NON-NEGATIVE MATRIX FACTORIZATION OF CLUSTERED DATA WITH MISSING VALUES

Rebecca Chen and Lav R. Varshney

Coordinated Science Laboratory and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

ABSTRACT

We propose the approximation-theoretic technique of optimal recovery for imputing missing values in clustered data, specifically for non-negative matrix factorization (NMF), and develop an algorithm for implementation. Under certain geometric conditions, we prove tight upper bounds on NMF relative error, which is the first bound of this type for missing values. Experiments on image data and biological data show that this technique performs as well as or better than other imputation techniques that account for local structure.

Index Terms— imputation, missing values, non-negative matrix factorization, optimal recovery

1. INTRODUCTION

Matrix factorization is commonly used for clustering and dimensionality reduction in computational biology, imaging, and other fields. Non-negative matrix factorization (NMF) is particularly favored by biologists because non-negativity constraints preclude negative values that are difficult to interpret in biological processes [1, 2]. NMF of gene expression matrices can discover cell groups and lower-dimensional manifolds in gene counts for different cell types.

Often, data exhibits local structure, e.g., different groups of cells follow different gene expression patterns. Due to physical and biological limitations of DNA- and RNA-sequencing techniques, gene-expression matrices are usually incomplete, and matrix imputation is often necessary before further analysis [2]. The local structure can be used to improve imputation.

Imputation accuracy is commonly measured using root mean-squared error (RMSE) or similar error metrics. However, Tuikkala et al. argue that "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" [3]. Here, we specifically consider imputation performance in the context of NMF.

We introduce a new imputation method based on *optimal recovery*, an approximation-theoretic approach for estimating linear functionals of a signal [4–6] previously applied in signal and image interpolation [7–9], to perform matrix imputation of clustered data. Pushing optimal recovery to imputation requires new geometric analysis. Our contributions include:

- A computationally-efficient imputation algorithm that performs as well as or better than other modern imputation methods, as demonstrated on hyperspectral remote sensing data and biological data; and
- A tight upper bound on the relative error of downstream analysis by NMF. This is the first such error bound for settings with missing values.

This work was supported in part by Air Force STTR Grant FA8650- 16-M-1819 and in part by grant number 2018-182794 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

2. RELATED WORK

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 [10, 11], 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) [12–14]. 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. These methods take a statistical approach (averaging or linear regression) rather than the geometric approach we develop; we test out algorithm against these methods in our experiments.

After imputation, downstream analysis such as NMF can be performed on data. Donoho and Stodden interpret NMF as the problem of finding cones in the positive orthant which contain clouds of data points [15]. Liu and Tan show that a rank-one NMF gives a good description of near-separable data and provide an upper bound on the relative reconstruction error [16]. Given that gene and protein expression data is often linearly separable on some manifold- or high-dimensional space [17], the bound given by rank-one NMF is valid. We extend these ideas to data with missing values and, for the first time, bound performance of downstream analysis of imputation. Loh and Wainwright have previously bounded linear regression error of data with missing values [18], but they do not consider imputation, and their proof is based on modifying the covariance matrix when data is missing. Our proof is based on the geometry of NMF.

3. OPTIMAL RECOVERY

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 $\hat{v}$ 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.

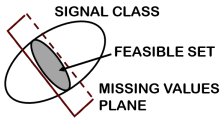

Fig. 1. Feasible set of estimators.

3.1. 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 observed values indicators with $\Omega_{ij} = 1$ if v_{ij} is observed and 0 otherwise. We make no assumptions on the missingness pattern, such as missing completely at random (MCAR) or missing at random (MAR) [11] 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. 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 $C_1, ..., C_k$. Let the *missing values plane* be the restriction set over $\mathbb{R}^F$ that satisfies the constraints on the observed values of v_{po}. We call this intersection the *feasible set* W:

$$W = \{\hat{v}_{po} \in C_k : P_\Omega(\hat{v}_{po}) = P_\Omega(v_{po})\} \text{ for some } k \in [K]. \quad (1)$$

Fig. 1 illustrates the feasible set of a three-dimensional vector with two missing samples when the signal class (convex space containing samples from kth cluster) covers an ellipsoid. If the the signal had only one missing sample, the feasible set would be a line segment.

All k for which (1) is satisfied are possible clusters from which the true v originated. Since W cannot be empty, there must be at least one C_k that has non-empty intersection with the set of all points satisfying the $P_\Omega(v_{po})$ constraint. The optimal recovery estimator $\hat{v}_{po}^*$ minimizes the maximum error over the feasible set of estimates:

$$\hat{v}_{po}^* = \arg\min_{\hat{v}_{po}\in C_k} \max_{v\in C_k} \|\hat{v}_{po} - v\|, \quad (2)$$

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

If W contains estimators belonging to more than one C_k, W can be partitioned into K disjoint sets, W_k, defined as

$$W_k = \{\hat{v}_{po} \in C_k : P_\Omega(\hat{v}_{po}) = P_\Omega(v_{po})\},\ k \in [K]. \quad (3)$$

Feasible clusters are those for which W_k is not empty, and we can find (2) over the C_k for which the corresponding W_k covers the largest volume: $k = \arg\max_k |W_k|$.

3.2. Application to non-negative matrix factorization

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 $\mathbf{V}$ [19]. We use the conical interpretation of NMF [15, 16], described as follows.

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 α:

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

or equivalently,

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

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,

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

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 [16, 20, 21]. 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 $\|\mathbf{y}\|_2 = \|\mathbf{z}\|_2(\sin\beta) \leq \|\mathbf{z}\|_2(\sin\alpha_k)$. We use this rank-one approximation to find error bounds [16].

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

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

Since the error is bounded by $\sin\alpha_k$, we choose our optimal recovery estimator to minimize α_k. This is equivalent to maximizing the inequality in (5):

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

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.

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 the its limit points to the origin. Then it is straightforward to find the center of the cone. 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 [16]. 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 $M \in \mathbb{N}_+$:

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

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. 4).

4. ALGORITHM AND ERROR BOUND

Now we considering clustering and NMF with missing values. If the geometric assumption (6) holds, a greedy clustering algorithm [16, 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.

Algorithm 1 (Greedy Clustering with Missing Values)
Input: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, $K \in \mathbb{N}$, $\Omega \in \{0,1\}^{F\times N}$
Output: Cone indices $J \in \{0,1,...,K\}^N$; $\alpha \in (0,\pi/2)^K$; $u \in \mathbb{R}_+^{F\times K}$

1. Partition columns in $\mathbf{V}$ into subsets $\mathbf{V}_{fo}$ and $\mathbf{V}_{po}$, where $\mathbf{V}_{fo}$ contains data columns for which $\sum_i \omega_{ij} = F$, and $\mathbf{V}_{po}$ contains remaining columns.
2. Normalize $\mathbf{V}_{fo}$ so that all columns have unit ℓ_2-norm. Let $\mathbf{V}'_{fo}$ be the normalized matrix
3. Cluster items in $\mathbf{V}'_{fo}$ using greedy clustering [16, Alg. 1] to obtain cluster indices J and run Alg. 2 on $\mathbf{V}'_{fo}$ to get $u_1,...,u_k$ from W^*.
4. **For** $v_{po} \in \mathbf{V}_{po}$
   Let Ω_j correspond to observed entries of $\mathbf{v}_{po}$. Find $k = \arg\max_{j\in[K]} \cos^{-1}\left(\frac{P_\Omega(\mathbf{z}_j)\cdot P_\Omega(\mathbf{v})}{\|P_\Omega(\mathbf{z}_j)\|\|P_\Omega(\mathbf{v})\|}\right)$. If this condition is maximized by more than one k, choose one at random. Add the index of v_{po} to J_k.
5. **For each** $k \in [K]$
   $\alpha_k = \max_{v_{po}} \cos^{-1}\left(\frac{P_\Omega(v_{po})\cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\|\|P_\Omega(u_k)\|}\right)$
6. Return cone indices J, u, α

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,*

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

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* (2) *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:

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

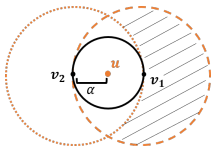

Fig. 2. Geometric proof of relative NMF error bound.

where Ω is the observed values indicator corresponding to v_{po}. Then $\gamma_k \le \hat{\gamma}_k$ since $P_\Omega(v_{po})\cdot u_k = P_\Omega(v_{po})\cdot P_\Omega(u_k)$ and $\|u_k\| \ge \|P_\Omega(u_k)\|$. 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. □

Algorithm 2 (Rank-1 NMF with Missing Values)
Input: Partially observed data $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, $\Omega \in \{0,1\}^{F\times N}$, $K \in \mathbb{N}$
Output: $\hat{\mathbf{W}}^* \in \mathbb{R}_+^{F\times K}$ and $\hat{\mathbf{H}}^* \in \mathbb{R}_+^{K\times N}$

1. Cluster data using Alg. 1
2. Impute data using (2)
3. Perform rank-1 NMF on imputed data using [16, Alg. 2]

We extend bound (7) on the relative NMF error to missing values (Alg. 2). 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 essentially allows the normalized perturbation matrix $\mathbf{E}_k$ to be upper-bounded by $\sin\alpha_k$. 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.

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. 1 correctly clusters data points and (2) 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. 2.

Proof. Suppose there are two points v_1 and v_2 in a cone, as indicated by the solid circle in Fig. 2. 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. 2 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 [22, 23], which is compact [24], 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 and is tight. □

5. EXPERIMENTAL RESULTS

To test our algorithm, we first generate conical data satisfying the geometric assumption, using $N = 10000$, $F = 160$, 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,

$$\Omega(i,j) \overset{i.i.d.}{\sim} Bern(\xi). \tag{14}$$

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

$$E[\mathbf{V}, \mathbf{W}^*_{po}\mathbf{H}^*_{po}] = \frac{\|\mathbf{V} - \mathbf{W}^*_{po}\mathbf{H}^*_{po}\|_F}{\|\mathbf{V}\|_F}. \tag{15}$$

Fig. 3 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.

In the next experiment, we impute the conical data with $\alpha = 0.1$ with other local imputation algorithms, including kNNimpute [25] with Euclidean, cosine, and Chebychev (L_∞) distances and iterated local least squares (itrLLS) [26]. 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. 2 as before (See Fig. 4). We found $k = 8$ neighbors gave 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 [16], the next experiment tests a subset of the hyperspectral imaging data set from Pavia [27]. 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. 5). We also run tests with mice protein data [28] (see Fig. 6). The original dataset contains 1077 measurements with 77 proteins. We remove the 9 proteins that had missing measurements, then introduced missing values. We found $k = 5$ neighbors gave us the best results for kNNimpute on these datasets. On the mouse data, we also test bicluster BPCA [14] in addition to the other methods. The conical and Pavia test data were not sufficiently well-conditioned to run bicluster BPCA. See Tab. 1 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.

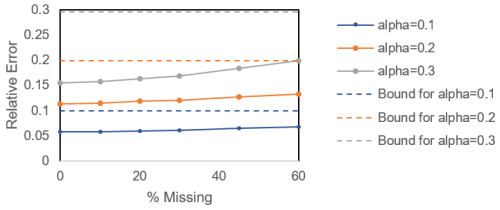

Fig. 3. Relative NMF error of imputed conical data with correct clustering.

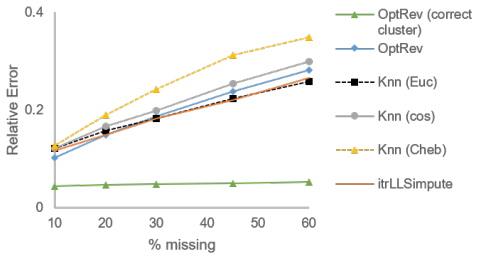

Fig. 4. Relative NMF error for Conical data

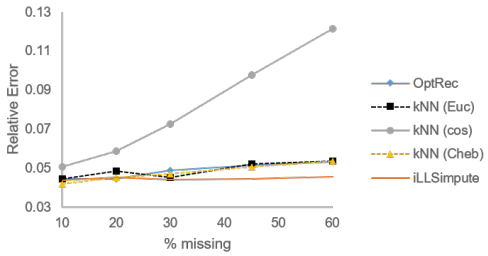

Fig. 5. Relative NMF error for Pavia data

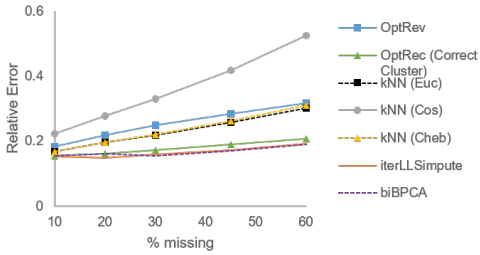

Fig. 6. Relative NMF error for Mouse data

Table 1. 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+				

6. CONCLUSION

We have extended classical approximation-theoretic *optimal recovery* for imputing missing values, specifically for NMF. We showed that imputation using optimal recovery minimizes relative NMF error under certain geometric assumptions. This required a novel reformulation of optimal recovery using the geometry of conic sections. Future work aims to extend optimal recovery to other settings of missing values in modern data science. We also aim to analyze the expected cone size and imputation error as a function of missingness fraction using a probabilistic approach, with minimum covering spheres as an analysis tool. On the experimental side, we plan to test our imputation algorithm on single-cell RNA sequencing data along with different clustering algorithms. We also aim to extend our algorithm to use local structure to take advantage of all observable data.

7. REFERENCES

[1] 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.

[2] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source Code for Biology and Medicine*, vol. 8, no. 10, Sep. 2013.

[3] J. Tuikkala *et al.*, "Missing value imputation improves clustering and interpretation of gene expression microarray data," *BMC Bioinformatics*, vol. 9, no. 202, Apr. 2008.

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

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

[6] ——, "Lectures on optimal recovery," in *Numerical Analysis Lancaster 1984*, ser. Lecture Notes in Mathematics, P. R. Turner, Ed. Berlin: Springer-Verlag, 1985, vol. 1129, pp. 21–93.

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

[8] D. L. Donoho, "Statistical estimation and optimal recovery," *The Annals of Statistics*, vol. 22, no. 1, pp. 238–270, Mar. 1994.

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

[10] S. van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, Dec. 2011.

[11] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley, 2002.

[12] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," *Technical Report, Division of Biostatistics, Stanford University.*, Oct. 1999.

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

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

[15] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. MIT Press, 2004, pp. 1141–1148.

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

[17] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, vol. 8, no. 1, pp. 37–49, Jan. 2008.

[18] P.-L. Loh and M. J. Wainwright, "Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression," in *Proc. 2012 IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2012.

[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[20] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *2017 IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017.

[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. 14th Int. Conf. Neur. Informat. Process. Syst. (NIPS)*. MIT Press, 2001, pp. 849–856.

[22] T. L. Heath, *Apollonius of Perga: Treatise on Conic Sections (Edited in Modern Notation)*. Cambridge University Press, 1986.

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

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

[25] 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." *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 498–513, Sep. 2011.

[26] 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.

[27] "Hyperspectral remote sensing scenes," http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes, accessed: 2019-10-29.

[28] 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: e0129126, 2015.