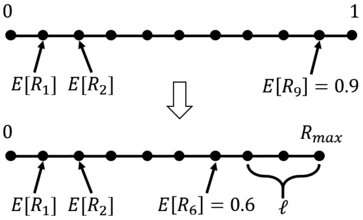

FIGURE 10. Example of $\mathbb{E}[\hat{R}]$ with $N = 9$ and $\ell = 3$.

Thus if ℓ of the outermost points are chosen to be missing,

$$\mathbb{E}[\hat{R}] = R_{max} - (\ell/N)R_{max} = \left(\frac{N-\ell}{N}\right) R_{max}. \tag{25}$$

We illustrate with an example in Fig. 10. We can substitute $\mathbb{E}[\ell]$ for ℓ, and since $\mathbb{E}[\ell]$ is a function of γ, we have derived the expected radius of the MCS as a function of missingness:

$$\mathbb{E}[\hat{R}] = \left(\frac{N-\mathbb{E}[\ell]}{N}\right) R_{max}. \tag{26}$$

Now we reverse the arrow in Fig. 9. Due to the random distribution of points in the sphere, removing the ℓ outermost points does not change the expected center u of the MCS. Transitioning from spheres back to cones, we get

$$\mathbb{E}[\hat{\alpha}] = \left(\frac{N-\mathbb{E}[\ell]}{N}\right) \alpha. \tag{27}$$

Thus

$$\alpha - \mathbb{E}[\hat{\alpha}] = \frac{\mathbb{E}[\ell]}{N} \cdot \alpha, \tag{28}$$

and the normalized Frobenius distance between $\mathbf{W}^*_{fo}\mathbf{H}^*_{fo}$ and $\mathbf{W}^*\mathbf{H}^*$ for a single cone is:

$$\mathbb{E}\left[\frac{\|\mathbf{W}^*_{fo}\mathbf{H}^*_{fo} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{W}^*\mathbf{H}^*\|_F}\right] \leq \sin\left(\frac{\mathbb{E}[\ell]}{N} \cdot \alpha\right). \tag{29}$$

If we assume $v_n \in \mathbf{V}$ are MCAR, the statistical mean of $\mathbf{V}_{fo}$ is the same as that of $\mathbf{V}$. Since v_n are uniformly distributed, the range of v_n remains centered on the mean, so the expected center of the MCS does not change. Thus the maximum difference between a point $v \in C_k$ and its imputed point $\hat{v}$ is $\sin \alpha_k$, and the theorem follows. ■

A. MCS WITH A DIFFERENT ASSUMPTION

If instead we assume points are uniformly distributed in the volume of the ball, we find the change in radius as follows. First, calculate the volume of a F-dimensional ball of radius $R = 1$:

$$V_F(R) = \frac{\pi^{F/2}}{\Gamma(F/2+1)} R^F. \tag{30}$$

Then we calculate radius $\hat{R}$ of an F-dimensional ball as:

$$\hat{R}_F(\hat{V}) = \frac{\Gamma(F/2+1)^{1/F}}{\sqrt{\pi}} \hat{V}^{1/F}, \tag{31}$$

where volume $\hat{V} = (\frac{1-\ell}{N})V_F(1)$.

The probability that a point x is in MCS_{po} is

$$\mathbb{P}(x \in \mathrm{MCS}_{po}) = \frac{V(\hat{R})}{V(R_{max})}. \tag{32}$$

Thus the expected radius given a missing parameter γ is given by

$$\mathbb{E}[\hat{R}] = \hat{R}_F \left(\frac{1 - \mathbb{E}[\ell]}{N} V_F(1) \right), \tag{33}$$

where $\mathbb{E}[\ell]$ is a function of γ, and the expected NMF error is

$$\mathbb{E}\left[\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{W}^*\mathbf{H}^*\|_F} \right] = \sin\left(\mathbb{E}[\hat{R}] \cdot \alpha\right). \tag{34}$$

B. MINIMUM COVERING SPHERICAL CAP FOR NORMALIZED DATA

If the data is normalized such that each vector has an L_2 norm of 1, all the points will fall on the surface of a sphere. Let there be N points $\{x_1, \ldots, x_N\} \in \mathbb{R}^F$, drawn at random from K F-dimensional spherical caps of a radius R F-ball, labeled $C_1, \ldots, C_K$. Let $d(x_i, x_j)$ be some distance between x_i and x_j. Assume there is at least one data point in each spherical cap, and that (6) holds.

The area of an F-dimensional spherical cap is

$$A(R, h) = \frac{1}{2} A_F R^{F-1} I_{2rh-h^2}/r^2 \left(\frac{F-1}{2}, \frac{1}{2} \right), \tag{35}$$

where $0 \leq h \leq R$, $A_n = 2\pi^{n/2}/\Gamma[n/2]$ is the area of the unit n-ball, h is the height of the cap, which can be calculated as a function of the angle α between the center and the edge of the cap, and $I_x(a, b)$ is the regularized incomplete beta function. Using the same style of analysis from the previous section, we can find the expected angle $\mathbb{E}[\alpha^{po}]$ given a parameter γ for partially observed points. Thus,

$$\mathbb{E}\left[\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{W}^*\mathbf{H}^*\|_F} \right] = \sin\left(\mathbb{E}[\alpha^{po}]\right). \tag{36}$$

VI. EXPERIMENTAL RESULTS

To test our algorithm, we first generate conical data satisfying the geometric assumption, using $N = 10\,000$, $F = 160$, and $K = 40$. We choose squared length of each v as a Poisson random variable with parameter 1, and we choose the angles of v uniformly. We then let $\mathbf{V}$ be partially-observed with Bernoulli parameter ξ to obtain $\mathbf{V}_{po}$. That is,

$$\Omega(i, j) \overset{i.i.d.}{\sim} Bern(\xi). \tag{37}$$

We run tests using $\xi \in \{0.4, 0.55, 0.7, 0.8, 0.9\}$ and find imputation relative error for NMF:

$$E[\mathbf{V}, \mathbf{W}_{po}^*\mathbf{H}_{po}^*] = \frac{\|\mathbf{V} - \mathbf{W}_{po}^*\mathbf{H}_{po}^*\|_F}{\|\mathbf{V}\|_F}. \tag{38}$$

Fig. 11 shows relative error of our optimal recovery imputation with different values of α when we enforce correct clustering. The error for all α values and missingness percentages lies within the bound given by (13). Note that because our data is drawn uniformly at random, the error does not approach the worst-case bound.

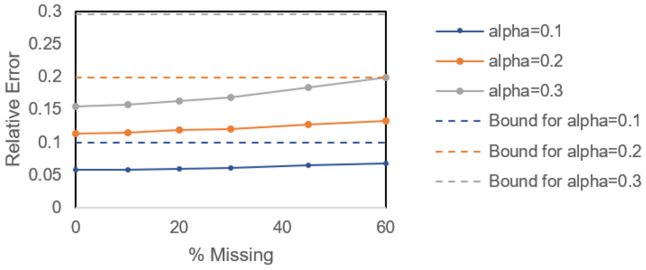

FIGURE 11. **Relative NMF error of imputed conical data with correct clustering.**

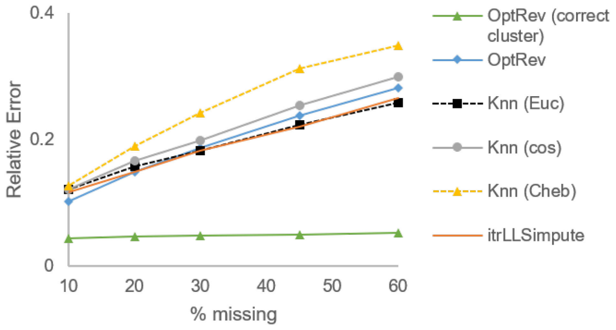

FIGURE 12. **Relative NMF error for Conical data.**

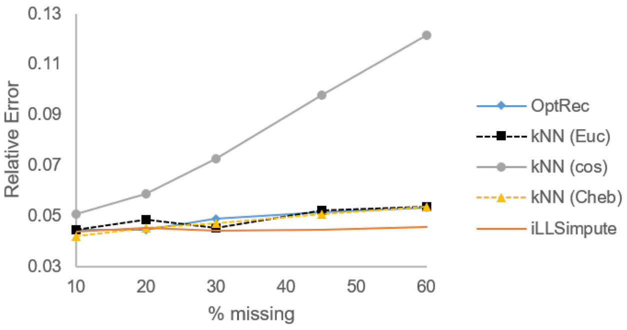

FIGURE 13. **Relative NMF error for Pavia data.**

In the next experiment, we impute the conical data with $\alpha = 0.1$ with other local imputation algorithms, including kNNimpute [44] with Euclidean, cosine, and Chebyshev (L_∞) distances and iterated local least squares (itrLLS) [45]. We perform two tests with optimal recovery: one with enforced correct clusterings and one without prior knowledge of the correct clusterings. We use $\alpha = 0.1$ and do not enforce correct clustering for Alg. 3 as before (see Fig. 12). We find $k = 8$ neighbors gives us the best results. Optimal recovery performs much better than other methods when clusters are known, and it performs similarly to other methods when they are not.

Following [8], the next experiment tests a subset of the hyperspectral imaging data set from Pavia [46]. We crop the 103 images to have 2000 pixels per image, set $K = 9$, corresponding to the different imagery categories, and introduce missing values in the same proportions as before (see Fig. 13).

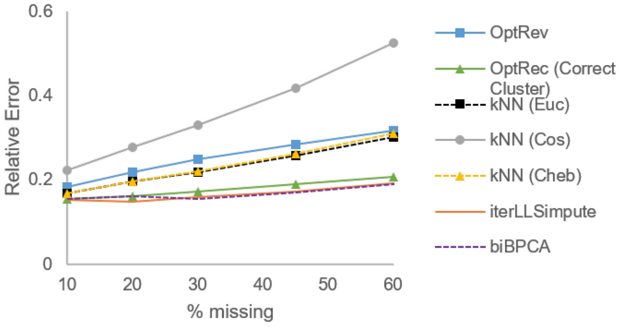

FIGURE 14. Relative NMF error for Mouse data.

TABLE I Average Imputation Times for Mouse Data in Seconds.

% Missing	10	20	30	45	60
OptRec	0.51	0.48	0.63	0.91	1.41
OptRec (Correct Clusters)	0.48	0.48	0.58	0.85	1.36
kNN (Euc)	0.11	0.17	0.29	0.34	0.51
kNN (Cos)	0.10	0.15	0.19	0.27	0.41
kNN (Cheb)	0.08	0.16	0.23	0.35	0.52
itrLLSimpute	43.9	30.4	25.7	20.5	14.5
biBPCA	5000+				

We also run tests with mice protein data [47] (see Fig. 14). The original dataset contains 1077 measurements with 77 proteins. We remove the 9 proteins that had missing measurements, then introduce missing values. We find $k = 5$ neighbors gives us the best results for kNNimpute on these datasets. On the mouse data, we also test bicluster BPCA [27] in addition to the other methods. The conical and Pavia test data were not sufficiently well-conditioned to run bicluster BPCA. See Tab. I for a comparison of run times. Our results demonstrate that optimal recovery performs similarly to kNN methods when clusters are not known beforehand. When clusters are known, optimal recovery performs similarly to more advanced methods (itrLLSimpute and biBPCA) in a fraction of the time.

VII. SUMMARY, DISCUSSION ON FAIRNESS, AND FUTURE WORK

We have extended classical approximation-theoretic *optimal recovery* to the setting of imputing missing values, specifically for NMF. We showed that imputation using optimal recovery minimizes relative NMF error under certain separability assumptions, and provided a straightforward algorithm for implementation. We gave a probabilistic error analysis of a clustering algorithm after minimax imputation. This analysis style can be extended to other clustering and imputation algorithms; various applications may require different model assumptions.

We now discuss the minimax approach and its implications on fairness. Missingness patterns themselves may carry information [48], and statistics-based imputation methods may introduce unfairness [49]. In certain social contexts, biases in algorithms can lead to unfair policy-making [50]. Researchers attempt to mitigate some of these biases using multiple imputation [24] or weighted estimators [51]. Philosopher John Rawls argues that in an effort to provide all individuals with equal opportunities, inequalities should only exist if they result in the worst off being better off [52]. In a scenario where one's place in society is chosen at random (including social status and other assets), one would prefer to land in a society that plays by a minimax rule, where the disadvantage of the worst off is minimized.

On the experimental side, we plan to test our imputation algorithm on single-cell RNA sequencing data along with different clustering algorithms, which will help us refine our algorithm for specific cases. We also aim to extend our algorithm to the scenario where complete cases for each cluster are not available. On the theoretical side, future work aims to study how minimax imputation impacts fairness in decision-making and clustering [53].

REFERENCES

[1] R. Chen and L. R. Varshney, "Non-negative matrix factorization of clustered data with missing values," in *Proc. IEEE Data Sci. Workshop*, Jun. 2019, pp. 180–184.

[2] J. Tuikkala *et al.*, "Missing value imputation improves clustering and interpretation of gene expression microarray data," *BMC Bioinf.*, vol. 9, no. 1, pp. 1–14, Apr. 2008.

[3] C.-C. Chiu *et al.*, "Missing value imputation for microarray data: A comprehensive comparison study and a web tool," *BMC Syst. Biol.*, vol. 7, no. 6, pp. 1–13, Dec. 2013.

[4] M. C. de Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–9, Dec. 2015.

[5] P.-L. Loh and M. J. Wainwright, "Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 2601–2605.

[6] Z. Charles, A. Jalali, and R. Willett, "Subspace clustering with missing and corrupted data," 2018, *arXiv:1707.02461*.

[7] V. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1602, Jul. 2013.

[8] Z. Liu and V. Y. F. Tan, "Rank-one NMF-based initialization for NMF and relative error bounds under a geometric assumption," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4717–4731, Sep. 2017.

[9] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2002.

[10] M. Golomb and H. F. Weinberger, "Optimal approximation and error bounds," in *On Numerical Approximation*, R. E. Langer Ed. Madison, WI, USA: Univ. Wisconsin Press, 1959, pp. 117–190.

[11] C. A. Micchelli and T. J. Rivlin, "A survey of optimal recovery," in *Optimal Estimation Approximation Theory*, C. A. Micchelli and T. J. Rivlin, Eds. New York, NY, USA: Plenum Press, 1976, pp. 1–54.

[12] C. A. Micchelli and T. J. Rivlin, "Lectures on optimal recovery," in *Numerical Analysis Lancaster 1984, Ser, Lecture Notes Mathematics*, P. R. Turner, Ed. Berlin, Germany: Springer-Verlag, 1985, pp. 21–93.

[13] R. G. Shenoy and T. W. Parks, "An optimal recovery approach to interpolation," *IEEE J. Signal Process.*, vol. 40, no. 8, pp. 1987–1996, Aug. 1992.

[14] D. L. Donoho, "Statistical estimation and optimal recovery," *Ann. Statist.*, vol. 22, no. 1, pp. 238–270, Mar. 1994.

[15] D. D. Muresan and T. W. Parks, "Adaptively quadratic (AQua) image interpolation," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 690–698, May 2004.

[16] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[17] K. Bhaskaran and L. Smeeth, "What is the difference between missing completely at random and missing at random?," *Int. J. Epidemiol.*, vol. 43, no. 4, pp. 1336–1339, 2014.

[18] M. Kolar and E. P. Xing, "Estimating sparse precision matrices from data with missing values," in *Proc. 29th Int. Conf. Mach. Learn.*, Jun. 2012, pp. 551–558.

[19] S. Oba *et al.*, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 6, pp. 2088–2096, Nov. 2003.

[20] K. Messer and L. Natarajan, "Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment," *Statist. Med.*, vol. 27, no. 30, pp. 6332–6350, Dec. 2008.

[21] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[22] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[23] S. van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, Dec. 2011.

[24] M. J. Azur *et al.*, "Multiple imputation by chained equations: What is it and how does it work?," *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, pp. 40–49, 2011.

[25] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," 1999.

[26] H. Kim, G. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: Local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, Jan. 2005.

[27] F. Meng, C. Cai, and H. Yan, "A bicluster-based Bayesian principal component analysis method for microarray missing value estimation," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 3, pp. 863–871, May 2014.

[28] A. C. O. Delalleau, A. Courville, and Y. Bengio, "Efficient EM training of Gaussian mixtures with missing data," 2018, *arXiv:1209.0521*.

[29] A. Robitzsch, S. Grund, and T. Henke, "Miceadds: Some additional multiple imputation functions, especially for mice," R package version 3.0-16. 2018. [Online]. Available: https://cran.r-project.org/web/packages/miceadds/index.html

[30] Q. Qi, Y. Zhao, M. Li, and R. Simon, "Non-negative matrix factorization of gene expression profiles: A plug-in for BRB-ArrayTools," *Bioinformatics*, vol. 25, no. 4, pp. 545–547, Feb. 2009.

[31] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source Code Biol. Med.*, vol. 8, no. 1, pp. 1–15, Sep. 2013.

[32] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, Mar. 2019.

[33] G. L. Stein-O'Brien *et al.*, "Enter the matrix: Factorization uncovers knowledge from omics," *Trends Genet.*, vol. 34, no. 10, pp. 790–805, Oct. 2018.

[34] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1141–1148.

[35] L. B. Thomas, "Rank factorization of nonnegative matrices," *SIAM Rev.*, vol. 16, no. 3, pp. 393–394, 1974.

[36] R. Clarke *et al.*, "The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data," *Nat. Rev. Cancer*, vol. 8, no. 1, pp. 37–49, Jan. 2008.

[37] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 18, pp. 788–791, Oct. 1999.

[38] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2017, pp. 988–992.

[39] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. 14th Int. Conf. Neur. Inf. Process. Syst.*, 2001, pp. 849–856.

[40] T. L. Heath, *Apollonius of Perga: Treatise on Conic Sections*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[41] M. S. Handlin, "Conic sections beyond $\mathbb{R}^2$," May 2013.

[42] Y. N. Kiselev, "Approximation of convex compact sets by ellipsoids. ellipsoids of best approximation," *Proc. Steklov Inst. Math.*, vol. 262, no. 1, pp. 96–120, Sep. 2008.

[43] T. H. Hopp and C. P. Reeve, "An algorithm for computing the minimum covering sphere in any dimension," Nat. Inst. Standards Technol., Gaithersburg, Maryland, NISTIR 5831, May 1996. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.5980&rep=rep1&type=pdf

[44] A. W.-C. Liew, N.-F. Law, and H. Yan, "Missing value imputation for gene expression data: Computational techniques to recover missing data from available information." *Brief Bioinf.*, vol. 12, no. 5, pp. 498–513, Sep. 2011.

[45] Z. Cai, M. Heydari, and G. Lin, "Iterated local least squares microarray missing value imputation," *J. Bioinform. Comput. Biol.*, vol. 4, no. 5, pp. 935–957, Oct. 2006.

[46] "Hyperspectral remote sensing scenes," Accessed: Oct. 29, 2019. [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Sensing_Scenes

[47] C. Higuera, K. Gardiner, and K. Cios, "Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome," *PLoS ONE*, vol. 10, no. 6, 2015, Art. no. e0129126.

[48] A. Ghorbani and J. Y. Zou, "Embedding for informative missingness: Deep learning with incomplete data," in *Proc. 56th Ann. Allerton Conf. Commun., Control, Comput.*, Oct. 2018, pp. 437–445.

[49] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo, "Fairness and missing values," 2019, *arXiv:1905.12728* .

[50] B. A. Williams, C. F. Brooks, and Y. Shmargad, "How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications," *J. Inf. Policy*, vol. 8, pp. 78–115, 2018.

[51] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proc. Conf. Fairness, Accountability, Transparency*, Atlanta, GA, USA, Jan. 2019, pp. 339–348.

[52] J. Rawls, *A Theory of Justice*. Belknap Press, 1971.

[53] F. Chierichetti *et al.*, "Fair clustering through fairlets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5029–5037.

REBECCA CHEN DEAN received the B.S. degree in biomedical engineering from the University of Texas, Austin, TX, USA, in 2015, and the M.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2019.

LAV R. VARSHNEY (Senior Member, IEEE) received the B.S. degree (*magna cum laude*) in electrical and computer engineering from Cornell University, Ithaca, New York, NY, USA, in 2004, and the S.M., E.E., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, 2008, and 2010, respectively.

He is currently an Associate Professor of electrical and computer engineering with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, with further appointments in computer science, neuroscience, industrial engineering, digital agriculture, and the Discovery Partners Institute. He is a Chief Scientist for AI in the IBM-ILLINOIS Center for Cognitive Computing Systems Research. He is also a Chief Scientist with Ensaras, Inc., Champaign, IL, USA, and the Founder of Kocree, Inc., Champaign, IL, USA. During 2010–2013, he was a Research Staff Member with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. During 2019–2020, he was a Principal Research Scientist of AI ethics and AI for social good with Salesforce Research, Palo Alto, CA, USA. His research interests include statistical signal processing, information theory, artificial intelligence, and creativity.

Dr. Varshney is a Member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He is currently on the Advisory Board of the AI XPRIZE. He was the recipient of the E. A. Guillemin Thesis Award and the J.-A. Kong Award Honorable Mention at Massachusetts Institute of Technology for his Ph.D.Thesis.