MISSING VALUES IMPUTATION AND IMAGE REGISTRATION FOR GENETICS APPLICATIONS

BY

REBECCA CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Assistant Professor Lav R. Varshney

ABSTRACT

In this thesis, we address several common scenarios of corrupted data in data and image processing pipelines. The first is in the setting of clustered data with missing values. We design an algorithm for imputing missing values using optimal recovery and derive an error bound for non-negative matrix factorization of the imputed data. Second, we consider missing values as erasure channels and show examples of using Fano's inequality to find lower bounds on missing values algorithms. Finally, we perform image registration of misaligned and noisy images using multiinformation and use finite rate of innovation sample to speed up registration while preserving optimality.

ACKNOWLEDGMENTS

I would like to thank my family, my husband, and my adviser for their support and guidance.

In addition, 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 the Silicon Valley Community Foundation.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Bioinformatics is an interdisciplinary field that has seen rapid growth in recent years due to the increased availability of biological data and advances in computer power. Modern technology has made it possible not only to store, but also to share large datasets. In addition, increases in computing power have made it possible to handle larger amounts of data than ever before. The goal is to extract useful information from the data that will allow us to understand biological systems.

The term "bioinformatics" was coined by Paulien Hogeweg and Ben Hesper in the 1970s to mean "the study of informatic processes in biotic systems" [1]. Hogeweg and Hesper proposed information storage, transmission, and processing as properties of living systems, and they considered this to be an important research area. This thinking was no doubt influenced by the development of the field of information theory in the 1950s and 60s, during which the Huffman code, the Reed-Solomon codes, and other landmark algorithms were developed. The idea that biological systems carried and transmitted information was reflected in the coinage of the "genetic code" [1]. From within the field of theoretical biology emerged mathematical models of enzyme dynamics and gene regulation. Alan Turing developed a theory for natural pattern formation (e.g. how a leopard gets its spots) [2]. Research in artificial intelligence led to the development of genetic algorithms for optimization and the now-ubiquitous neural networks for pattern recognition.

The goal of bioinformatics soon evolved into one of data analysis and interpretation. This was due to a massive increase in public data brought about by advances in DNA sequencing techniques, including a rapid DNA sequencing technique in 1977 and the polymerase chain reaction (PCR) technique for amplifying DNA in 1983. The U.S. Department of Energy (DOE) and the National Institutes of Health (NIH) set forth a plan to sequence the entire human genome, and in 1990, public funding for the Human Genome Project

(HGP) began. A parallel sequencing effort by Celera Genomics was formally launched in 1998. Celera developed a shotgun strategy that was able to sequence DNA more quickly and cost-effectively. This drove the HGP to improve their own technology, and both projects published their drafts of the completed human genome in 2000. The genome sequencing race spurred development not only in sequencing technology, but also in data mining, pattern recognition, and other analysis techniques.

Since then, technology has been developed for DNA methylation sequencing, mRNA sequencing, and protein sequencing, allowing scientists to study genetic processes in a more nuanced way [3]. While every cell belonging to an individual contains the same set of DNA, the mRNA differ from cell to cell depending on the cell's state and function (as do methylation and protein expression). In 2016, an initiative called the Human Cell Atlas Project began. The project aims to create a comprehensive reference map of all the cells in the human body, with the goal of understanding human health and, ultimately, diagnosing and treating disease. Single-cell RNA sequencing (scRNA-seq) has made it possible to capture gene expression patterns in individual cells. Beyond cell-level interactions, researchers hope to study tissue-level and eventually organ-level interactions, both spatially and functionally. Variations in gene expression patterns in healthy and diseased states can be compared, and temporal responses to drugs can be captured [4].

In this thesis we consider two facets of this genomics research: cell clustering and image-based transcriptomics. Clustering can be performed on gene-expression count matrices, which are matrices with cells on one axis and genes on the other axis. Matrix entries indicate the number of times a gene is expressed in a cell. Cells can then be clustered according to gene-expression patterns. Presumably, cells of different types or in different states will express genes differently. The second area we consider is image-based transcriptomics. Unlike gene counts, which only describe the abundance of genes in a cell, image-based transcriptomics captures spatial patterns of gene expression. In fluorescence in situ hybridization (FISH), genes are tagged by fluorescent markers, and images are taken in situ. Thus cellular microenvironments, as well as localization patterns, are preserved. Single molecule FISH (sm-FISH) allows for single molecule sensitivity.

In both cases, we run into the problem of noise. Gene expression counts may be inaccurate, and in the case of scRNA-seq, some genes may be missed

entirely. Low gene counts may be incorrectly recorded as zeros, but there is no easy way to determine whether a zero indicates the absence of a gene or a gene that was missed by the sequencing method. In chapter 2 we introduce patterns of missingness and describe methods of imputing, or filling in, missing values. We cluster and impute data with missing values using optimal recovery and find error bounds in the context of non-negative matrix factorization, which is a popular method for analyzing gene-expression matrices. In chapter 3 we discuss missing mechanisms as erasure channels and show an example of Fano's inequality in a missing value setting.

In image-based transcriptomics, images must be aligned, or registered, so that the spatial configuration of the cell on the pixel grid matches across consecutive images. In chapter 4, we consider misaligned images as noisy copies of the original image and multivariate information functionals to perform theoretically optimal registration. We then use finite-rate-of-innovation sampling to extract salient features, which speeds up registration while preserving optimality properties. Chapter 5 summarizes our findings and outlines future work.

The basic approaches we develop are useful not just in bioinformatics but in other parts of data science and image processing also. Missing data is common in real-world settings, and image registration is used in hyperspectral remote sensing and many computer vision applications. Our methods can be readily applied to other applications.

Bibliographical Note

Part of Chapter 2 appears in

R. Chen, L. R. Varshney, "Non-negative matrix factorization of clustered data with missing values," *Proc. IEEE 2019 Data Sci. Workshop*, 2019.

Part of Chapter 4 appears in

R. Chen, A. B. Das, and L. R. Varshney, "Registration for image-based transcriptomics: Parametric signal features and multivariate information measures," *Proc. 53rd Conf. Inform. Sci. Syst.*, 2019.

CHAPTER 2

NON-NEGATIVE MATRIX FACTORIZATION OF CLUSTERED DATA WITH MISSING VALUES

2.1 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 [5, 6]. NMF of gene expression matrices can discover cell groups and lower-dimensional manifolds in gene counts for different cell types. According to Stein-O'Brien et al., "newer MF algorithms that model missing data are essential for [single-cell RNA sequence] data" [7].

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 [6]. 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" [8]. 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 [9, 10, 11] previously applied in signal and image interpolation [12, 13, 14], to perform matrix imputation of clustered data. Analysis with missing data has been performed in the settings of high-dimensional regression [15] and subspace clustering [16]. 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.

2.2 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) [17]. When data is MCAR, the missing data is a random subset of all data, and the missing and observed values have similar distributions [18]. The MCAR condition is described in (2.1). This may occur if a researcher forgets to collect certain information for certain subjects, or if certain data samples are collected only for a random subset of test subjects. When data is MAR, the distribution of missing data is dependent on the observed data (2.2). 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 (2.3). For example, people with very high incomes may be less likely to report their incomes.

$$\text{MCAR: } \mathbb{P}(Y\text{is missing}|X,Y) = \mathbb{P}(Y\text{is missing}) \tag{2.1}$$

$$\text{MAR: } \mathbb{P}(Y\text{is missing}|X,Y) = \mathbb{P}(Y\text{is missing}|X) \tag{2.2}$$

$$\text{MNAR: } \mathbb{P}(Y\text{is missing}|X,Y) = \mathbb{P}(Y\text{is missing}|Y) \tag{2.3}$$

It is important to understand the missingness mechanism when analyzing the data. When data is MCAR, the statistics 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. If data is MAR or MNAR, the complete cases may be a biased representation of the dataset. One can also estimate statistics such as means, variances, and covariances based on all

available non-missing observations of a variable. Then the sample size reduction may be less severe for certain variables. However, this also introduces bias when data is MAR or MNAR, and there is the additional problem of inconsistent sample sizes. 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.

Ding and Simonoff argue that missingness mechanisms are more nuanced than the three basic categories described by Rubin [19]. They claim that missingness is dependent on any combination of the missing values, the observed predictors, and the response variable (e.g. a category label). In the cases where the missingness pattern contains information about the response variable, the missingness is *informative* [20]. Ghorbani and Zou use informative missingness, using the missingness patterns themselves as an additional feature for data classification [21].

2.3 Missing values imputation

In many cases, it is advantageous to impute the missing data for 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. The variance in the multiple imputations of each missing observation reflects the uncertainty of the estimates, and *all* imputed datasets are used in the downstream analysis, which increases statistical power.

2.3.1 Single imputation

One of the simplest imputation techniques is *mean imputation.* Missing values of each variable are imputed with the mean value of that variable. Since all missing observations of a variable are imputed with the same value, variance is reduced, and other statistics may be skewed in the MAR and MNAR cases. The reduced variability in the imputed variable also decreases correlation with other variables [22].

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. This method also underestimates variances, but it overestimates correlations. *Stochastic regression* attempts to add the variance back by distributing imputed points above and below the regression line using a normal distribution.

Bayesian imputation approaches also exist, including *Bayesian PCA* [23] and *maximum likelihood imputation* [24]. 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 (MCMC) method, which may be prohibitively time-consuming for large datasets.

2.3.2 Multiple imputation

Multiple imputation attempts to preserve the variance/covariance matrix of the data. Multiple imputation generates several imputations of the set, resulting in multiple complete datasets. Imputed datasets are then analyzed and results are pooled. The different imputations introduce variance into the data, but the variance may still be an underestimate since the imputations assume correlation between the variables. One of the more popular algorithms for multiple imputation is *multiple imputation by chained equations* (MICE) [25]. The steps of MICE are as follows:

1. Perform single imputation (e.g. mean imputation) for each missing value as a temporary value.
2. Choose one variable, set all the temporary values for that variable back to missing and regress this variable on other variables (these variables are user-specified).
3. Impute the missing values of that variable based on the previous regression.
4. Repeat steps 2-3 for each variable with missing values.
5. Repeat steps 2-4 for a specified number of cycles (i.e. until convergence). This results in one complete dataset.
6. Repeat steps 1-5 to obtain multiple imputed datasets.

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 [26]. A stepwise regression can be performed so that the missing variable is regressed on the best predictors.

2.3.3 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 [25, 27], 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) [28, 29, 30]. 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) [31]. By itself, MICE does not address clusters, but cluster-specific (group-wise) regression can be performed [32].

Tuikkala et al.'s clustering results on cDNA microarray datasets showed that "even when there are marked differences in the measurement-level imputation accuracies across the datasets, these differences become negligible when the methods are evaluated in terms of how well they can reproduce the original gene clusters or their biological interpretations" [8]. They used the Average Distance Between Partition (ADBP) to calculate clustering error, and they showed that BPCA, LLS, and kNN gave similar clustering results. Chiu et al. found that LLS-like algorithms performed better than kNN-like algorithms in terms of downstream clustering accuracy (measured using cluster pair proportions) [33]. De Souto et al. evaluated whether the effect of different imputation methods on clustering and classification were statistically significant [34]. They removed all genes with more than 10% missing

values and compared classification using the corrected Rand index. They found that simple methods such as mean and median imputation performed as well as weighted kNN and BPCA.

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 [35]. 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 [36]. Given that gene and protein expression data is often linearly separable on some manifold- or high-dimensional space [37], 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 [15], 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.

2.4 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.

2.4.1 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 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) [27] because we take a

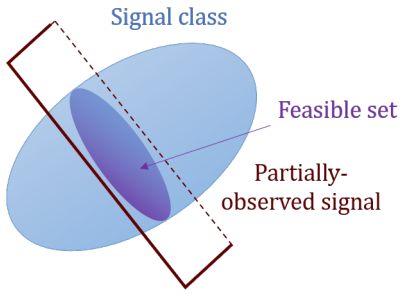

Figure 2.1: Feasible set of estimators.

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:

$$W = \{\hat{v}_{po} \in C_k : P_\Omega(\hat{v}_{po}) = P_\Omega(v_{po})\} \text{ for some } k \in [K]. \tag{2.4}$$

Fig. 2.1 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 signal had only one missing sample, the feasible set would be a line segment.

All k for which (2.4) 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:

$$\hat{v}_{po}^{*} = \arg\min_{\hat{v}_{po} \in C_k} \max_{v \in C_k} \|\hat{v}_{po} - v\|, \tag{2.5}$$

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

$$W_k = \{\hat{v}_{po} \in C_k : P_\Omega(\hat{v}_{po}) = P_\Omega(v_{po})\},\ k \in [K]. \tag{2.6}$$

Feasible clusters are those for which W_k is not empty, and we can find (2.5) over the C_k for which the corresponding W_k covers the largest volume: $k = \arg\max_k |W_k|$.

2.4.2 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 $\mathbf{V}$ [38]. We use the conical interpretation of NMF [35, 36], 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:

$$C(u, \alpha) := \left\{ x \in \mathbb{R}^F \backslash \{0\} : \frac{x \cdot u}{\|x\|_2} \geq \cos\alpha \right\}, \tag{2.7}$$

or equivalently,

$$C(u, \alpha) := \left\{ x \in \mathbb{R}^F \backslash \{0\} : (x \cdot u)^2 - (x \cdot x)\cos^2(\alpha) \geq 0 \right\}. \tag{2.8}$$

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

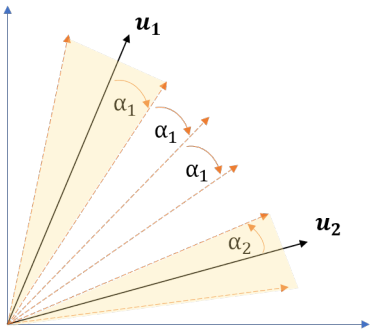

Figure 2.2: Geometric assumption for greedy clustering.

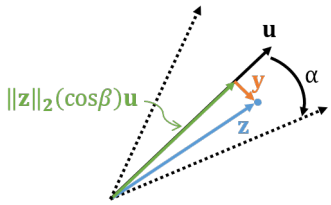

Figure 2.3: Decomposition of vectors in a circular cone.

well-separated cones, that is,

$$\min_{i,j\in[K],i\neq j} \beta_{ij} > \max_{i,j\in[K],i\neq j} \{\max\{\alpha_i + 3\alpha_j, 3\alpha_i + \alpha_j\}\}. \tag{2.9}$$

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 [36, 39, 40] (see Fig. 2.2). 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 [36] (see Fig. 2.3).

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 (2.7), 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

$$\frac{\|\mathbf{V} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{V}\|_F} \leq \max_{k\in[K]}\{\sin\alpha_k\}. \tag{2.10}$$

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

(2.8):

$$\hat{v}_{po}^{*} = \underset{\hat{v}_{po} \in C_k}{\arg\max} \{(\hat{v}_{po} \cdot u_k)^2 - (\hat{v}_{po} \cdot \hat{v}_{po}) \cos^2(\alpha_k)\}. \tag{2.11}$$

We can solve (2.11) analytically using the Lagrangian with known values of v_{po} as equality constraints. We can also solve (2.11) 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 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 [36]. 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 $M \in \mathbb{N}_{+}$:

$$\begin{aligned} &\text{minimize}_{(0,\pi/2)} && \alpha \\ &\text{subject to} && \mathbf{x}_m^T \mathbf{u} \geq \cos\alpha,\ m \in [M], \\ &&& \mathbf{u} \geq 0,\ \|\mathbf{u}\|_2 = 1,\ \alpha \geq 0. \end{aligned} \tag{2.12}$$

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. 2.5).

2.5 Algorithm and error bound

Now we considering clustering and NMF with missing values. If the geometric assumption (2.9) holds, a greedy clustering algorithm [36, 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,*

$$\arccos\left(\frac{P_\Omega(v_{po}) \cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\| \|P_\Omega(u_k)\|}\right) \leq P_\Omega(\alpha_k) \tag{2.13}$$

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 ℓ_2-norm. Let $\mathbf{V}'_{fo}$ be the normalized matrix ;

Cluster items in $\mathbf{V}'_{fo}$ using greedy clustering [36, 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 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. ;

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, α ;

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* (2.5) *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:

$$\gamma_k = \arccos\left(\frac{P_\Omega(v_{po}) \cdot u_k}{\|P_\Omega(v_{po})\|\|u_k\|}\right), \tag{2.14}$$

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

$$\hat{\gamma}_k = \arccos\left(\frac{P_\Omega(v_{po}) \cdot P_\Omega(u_k)}{\|P_\Omega(v_{po})\|\|P_\Omega(u_k)\|}\right), \tag{2.15}$$

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. □

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 $\hat{\mathbf{H}}^* \in \mathbb{R}_+^{K \times N}$

1 Cluster data using Alg. 1 ;
2 Impute data using (2.5) ;
3 Perform rank-1 NMF on imputed data using [36, Alg. 2] ;

We extend bound (2.10) on the relative NMF error to missing values (Alg. 3). Note that the original bound allows for overlapping cones and does not assume (2.9) 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. 2 correctly clusters data points and Alg. 1 is used to perform imputation, then*

$$\frac{\|\mathbf{V} - \mathbf{W}^*_{po}\mathbf{H}^*_{po}\|_F}{\|\mathbf{V}\|_F} \leq \max_{k\in[K]}\{\sin 2\alpha_k\}, \tag{2.16}$$

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. 2.4. 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 (2.11), 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. 2.4 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 [41, 42], which is compact [43], 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. □

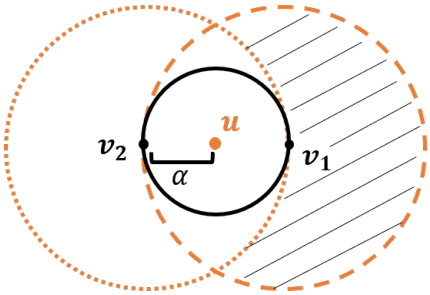

Figure 2.4: Geometric proof of relative NMF error bound.

2.6 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,

$$\Omega(i,j) \overset{i.i.d.}{\sim} Bern(\xi). \tag{2.17}$$

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

$$E[\mathbf{V}, \mathbf{W}^*_{po}\mathbf{H}^*_{po}] = \frac{\|\mathbf{V} - \mathbf{W}^*_{po}\mathbf{H}^*_{po}\|_F}{\|\mathbf{V}\|_F}. \tag{2.18}$$

Fig. 2.5 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 (2.16). 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 [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. 2.6). 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 [36], the next experiment tests a subset of the hyperspectral

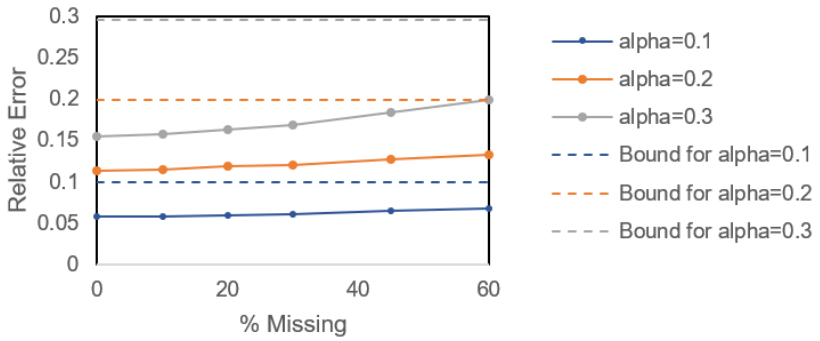

Figure 2.5: Relative NMF error of imputed conical data with correct clustering.

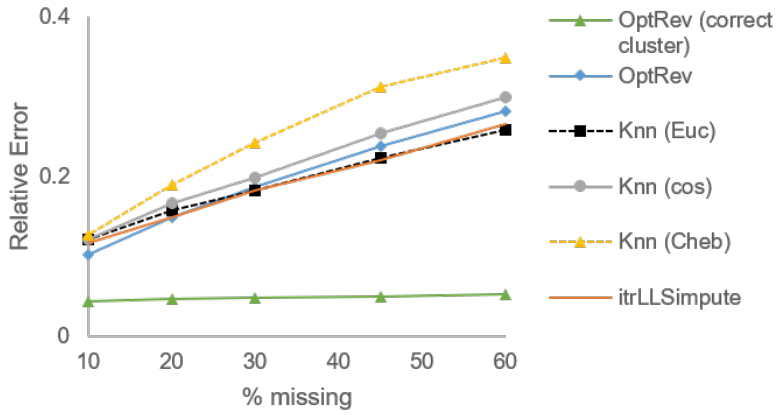

Figure 2.6: Relative NMF error for Conical data.

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. 2.7). We also run tests with mice protein data [47] (see Fig. 2.8). 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 [30] in addition to the other methods. The conical and Pavia test data were not sufficiently well-conditioned to run bicluster BPCA. See Tab. 2.1 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.

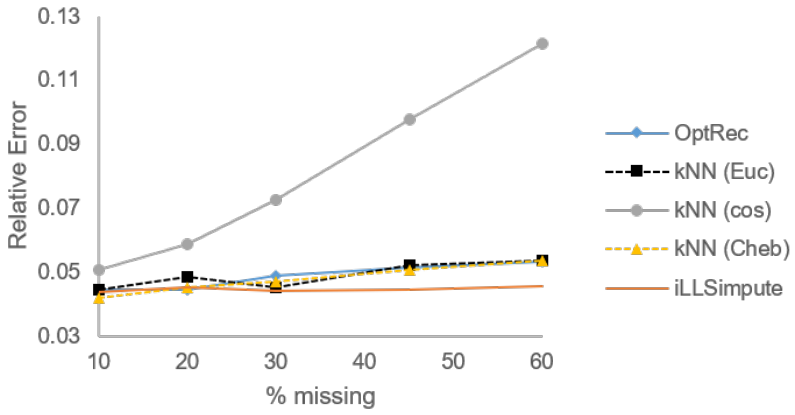

Figure 2.7: Relative NMF error for Pavia data.

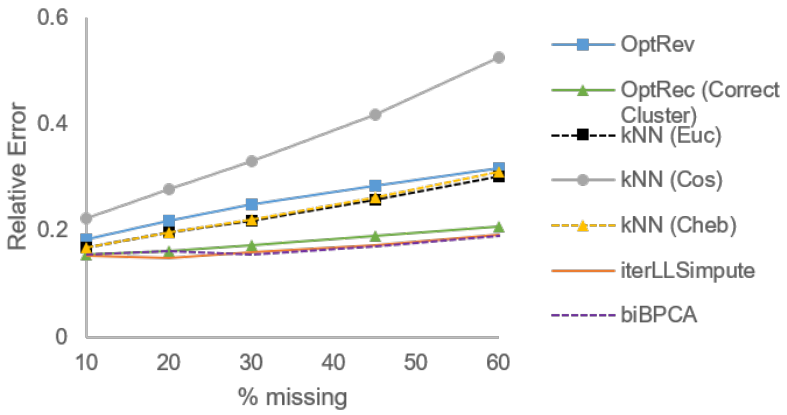

Figure 2.8: Relative NMF error for Mouse data.

Table 2.1: 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+				

Algorithm 3: SVD with missing values

Data: Data matrix $\mathbf{V}_{(0)} \in \mathbb{R}_+^F, K \in \mathbb{N}$
Result: $\mathbf{U} \in \mathbb{R}^{F \times K}, \mathbf{\Sigma} \in \mathbb{R}^{K \times K}, \mathbf{X} \in \mathbb{R}^{K \times N}$

Initialize $\mathbf{U}_{(0)}, \mathbf{\Sigma}_{(0)}, \mathbf{X}_{(0)}$;
for $t = 1, 2, ...$ *until convergence* **do**
  $\mathbf{M}_{(t)} = \mathbf{U}_{(t-1)}\mathbf{\Sigma}_{(t-1)}\mathbf{X}_{(t-1)}$;
  $\mathbf{V}_{(t)} = P_\Omega(\mathbf{V}_{(t-1)}) + P_{\Omega^C}(\mathbf{M}_{(t)})$;
  $\mathbf{U}_{(t)}, \mathbf{\Sigma}_{(t)}, \mathbf{X}_{(t)} = \texttt{svd}(\mathbf{V}_{(t)}, K)$;
end
Return $\mathbf{U}_{(t)}, \mathbf{\Sigma}_{(t)}, \mathbf{X}_{(t)}$;

Algorithm 4: Rank-1 NMF

Data: Data matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}, J \in [K]^N$
Result: $\mathbf{U} \in \mathbb{R}^{F \times K}, \mathbf{\Sigma} \in \mathbb{R}^{K \times K}, \mathbf{X} \in \mathbb{R}^{K \times N}$

for $k = 1$ *to* K **do**
  $\mathbf{V}_k := V(:, J_k)$;
  $[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{X}_k] := \texttt{svd}(\mathbf{V}_k)$;
  $\mathbf{w}_k^* := |\mathbf{U}_k(:, 1)|$, $\mathbf{h}_k := \mathbf{\Sigma}_k(1,1)|\mathbf{X}_k(:, 1)|$;
  $\mathbf{h}_k^* := \text{zeros}(1, N)$, $\mathbf{h}_k^*(J_k) := \mathbf{h}_k$;
end
$\mathbf{W}^* := [\mathbf{w}_1^*, ..., \mathbf{w}_K^*]$, $\mathbf{H}^* := [(\mathbf{h}_1^*)^T, ..., (\mathbf{h}_K^*)^T]$;
Return $\mathbf{W}^*, \mathbf{H}^*$;

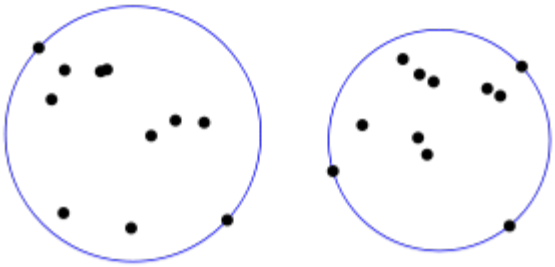

Figure 2.9: Minimum covering sphere in two dimensions.

2.7 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. 2.4.

Let there be N points $\{x_1, ..., x_N\} \in \mathbb{R}^F$, drawn uniformly at random from K F-dimensional balls, labeled $B_1, ..., B_K$. 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:

For any $i, j \in [N], i \neq j$,

$$\max_{i,j \in B_k} d(x_i, x_j) < \min_{i \in B_k, j \notin B_k} d(x_i, x_j) \qquad \text{for all } k = 1, ..., K. \tag{2.19}$$

This is equivalent to the geometric assumption in (2.9), and we can correctly cluster any points drawn from such balls using the greedy clustering algorithm already described. After obtaining the clusters, we can compute the minimum covering sphere (MCS) on the points in each cluster [48] (Fig. 2.9). 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 uncertainly 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 calculated the expected change in its corresponding cone.

If we assume the N points are drawn uniformly 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

$$\mathbb{E}[|X_{k,fo}|] = \gamma N_k \qquad \text{and} \qquad \mathbb{E}[|X_{k,po}|] = (1-\gamma)N_k. \tag{2.20}$$

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 pick points to be partially observed uniformly at random. Let

$$N_{po} = \lceil (1-\gamma)N \rceil. \tag{2.21}$$

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 [36] that the angles are drawn uniformly at random on $[0, \alpha]$ (see Fig. 2.10). Assuming a continuous distribution, almost surely no two points have exactly the same radius, and the probability of picking the ℓ outermost points is

$$\mathbb{P}(\ell) = \frac{\binom{N-\ell}{N_{po}-\ell}}{\binom{N}{N_{po}}}, \text{ where } \ell = 0, 1, ..., N_{po} \ . \tag{2.22}$$

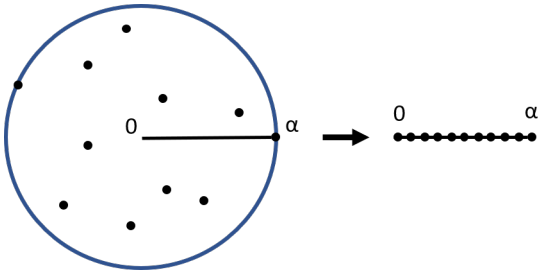

Figure 2.10: Assumption that points are uniformly random on the radius.

This gives us

$$\mathbb{E}[\ell] = \sum_{\ell=1}^{N_{po}} \ell \cdot \mathbb{P}[\ell] \tag{2.23}$$

$$= \sum_{\ell=1}^{N_{po}} \ell \cdot \frac{\binom{N-\ell}{N_{po}-\ell}}{\binom{N}{N_{po}}} \tag{2.24}$$

$$= \frac{1}{\binom{N}{N_{po}}} \sum_{\ell=1}^{N_{po}} \ell \cdot \binom{N-\ell}{N_{po}-\ell} \tag{2.25}$$

$$= \frac{\binom{N-1}{N_{po}-1} N(N+1)}{\binom{N}{N_{po}}(N-N_{po}+1)(N-N_{po}+2)}, \tag{2.26}$$

where N_{po} is dependent on γ, as defined in (2.21).

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 $x_1, ..., x_n$ 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

$$R_n \sim B(n, 1), \tag{2.27}$$

and

$$\mathbb{E}[R_n] = \frac{n}{n+1}. \tag{2.28}$$

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

$$\mathbb{E}[\hat{R}] = R_{max} - (\ell/N)R_{max} = \left(\frac{N-\ell}{N}\right) R_{max} \ . \tag{2.29}$$

We illustrate with an example in Fig. 2.11. We can substitute $\mathbb{E}[\ell]$ for ℓ, and

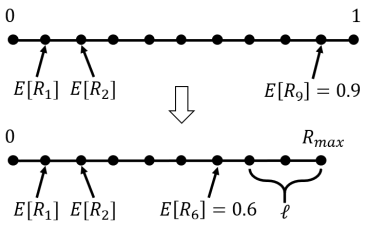

Figure 2.11: Example of $\mathbb{E}[\hat{R}]$ with $N = 9$ and $\ell = 3$.

since $\mathbb{E}[\ell]$ is a function of γ, we have derived the expected radius of the MCS as a function of missingness:

$$\mathbb{E}[\hat{R}] = \left(\frac{N - \mathbb{E}[\ell]}{N}\right) R_{max} \ . \tag{2.30}$$

Now we reverse the arrow in Fig. 2.10. 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

$$\mathbb{E}[\hat{\alpha}] = \left(\frac{N - \mathbb{E}[\ell]}{N}\right) \alpha \ . \tag{2.31}$$

Thus

$$\alpha - \mathbb{E}[\hat{\alpha}] = \frac{\mathbb{E}[\ell]}{N} \cdot \alpha, \tag{2.32}$$

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

$$\frac{\|\mathbf{W}^*_{fo}\mathbf{H}^*_{fo} - \mathbf{W}^*\mathbf{H}^*\|_F}{\|\mathbf{W}^*\mathbf{H}^*\|_F} \leq \sin\left(\frac{\mathbb{E}[\ell]}{N} \cdot \alpha\right). \tag{2.33}$$

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 $\sin \alpha_k$. Thus after imputing with Alg. 3, we can tighten the bound in (2.16) to

$$\frac{\|\mathbf{V} - \mathbf{W}^*_{po}\mathbf{H}^*_{po}\|_F}{\|\mathbf{V}\|_F} \leq \max_{k \in [K]} \{\sin \alpha_k\}. \tag{2.34}$$

2.7.1 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 $R = 1$:

$$V_F(R) = \frac{\pi^{F/2}}{\Gamma(F/2+1)} R^F. \tag{2.35}$$

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

$$\hat{R}_F(\hat{V}) = \frac{\Gamma(F/2+1)^{1/F}}{\sqrt{\pi}} \hat{V}^{1/F}, \tag{2.36}$$

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

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

$$\mathbb{P}(x \in \text{MCS}_{po}) = \frac{V(\hat{R})}{V(R_{max})}. \tag{2.37}$$

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

$$\mathbb{E}[\hat{R}] = \hat{R}_F \left(\frac{1 - \mathbb{E}[\ell]}{N} V_F(1) \right), \tag{2.38}$$

where $\mathbb{E}[\ell]$ is a function of γ.

2.7.2 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 $\{x_1, ..., x_N\} \in \mathbb{R}^F$, drawn at random from K F-dimensional spherical caps of a radius R F-ball, labeled $C_1, ..., C_K$. 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 Assumption 1 holds.

The area of an F-dimensional spherical cap is

$$A(R, h) = \frac{1}{2} A_F R^{F-1} I_{2rh-h^2}/r^2 \left(\frac{F-1}{2}, \frac{1}{2} \right), \tag{2.39}$$

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.

2.8 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. 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.

CHAPTER 3

MISSING VALUES AS NOISY CHANNELS

3.1 Introduction

We can consider the missingness mechanisms described in Chapter 2 as different types of channel noise. In the simplest case, MCAR mechanisms can be modeled as an erasure channel (Fig. 3.1). MAR and MNAR mechanisms can be modeled as signal-dependent noise channels. Ding and Simonoff describe X missingness, or missingness that is dependent on observed variables (equivalent to MAR), M missingness, or missingness dependent on missing variables (equivalent to MNAR), and Y missingness, or missingness dependent on an output (such as signal class or some function of the signal observations). Missingness can also be mixed (MX, MY, XY, XMY). Consider a case of MNAR missingness in gene testing where small gene counts are missed and recorded as zeros. This can be modeled as a channel that distorts or attenuates small signals with high probability and leaves large signals unchanged. One can define other similar signal-dependent or class-dependent erasure channels and derive channel capacities to obtain information theoretic bounds. We use an erasure channel to model MCAR mechanisms, leaving the other mechanisms as an area for future work.

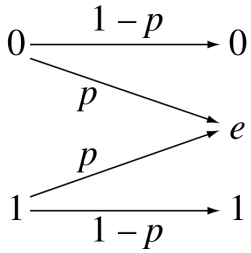

Figure 3.1: Binary erasure channel with error probability p.

3.2 Missing values as a binary erasure channel: Fano's inequality and multiple hypothesis testing

Consider a set of n samples $\mathbf{Y} = (Y_1, ..., Y_n)$ drawn from joint distribution $P_\theta^n(\mathbf{y})$ parametrized by θ, where θ lies in some set Θ. If inputs $\mathbf{X} = (X_1, ..., X_n)$ are present, samples are drawn from joint distribution $P_{\theta,\mathbf{X}}^n(\mathbf{y})$ parametrized by $(\theta, \mathbf{X})$. An algorithm forms an estimate $\hat{\theta}$ of θ, with the goal that some loss function $\ell(\theta, \hat{\theta})$ is small.

We set up a multiple hypothesis test. Let $V \in \{1, ..., M\}$ be an index corresponding to the parameters θ_V. Suppose V is drawn from a prior distribution P_V, and a sequence of samples $\mathbf{Y}$ is drawn from $P_{\mathbf{Y}|V}$. The goal is to identify V with high probability given $\mathbf{Y}$. If our estimation algorithm correctly outputs $\hat{\theta} \approx \theta_V$, then we should be able to recover the index V from $\hat{\theta}$. This is the problem of multiple hypothesis testing, where the vth hypothesis is that the underlying parameter is θ_v. If the hypothesis test cannot be successful, then the algorithm cannot perform well [49].

We use Fano's inequality to lower bound the error probability. For exact recovery, we define error probability as

$$P_e = \mathbb{P}[\hat{V} \neq V]. \tag{3.1}$$

Theorem 2 (Fano's inequality)**.** *For discrete random variables V and $\hat{V}$ on a common alphabet $\mathcal{V}$,*

$$H(V|\hat{V}) \leq H_2(P_e) + P_e \log(|\mathcal{V}| - 1), \tag{3.2}$$

where $H_2(\cdot)$ denotes the binary entropy. If V is uniform on $\{1, ..., M\}$, then

$$I(V; \hat{V}) \geq (1 - P_e) \log |\mathcal{V}| - \log 2, \tag{3.3}$$

or

$$P_e \geq 1 - \frac{I(V; \hat{V}) + \log 2}{\log |\mathcal{V}|}. \tag{3.4}$$

Proof. The proof of (3.2) can be found in [50]. To get (3.3), we use $|\mathcal{V}| - 1 \leq |\mathcal{V}|$ and $H_2(P_e) \leq \log 2$, and we subtract $H(V) = \log |\mathcal{V}|$ from both sides [49]. □

Since $V \to \mathbf{Y} \to \hat{V}$ is a Markov chain, we can use the data-processing

inequality to bound $I(V;\hat{V}) \leq I(V;\mathbf{Y})$. If each of the n observations can take on b values, $I(V;\mathbf{Y}) \leq H(\mathbf{Y}) \leq n \log b$, so

$$P_e \geq 1 - \frac{n \log b + \log 2}{\log M}. \tag{3.5}$$

Thus to achieve $P_e \leq \delta$, we need

$$n \geq \frac{(1-\delta)\log M - \log 2}{\log b} \tag{3.6}$$

samples. Suppose samples are missing completely at random with probability ξ. The problem can be modeled as an erasure channel, and $\max I(V;\hat{V})$ is scaled by a factor of $(1-\xi)$. Substituting $(1-\xi)I(V;\hat{V})$ for $I(V;\hat{V})$ in Fano's inequality, we now need

$$n \geq \frac{(1-\delta)\log M - \log 2}{(1-\xi)\log b} \tag{3.7}$$

samples to achieve $P_e \leq \delta$.

For approximate recovery, the error probability is defined as

$$P_e(t) = \mathbb{P}[d(V,\hat{V}) > t], \tag{3.8}$$

where $d(v,\hat{v})$ is some real-valued function and $t \in \mathbb{R}$. We define minimum and maximum neighborhood sizes

$$N_{max}(t) = \max_{\hat{v}\in\hat{\mathcal{V}}} N_{\hat{v}}(t) \qquad \text{and} \qquad N_{min}(t) = \min_{\hat{v}\in\hat{\mathcal{V}}} N_{\hat{v}}(t), \tag{3.9}$$

where $N_{\hat{v}}(t)$ is the number of $v \in \mathcal{V}$ for which $d(v,\hat{v}) \leq t$.

Theorem 3 (Fano's inequality with approximate recovery)**.** *For any random variables V and $\hat{V}$ on the finite alphabets $\mathcal{V}$ and $\hat{\mathcal{V}}$,*

$$H(V|\hat{V}) \leq H_2(P_e(t)) + P_e(t)\log\frac{|\mathcal{V}| - N_{min}(t)}{N_{max}(t)} + \log N_{max}(t). \tag{3.10}$$

If V is uniform on $\mathcal{V}$, then

$$I(V;\hat{V}) \geq (1-P_e(t))\log\frac{|\mathcal{V}|}{N_{max}(t)} - \log 2, \tag{3.11}$$

or equivalently,

$$P_e(t) \geq 1 - \frac{I(V;\hat{V}) + \log 2}{\log \frac{|\mathcal{V}|}{N_{max}(t)}}. \tag{3.12}$$

Proof. The proof can be found in [51]. □

Again, we can use the data-processing inequality to bound $I(V;\hat{V}) \leq I(V;\mathbf{Y})$. If each of the n observations can take on b regions of values, where each region $d(v,\hat{v}) \leq t$, then $I(V;\mathbf{Y}) \leq H(\mathbf{Y}) \leq n \log b$, so

$$P_e \geq 1 - \frac{n \log b + \log 2}{\log \frac{|\mathcal{V}|}{N_{max}(t)}}. \tag{3.13}$$

To achieve $P_e(t) \leq \delta$, we require

$$n \geq \frac{1}{\log b} \left((1-\delta) \log \frac{|\mathcal{V}|}{N_{max}(t)} - \log 2 \right). \tag{3.14}$$

To achieve $P_e(t) \leq \delta$ with samples MCAR with probability ξ, we need

$$n \geq \frac{1}{(1-\xi) \log b} \left((1-\delta) \log \frac{|\mathcal{V}|}{N_{max}(t)} - \log 2 \right). \tag{3.15}$$

3.3 Group testing with missing outcomes

We describe the group testing problem and directly apply (3.7) and (3.15) to existing bounds outlined in [49]. Suppose we have a set of p items. Of these items, k are defective. The set of defective items $S \subseteq \{1, ..., p\}$ is uniform over the $\binom{p}{k}$ possible subsets containing k items. For each test, a subset of the p items are polled, and the test produces a binary outcome indicating if the polled subset contains at least one defective item. The goal is to identify S with the fewest tests.

We denote the test matrix $\mathbf{X} \in \{0,1\}^{n \times p}$ where the $X_{i,j}$ is 1 if item j is included in test i. We let $\mathbf{X}$ be chosen in advance and randomly distributed (e.g., i.i.d. Bernoulli).

Scarlett and Cevher [49] consider passing the noiseless test outcome through a binary symmetric channel (BSC), which corresponds to incorrect test outcomes. In their model, the observed outcome is given by

$$Y_i = \left(\bigvee_{j \in S} X_{ij} \right) \oplus Z_i, \tag{3.16}$$

where $\bigvee_{j \in S} X_{ij}$ is the noiseless test outcome, $\oplus$ denotes modulo-2 addition, and $\vee$ is the "OR" operation. Let Z_i be i.i.d. Bernoulli(ϵ) for some $\epsilon \in [0, \frac{1}{2})$ and independent of $\mathbf{X}$. Let the vector of test outcomes $\mathbf{Y} = \{Y_1, ..., Y_n\}$. Given $\mathbf{X}$ and $\mathbf{Y}$, we estimate $\hat{S}$ from $\hat{S}$. In the exact recovery setting, the probability of error is given by

$$P_e = \mathbb{P}[\hat{S} \neq S]. \tag{3.17}$$

Theorem 4 (Group testing under BSC with exact recovery)**.** *Under the noisy group testing setup described above, in order to achieve $P_e \leq \delta$, we must have*

$$n \geq \frac{k \log \frac{p}{k}}{\log 2 - H_2(\epsilon)}(1 - \delta - o(1)) \tag{3.18}$$

as $p \to \infty$, possibly with $k \to \infty$ simultaneously.

Proof. We can directly formulate this problem as a multiple hypothesis test with $V = S$. Applying Fano's inequality with conditioning on $\mathbf{X}$, we obtain

$$I(S; \hat{S}|\mathbf{X}) \geq (1 - \delta) \log \binom{p}{k} - \log 2. \tag{3.19}$$

We upper bound $I(S; \hat{S}|\mathbf{X}) \leq I(S; \mathbf{Y}|\mathbf{X})$ using the data processing inequality, since $S \to \mathbf{Y} \to \hat{S}$ when conditioned on $\mathbf{X}$. Since the noise variables Z_i are independent, and Y_i are conditionally independent of X_i and S given $\bigvee_{j \in S} X_{ij}$, we have

$$I\left(S; \mathbf{Y}|\mathbf{X}\right) \leq \sum_{i=1}^{n} I(\bigvee_{j \in S} X_{ij}; Y_i). \tag{3.20}$$

Since Y_i is generated from $\bigvee_{j \in S} X_{ij}$ using a BSC, which has capacity $\log 2 - H_2(\epsilon)$, we have

$$I\left(S; \mathbf{Y}|\mathbf{X}\right) \leq n(\log 2 - H_2(\epsilon)). \tag{3.21}$$

Substituting the inequality $\binom{p}{k} \geq (\frac{p}{k})^k$ and (3.21) into (3.19) gives us Thm. 4 [49].

□

Similarly, we can model the error using a BEC to obtain the following theorem. This would correspond to the case in which test outcomes are undetermined or missing with probability P_ϵ.

Theorem 5 (Group testing under BEC with exact recovery)**.** *Under the group testing setup with missing test outcomes, in order to achieve $P_e \leq \delta$, we must have*

$$n \geq \frac{k \log \frac{p}{k}}{\log 2 - P_\epsilon}(1 - \delta - o(1)) \tag{3.22}$$

as $p \to \infty$, possibly with $k \to \infty$ simultaneously.

Proof. The proof is similar to the proof of Thm. 4, except that the BEC has a capacity of $\log 2 - P_\epsilon$. □

For the approximate recovery case, the decoder outputs a list $\mathcal{L} \subseteq \{1, ..., p\}$ of cardinality $L \geq k$. We can then define

$$P_e(\alpha k) = \mathbb{P}[|S \setminus \mathcal{L}| > \alpha k]. \tag{3.23}$$

In other works, if the number of missed defective items exceeds some fraction of the total number of defective items, then the decoder is wrong.

Theorem 6 (Group testing under BSC with approximate recovery)**.** *Under the noisy group testing setup with approximate recovery, with list size $L \geq k$, in order to achieve $P_e(\alpha k) \leq \delta$ for some $\alpha \in (0, 1)$, we need*

$$n \geq \frac{(1-\alpha)k \log \frac{p}{L}}{\log 2 - H_2(\epsilon)}(1 - \delta - o(1)) \tag{3.24}$$

as $p \to \infty, k \to \infty$, and $L \to \infty$ simultaneously with $L = o(p)$.

Proof. We apply Thm. 3. The number of $\mathcal{L}$ with cardinality L within a neighborhood αk of S is given by

$$N_{max}(\alpha k) = \sum_{j=0}^{\lfloor \alpha k \rfloor} \binom{p-L}{j} \binom{L}{k-j}, \tag{3.25}$$

which is the number of ways to place up to αk items in $\mathcal{L}$. Hence, conditioning on $\mathbf{X}$ and the data processing inequality as in Thm. 4, we get

$$I(S; \mathbf{Y}|\mathbf{X}) \geq (1 - \delta) \log \frac{\binom{p}{k}}{\sum_{j=0}^{\lfloor \alpha k \rfloor} \binom{p-L}{j} \binom{L}{k-j}} - \log 2. \tag{3.26}$$

Using some asymptotic simplifications (described in [49]), we combine (3.26) and (3.21) to obtain Thm. 7.

□

Again, this can be extended to a BEC corresponding to missing test outcomes.

Theorem 7 (Group testing under BEC with approximate recovery)**.** *Under the BEC group testing setup with approximate recovery, with list size $L \geq k$, in order to achieve $P_e(\alpha k) \leq \delta$ for some $\alpha \in (0, 1)$, we need*

$$n \geq \frac{(1-\alpha)k \log \frac{p}{L}}{\log 2 - P_\epsilon}(1 - \delta - o(1)) \tag{3.27}$$

as $p \to \infty, k \to \infty$, and $L \to \infty$ simultaneously with $L = o(p)$. Note that P_ϵ is the error rate of the BEC.

Proof. The proof is the same as that of Thm. 7 but with the BEC channel capacity. □

In this example, we assumed i.i.d. Bernoulli missingness, corresponding to the MCAR case. To account for the MAR and MNAR cases, there has been some work on the capacity of signal-dependent noise channels [52], but this is an open area for future research.

CHAPTER 4

REGISTRATION FOR IMAGE-BASED TRANSCRIPTOMICS: PARAMETRIC SIGNAL FEATURES AND MULTIVARIATE INFORMATION MEASURES

Image-based transcriptomics involves determining spatial patterns in gene expression across cells and tissues. Image registration is a necessary component of data analysis pipelines that characterize gene expression levels across different cells and intracellular structures. We consider images from multiplexed single molecule fluorescent *in situ* hybridization (smFISH) and multiplexed *in situ* sequencing (ISS) datasets from the Human Cell Atlas project and demonstrate a novel approach to groupwise image registration using a parametric representation of images based on finite rate of innovation sampling, together with practical optimization of empirical multivariate information measures.

4.1 Introduction

The transcriptome of a cell (or an organism) is the portion of DNA that is expressed as RNA in that cell (or organism). The subset of genes that are expressed in a cell varies depending on cell type and cell state, and recognizing patterns in the transcriptome is an important part of understanding cell function and pathology. The RNA molecules present in a cell can be tagged using fluorescent markers, which can be observed *in situ*, and spatial patterns of gene expression within cells and tissues can be studied [53]. Different combinations of colored fluorescent markers serve as tags for different RNA sequences, and images of different color spectra (multispectral images) must be aligned for analysis.

Researchers have developed methods and algorithms to extract transcript molecule feature sets, localization, and patterns, in tens of thousands of single cells across the human transcriptome [54, 55]. Starfish, a community of computational biologists and software engineers, has developed standardized

file formats for input data and analysis [56]. They are building a standard library that consolidates different methods from different steps of the analysis pipeline. The goal is to contribute to the Human Cell Atlas, a project to create comprehensive reference maps of all human cells to understand human health and to diagnose, monitor, and treat disease [57].

An important part of the image-based transcriptomics analysis pipeline is image registration. Image registration in the Starfish pipeline currently addresses pixel-level translational error using an FFT-based phase correlation approach [58]. Phase correlation has also been extended to rotational and scaling error [59] as well as to subpixel registration [60]. However, phase correlation does not account for non-linear intensity changes, which can arise when the images to be registered are taken under different conditions (e.g. multimodal, multispectral, different lighting, etc.). To address such intensity nonlinearities, mutual information has been proposed elsewhere as an image similarity measure [61, 62], and under the assumption that images are statistically dependent, we have shown that maximizing mutual information is a theoretically optimal method for registering image pairs [63].

What about registering not just a pair of images but a larger set of images, as in transcriptomics? A natural extension to mutual information is multiinformation, which can be used to jointly register multiple images and has been shown to be superior to sequential pairwise registrations and asymptotically optimal [63]. However, computing the multiinformation of several images is computationally expensive—$O(N^n)$ for n images with N pixels. To address this computational issue we develop a novel parametric signal representation for image-based transcriptomics data using finite rate of innovation sampling [64]. We propose a feature- and information-based method which is $O(nN)$ for n images. This algorithm registers images in a joint rather than pairwise manner and has the same output as multiinformation in certain settings when features are properly extracted; see Thm. 8 for a formal statement.

4.2 Registration using information

4.2.1 Mutual information

Maximizing empirical mutual information for image registration [61, 62] has been commonly used in fields such as medical imaging and remote sensing. Mutual information between random variables X and Y is defined as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{4.1}$$

where $H(X)$ and $H(Y)$ are marginal entropies and $H(X,Y)$ is the joint entropy of X and Y. Mutual information can also be formulated as the Kullback-Liebler divergence (relative entropy) between the joint distribution and the product of the marginal distributions:

$$I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right). \tag{4.2}$$

Given two r-dimensional images X_1 and X_2 defined over a discrete spatial region Ω, we define an image transformation $T : \Omega \to \Omega$. If X_2 is a transformed version of X_1, we can register X_1 and X_2 by finding

$$T^* = \arg\max_T I(X_1(x,y); X_2(T(x,y))). \tag{4.3}$$

There are several ways to estimate the joint and marginal distributions of images. Two popular ones are the joint histogram method [65] and the Parzen windowing method [66, 61]. After estimating the distributions, the maximization problem (4.3) can then solved using an appropriate optimization algorithm. If properly initialized, a local optimization algorithm such as gradient descent can be used [65]. Global optimization algorithms such as genetic algorithms, simulated annealing, and particle swarm optimization have also been used to maximize mutual information [67, 68, 69, 70].

4.2.2 Multiinformation

Several groups have proposed using multiinformation to perform groupwise registration [71, 72, 73]. Guyader et al. show that groupwise multiinforma-

tion yields better registration performance than pairwise mutual information for medical imaging settings. Multiinformation, like mutual information, is defined as the KL divergence between a joint distribution and a product of marginal distributions. Given n random variables $X_1, X_2, ..., X_n$, multiinformation is defined as

$$I(X_1; X_2; \ldots; X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n)$$

$$= \sum p(x_1, x_2, \ldots, x_n) \log \left(\frac{p(x_1, x_2, \ldots, x_n)}{p(x_1)p(x_2) \cdots p(x_n)} \right). \tag{4.4}$$

To register n images $X_1, X_2, \ldots, X_n$, we find the transformations that maximize multiinformation:

$$T_2^*, \ldots, T_n^* = \underset{T_2,\ldots,T_n}{\arg\max}\, I(X_1(\mathbf{x}); X_2(T_2(\mathbf{x})); \ldots; X_n(T_n(\mathbf{x}))), \tag{4.5}$$

where $\mathbf{x} = (x, y)$ for a 2-D image. Thus a single optimization can be used to register any number of images. We call this the *MM algorithm*. We have recently shown that MM is exponentially consistent for image registration (probability of error goes to 0 exponentially fast as number of pixels goes to infinity) [63]. In fact, we showed that MM is asymptotically optimal in the sense of achieving the same error exponent as maximum likelihood registration. Thus

$$\mathbf{T}_{MM}^* \doteq \mathbf{T}_{ML}^* = \underset{T_1,\ldots,T_n}{\arg\max} \prod_{i=1}^{n} P(X_i(T_i(x, y))|X_i), \tag{4.6}$$

where "$\doteq$" indicates equivalent error exponents.

Guyader et al. assume images are jointly Gaussian, which greatly simplifies the empirical distribution estimation [71]. This is usually not a valid assumption for natural images or image-based transcriptomics (although they show that it works for the images they tested). Kern et al. [72] use the approximation

$$I(X_1; X_2; \ldots; X_n) \approx \sum_{i,j \in \{1,\ldots,n\}; i \neq j} I(X_i, X_j). \tag{4.7}$$

This approximation can greatly overestimate multiinformation in cases where mutual information of image pairs is large.

Rather than approximating the multiinformation function, we initially implement the original objective in (4.5) using full image histograms. We perform our optimization using particle swarm optimization [74]. The details of our implementation are covered in Sec. 4.4.

4.3 Feature-based registration

4.3.1 Background and related work

While MM performs groupwise registration optimally [63], it is computationally expensive. Feature-based algorithms tend to be more efficient. Many registration methods first extract features and then match only the extracted features across images. Others have used corner detectors and edge detectors for registration [75, 76]. Phase correlation in the Fourier domain is applied to these features to retrieve translations and rotations. A particularly successful method is based on Lowe's scale-invariant feature transform (SIFT) [77]. SIFT extracts local scale-invariant features, called keypoints, from images. Key points are image features that are visually interesting (e.g., corners and curved edges), and feature matching and clustering is performed to detect and register objects. Numerous variants of SIFT have been used in multimodal image registration (we reference just a few) [78, 79, 80, 81].

Baboulaz and Dragotti formulate feature-based registration as a finite rate of innovation problem, with applications in super-resolution [82]. They use a Canny-like edge detector and use the intersections of those edges (corners) as features. They then match corners across images using correlation and RANSAC methods (also used in SIFT) and demonstrate that exact registration is possible using only the detected corners.

We apply finite rate of innovation sampling to multispectral registration by representing smFISH images as sums of delta functions. We demonstrate that these features alone are enough to perform registration. We also use this sparse representation in conjunction with MM to perform groupwise registration.

4.3.2 Finite rate of innovation sampling

Certain signals that have a *finite rate of innovation* (FRI) [64] can be written in the form

$$s(t) = \sum_k c_k \phi(t - t_k), \tag{4.8}$$

where $\phi(t)$ is a known kernel and the number of t_k values per unit time (the rate of innovation) is finite. Then the *innovative* part of the signal lies in c_k. Given c_k and t_k, we can reconstruct $s(t)$. Various kinds of filters can be used to identify the innovative part of the signal, and the signal can be perfectly reconstructed using just these samples [83, 84]. Examples of FRI signals include delta trains and piecewise polynomials.

For a 2-dimensional image, we can write (4.8) as

$$s(x, y) = \sum_{x', y' \in \Omega} c_{x', y'} \phi(x - x', y - y'). \tag{4.9}$$

Baboulaz and Dragotti register multiview images using FRI sampling. They model images as sums of polynomials, using a B-spline sampling kernel [82]. The images captured by digital imaging technologies result from the point spread function (PSF) of a lens, which can be used as the sampling kernel $\phi(x, y)$ in (4.9). An object in space $o(x, y)$ is filtered through the lens as

$$s(x, y) = o(x, y) * \phi(-x/T, -y/T), \tag{4.10}$$

where T is the sampling period. For registration, we must obtain the features of $o(x, y)$ from $s(x, y)$. This can be done by deconvolving $s(x, y)$ with the PSF of the imaging system to give $\hat{o}(p, q)$. The image $\hat{o}(x, y)$ is then processed to extract the features of interest, giving us a weighted sum of spikes:

$$d(x, y) = \sum_{x', y' \in \Omega} c_{x', y'} \delta(x - x', y - y'). \tag{4.11}$$

For every feature, there will be a corresponding spike in each misaligned image. To perform groupwise registration, we group these features across images and find the transformations so that spikes corresponding to a feature have the same location in each image.

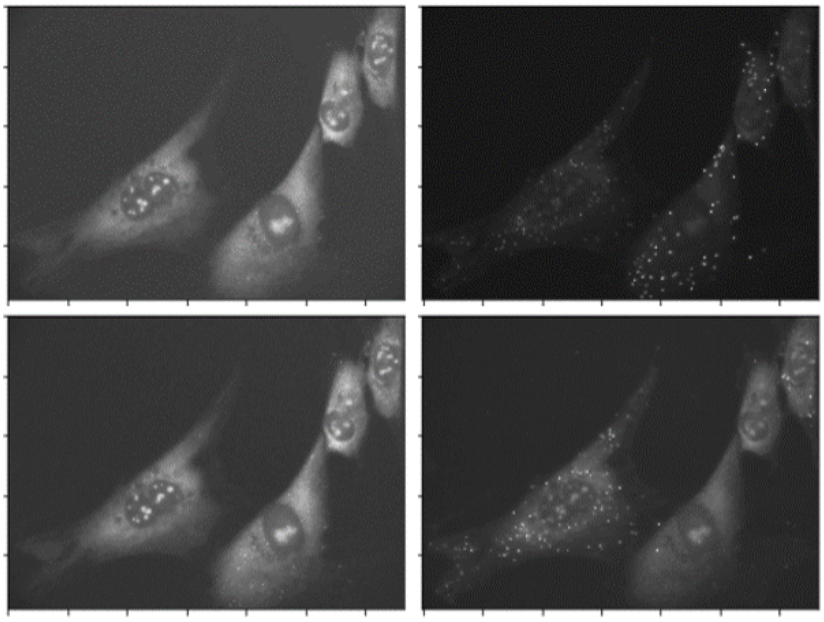

Figure 4.1: Example smFISH images. Columns are different color channels. Rows are imaging rounds.

4.4 Methods and experiments

In this section, we demonstrate our algorithmic approach on manually shifted and rotated multispectral images, and then on a single smFISH image which we manually shift and rotate.[1] Fig. 4.1 shows some example smFISH images. Columns correspond to a color channel (there are three total), while rows corresponds to a round of imaging. Disturbances between imaging rounds may cause images to shift, and image registration is necessary. We perform a maximum intensity projection across channels to obtain one image per imaging round, and we perform registration with these images. In Sec. 4.4.3 we present our results on misaligned smFISH images, for which we do not have ground truth data.

4.4.1 Maximizing multiinformation

We implement MM using particle swarm optimization (PSO). To maximize (4.5), we must search over the space $T_2 \times T_3 \times \cdots \times T_n$, where each T_i can have multiple degrees of freedom. For example, if we restrict T to include only

[1] Code can be found at `https://github.com/toby2476/fish_register`

vertical and horizontal translations, each image has two degrees of freedom, giving us a 2^{n-1}-dimensional search space. If we include rotations, it becomes a 3^{n-1}-dimensional search space. We can also include scaling, shearing, etc. Clearly (4.5) is much more difficult to optimize than (4.3). We plot the multiinformation of 30 multispectral images from the CAVE database [85] with random horizontal and vertical shifts, estimating distributions using histograms (see Fig. 4.2). The maximum shift M is given on the x-axis, and images were shifted with a x-shift and a y-shift drawn uniformly on $[0, M]$. This matches the plots in [72] showing that mutual information of two images is maximized when they are properly aligned.

To test registration, we use five multispectral images, shown in Fig. 4.3. We randomly rotate them with angles between $[-5, 5]$ degrees. We find the empirical image intensity distributions for each image using image histograms, and compute the empirical joint distribution using a joint histogram. Using numerous histogram bins causes joint probabilities to become too small, so we use four bins. We calculate entropies and joint entropies using the empirical distributions, obtain multiinformation using (4.2), and maximize (4.3) using PSO. Because histograms are computationally expensive, we resized the images to decrease image sizes. We begin with exhaustive search of multiinformation over all rotations and apply the transformation that maximizes multiinformation. (We restrict to rotations—only one degree of freedom per image—so we can test exhaustive search.) Comparing this to PSO yields the same results. Fig. 4.4 shows images transformed with shifts drawn uniformly from $[-10, 10]$ pixels and rotations drawn uniformly from $[-10, 10]$ degrees. This example was initialized with 300 particles and converged after 211 iterations.

We also test a single maximum projection smFISH image; rather than resizing, we crop images and perform registration on areas with high densities of fluorescent markers. To increase the speed of PSO convergence, we use a dictionary to store the result of every computation so that the same calculations need not be repeated, and we introduce a spread factor to the PSO to improve convergence speed [86].

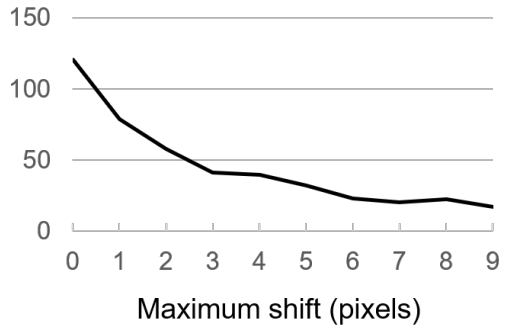

Figure 4.2: Multiinformation of 30 multispectral images as a function of random shifts.

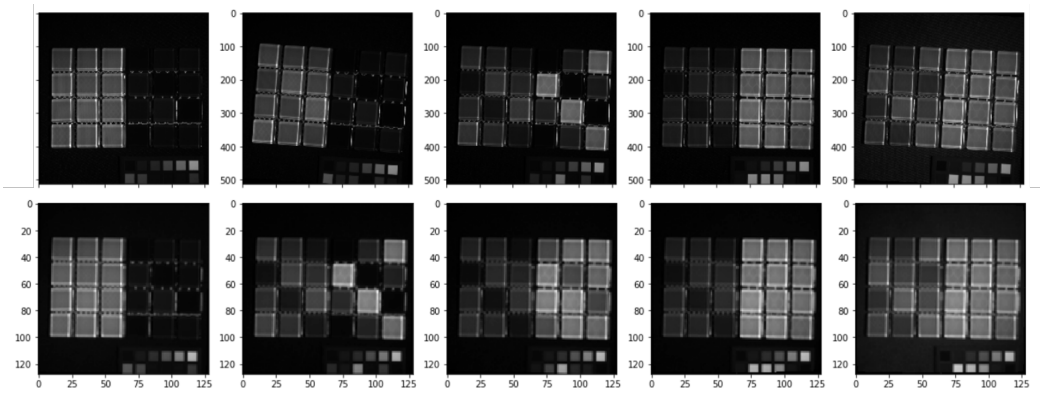

Figure 4.3: Top: Rotated test images. Bottom: Images registered using MM.

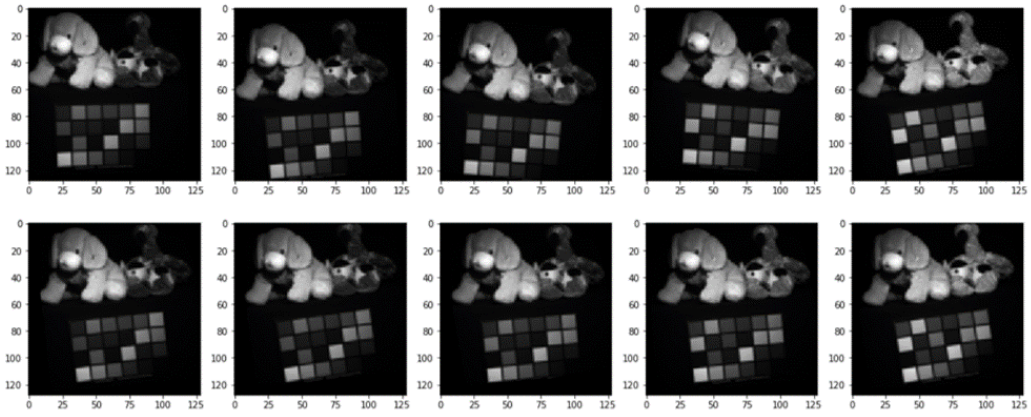

Figure 4.4: Top: Rotated and shifted test images. Bottom: Images registered using MM.

4.4.2 Finite rate of innovation sampling

If the weighted spikes corresponding to innovative features are correctly located across each image, the sampled spike images will be transformed versions of the reference image, with transformations $\mathbf{T}_\delta$ that are exactly the original transformations $\mathbf{T}$. If a spike δ_i has moved an $L-2$ distance less than D from the reference spike δ_1 for all images $i = 2, \ldots, n$, and if any spike corresponding to a feature is significantly greater than distance D from all other spikes in the image (that is, spikes are far apart and shifts and rotations were small), then we can use a nearest neighboring clustering algorithm to group spikes corresponding to the same feature across images.

We first demonstrate that we can register a collection of randomly shifted spikes. We randomly generate a 1000×1000 pixel image of 500 maximum intensity pixels (spikes) on a black background. We translate the image and use a nearest neighbor search to find where a spike has shifted. Both our algorithm and phase correlation register the images correctly with no error. We then add noise by randomly removing 50 spikes and randomly introducing 50 spikes. We then randomly shift individual spikes by 1 pixel to any of its eight adjacent locations. We find that both our algorithm and phase correlation produce an error of 1 pixel in any direction for about half of the images.

We test registration of multispectral images by randomly applying a random horizontal and vertical shift drawn from $[-25, 25]$, and we sample using Rosten and Drummond's fast corner detection [87]. We represent corner locations as spikes (see Fig. 4.5) and perform a nearest neighbor search using a search neighborhood of 30. FRI sampling followed by clustering again registers the images correctly around 40% of the time, and is off by 1 pixel in any direction 60% of the time. On the other hand, cross-correlation gives poor results when the images look drastically different in the various spectra. Fig. 4.6 shows a typical example when MM and our FRI method correctly register images, but phase correlation does not.

Next we find a sampling procedure that works well with smFISH images, taking fluorescent markers as features. We begin again by performing a maximum intensity projection across all three color channels to obtain one image per imaging round. We locate the fluorescent markers using a White Top Hat filter, which extracts small points that are brighter than their surrounding

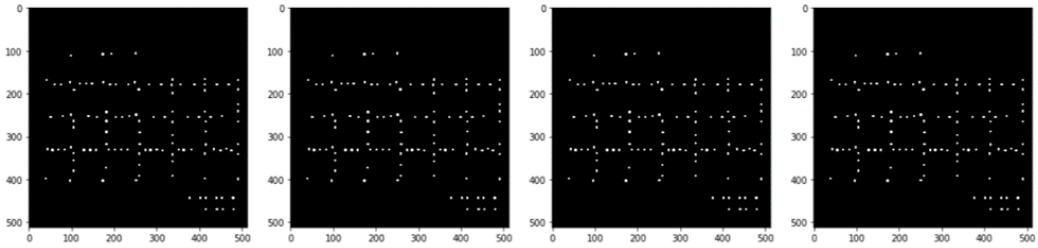

Figure 4.5: Spikes found by corner detection.

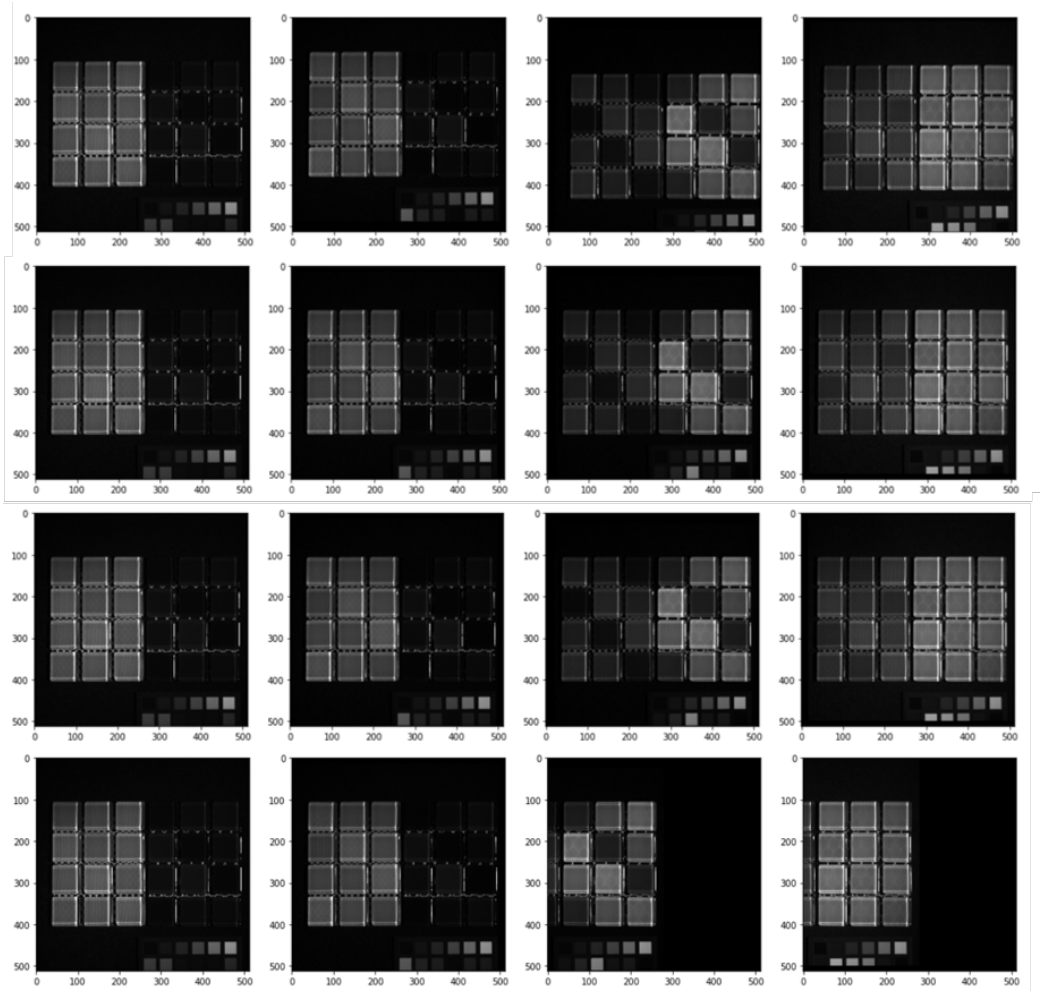

Figure 4.6: From top to bottom: Manually shifted test images, registration using MM (correctly registered), registration using FRI sampling (correctly registered), registration using phase correlation (incorrectly registered).

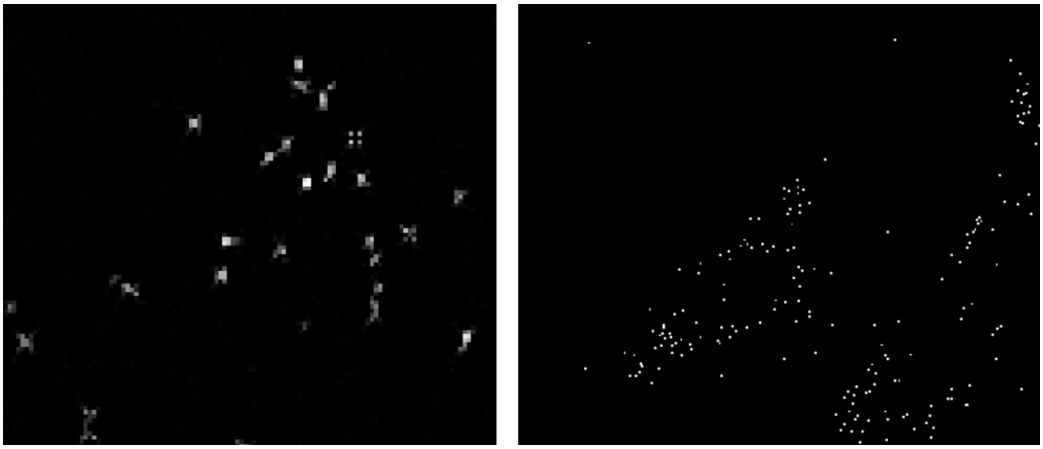

Figure 4.7: Left: Groups of pixels corresponding to fluorescent markers (enlarged for visibility). Right: spike deltas corresponding to centroids of clusters.

and threshold to remove any low-intensity points generated by noise.

Since each fluorescent marker is several pixels large, we must find a single pixel location that best approximates each fluorescent marker. We use density-based spatial clustering of applications with noise (DBSCAN) since the number of clusters does not need to be specified and since it does not cluster outliers [88]. DBSCAN clusters pixels corresponding to a single fluorescent marker (see Fig. 4.7). Once clusters are identified, we find the coordinates of their centroids.

We obtain a list of reference spikes from the reference image and we align the spikes in each misaligned image to the reference spikes using nearest neighbor clustering across images. We find that results are more accurate when the search range for each point was small, so we use a successive refinement of the search radius. We perform the search first with a larger search radius of 50 and then with a smaller search radius of 25.

4.4.3 Experimental results

Registration results for smFISH images are shown in Figs. 4.8 and 4.9. Since there is no ground truth, we closely inspect image regions to visually confirm correct registration (see Fig. 4.10). Comparing our results to results using phase correlation indicate improved performance. Registration of four 980×1330 images takes under a second with both FRI and phase correlation and

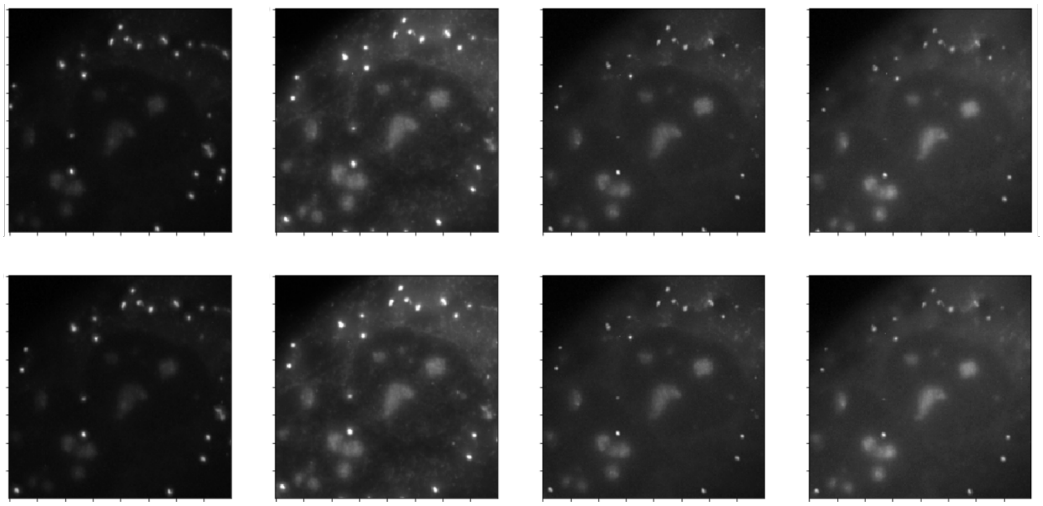

Figure 4.8: Registration results. Top: Before registration. Bottom: After registration with FRI

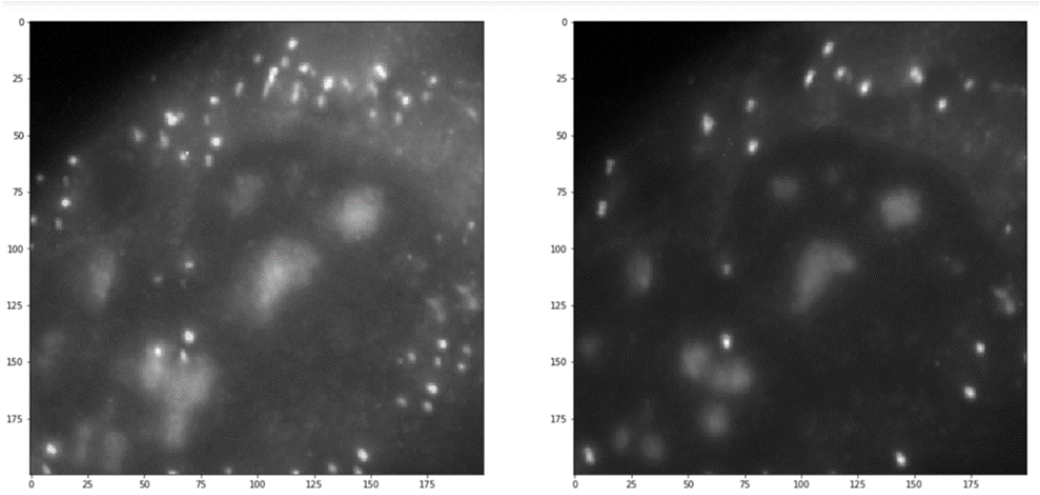

Figure 4.9: Left: Overlay of image rounds before registration. Right: Overlay of image rounds after registration with FRI.

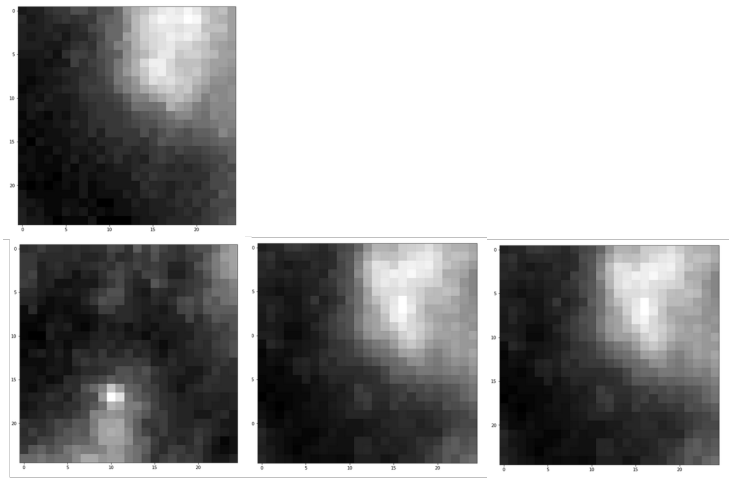

Figure 4.10: Close of up pairwise registration results. Top: Reference image. Bottom left: Before registration. Bottom center: After registration with MM. Bottom right: After registration with FRI.

15 minutes with MM on the same machine.

4.5 Discussion

Not only does finite rate of innovation sampling improve computational efficiency of groupwise image registration, as we have shown, but it is asymptotically optimal for image registration if it is used with multiinformation.

Theorem 8. *If images $X_1, \ldots, X_n$ are FRI signals, maximizing multiinformation of the images sampled at FRI is asymptotically optimal for image registration.*

Proof. The maximum likelihood estimator of image transformations when transformations are independent is given by:

$$\hat{\mathbf{T}}_{ML} = \arg\max_{\mathbf{T}} \prod_{i=2}^{n} \mathbb{P}[X_i(T_i(\mathbf{x}))|X_1(\mathbf{x})], \tag{4.12}$$

which is Bayes optimal. The MM estimate for image registration

$$\hat{\mathbf{T}}_{MM} = \arg\max_{\mathbf{T}} I\left(X_1(\mathbf{x}); X_2(T_2(\mathbf{x})); \ldots; X_n(T_n(\mathbf{x}))\right) \tag{4.13}$$

is exponentially consistent with $\hat{\mathbf{T}}_{ML}$, as we showed in [63, Theorem 7]. Let

$\hat{X}_i$ be X_i sampled at or above the rate of innovation for $i = 1, \ldots, n$. Then there is no loss of information from $X_i(T_i(\mathbf{x}))$ to $\hat{X}_i(T_i(\mathbf{x}))$ [64], and

$$I\left(X_1(\mathbf{x}); X_2(T_2(\mathbf{x})); \ldots; X_n(T_n(\mathbf{x}))\right) = \tag{4.14}$$

$$I\left(\hat{X}_1(\mathbf{x}); \hat{X}_2(T_2(\mathbf{x})); \ldots; \hat{X}_n(T_n(\mathbf{x}))\right). \tag{4.15}$$

Thus $\hat{\mathbf{T}}_{MM}$ of the FRI-sampled images is equivalent to $\hat{\mathbf{T}}_{MM}$ of the original images, which is asymptotically optimal. □

Although MM is theoretically optimal, multiinformation is costly to compute even for FRI-sampled signals, so in practice we have used nearest neighbor clustering of sampled images. We have demonstrated that this heuristic approximation is effective. In the future, we aim to examine efficacy and computational efficiency of our technique for image-based transcriptomics at the massive scale of the Human Cell Atlas.

CHAPTER 5

CONCLUSION

As data collection methods become more efficient and computer processors more powerful, researchers are able to analyze larger amounts of data and better understand the complex interactions in different types of systems, biological or otherwise. New algorithms must be developed to better understand the data, and new theoretical bounds should be found to benchmark the algorithms.

In this thesis, we first introduced missing values as a common form of data corruption and presented an imputation algorithm based on optimal recovery. This is a useful approach when data is clustered, and it does not rely on any statistical assumptions. We found deterministic and probabilistic error bounds for our algorithm when used in the context of non-negative matrix factorization and tested our algorithm on imaging and mouse protein data.

We then showed that missingness mechanisms can be modeled as erasure channels, and demonstrated the use of Fano's inequality to find lower bounds for missing values settings. We illustrated with a group testing example.

Finally, we looked at mis-registered images as a form of corrupted data. We performed multi-image registration by maximizing multiinformation, which is a theoretically optimal approach. We retained optimality while speeding up registration by using finite rate of innovation sampling, and we tested our algorithm on multispectral images and fluorescent RNA images.

Future work includes characterizing MAR and MNAR mechanisms as noisy channels to find new lower bounds. This can be used to find missing value bounds in the NMF context for non-MCAR settings. On the image registration side, we assumed that certain features (e.g. corners and fluorescent markers) in images carried most of the information. One can compare the information content of different features in a more systematic way or discover the features that carry the most information [89].

REFERENCES

[1] P. Hogeweg, "The roots of bioinformatics in theoretical biology," *PLoS Computational Biology*, vol. 7, no. 3, Mar. 2011.

[2] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London*, vol. 237, no. 641, pp. 37–72, Aug. 1952.

[3] T. K. Olsen and N. Baryawno, "Introduction to single-cell RNA sequencing," *Current Protocols in Molecular Biology*, vol. 122, no. 1, Apr. 2018.

[4] O. Rozenblatt-Rosen et al., "The human cell atlas: from vision to reality," *Nature*, vol. 550, no. 7677.

[5] 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.

[6] 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.

[7] G. L. Stein-O'Brien, R. Arora, A. C. Culhane, A. V. Favorov, L. X. Garmire, C. S. Greene, L. A. Goff, Y. Li, A. Ngom, M. F. Ochs, Y. Xu, and E. J. Fertig, "Enter the matrix: factorization uncovers knowledge from omics," *Trends in Genetics*, vol. 34, no. 10, pp. 790–805, Oct. 2018.

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

[9] 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.

[10] 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.

[11] C. A. Micchelli and T. J. Rivlin, "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.

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

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

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

[15] 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)*, July 2012.

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

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

[18] K. Bhaskaran and L. Smeeth, "What is the difference between missing completely at random and missing at random?" *International Journal of Epidemiology*, vol. 43, no. 4, pp. 1336–1339, 2014.

[19] Y. Ding and J. S. Simonoff, "An investigation of missing data methods for classification trees applied to binary response data," *J. Machine Learning Research*, vol. 11, pp. 131–170, Jan. 2010.

[20] P. J. García-Laencina and J.-L. Sancho-Gómez and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

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

[22] C. K. Enders, *Applied Missing Data Analysis*. The Guilford Press, 2010.

[23] 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.

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

[25] 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.

[26] M. J. Azur et al., "Multiple imputation by chained equations: What is it and how does it work?" *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011.

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

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

[29] 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.

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

[31] A. C. Olivier Delalleau and Y. Bengio, "Efficient EM training of Gaussian mixtures with missing data," *arXiv:1209.0521 [cs.LG]*, Jan. 2018.

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

[33] C.-C. Chiu et al., "Missing value imputation for microarray data: a comprehensive comparison study and a web tool," *BMC Systems Biology*, vol. 7, no. Suppl 6:S12, Dec. 2013.

[34] M. C. de Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC Bioinformatics*, vol. 16, no. 64, Feb. 2015.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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)*, June 2017.

[40] 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.

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

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

[43] 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.

[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." *Briefings in Bioinformatics*, 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: 2019-10-29. [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_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: e0129126, 2015.

[48] T. H. Hopp and C. P. Reeve, "An algorithm for computing the minimum covering sphere in any dimension," *NISTIR*, no. 5831, 1996.

[49] J. Scarlett and V. Cevher, "An introductory guide to Fano's inequality with applications in statistical estimation," in *Information-Theoretic Methods in Data Science*, Y. Eldar and M. Rodrigues, Eds. Cambridge University Press, 2019 (Expected).

[50] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, Inc., 2006.

[51] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation," *arXiv:1311.2669 [cs.IT]*, Dec. 2013.

[52] H. Ghourchian et al., "On the capacity of a class of signal-dependent noise channels," *IEEE Informat. Th.*, vol. 64, no. 12, pp. 7828–7846, Dec. 2018.

[53] P. R. Langer-Safer, M. Levine, and D. C. Ward, "Immunological method for mapping genes on drosophila polytene chromosomes." *Proc. Nat. Acad. Sci.*, vol. 79, no. 14, pp. 4381–4385, July 1982.

[54] T. Stoeger, N. Battich, M. Herrmann, Y. Yakimovich, and L. Pelkmans, "Computer vision for image-based transcriptomics," *Methods*, vol. 85, pp. 44–53, Sep. 2015.

[55] N. Battich, T. Stoeger, and L. Pelkmans, "Image-based transcriptomics in thousands of single human cells at single-molecule resolution," *Nature Methods*, vol. 10, pp. 1127–1133, 2013.

[56] S. Axelrod et al., "Starfish: Open source image based transcriptomics and proteomics tools," 2018. [Online]. Available: http://github.com/spacetx/starfish

[57] Human Cell Atlas, "Human cell atlas," 2018. [Online]. Available: https://www.humancellatlas.org/

[58] C. D. Kuglin and D. C. Hines, "The phase correlation image alignment method," *Proc. IEEE 1975 Int. Conf. Cybernet. Soc.*, pp. 163–165, July 1975.

[59] B. Reddy and B. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, Aug. 1996.

[60] A. Alba, R. M. Aguilar-Ponce, J. F. Vigueras-Gómez, and E. Arce-Santana, "Phase correlation based image alignment with subpixel accuracy," in *MICAI 2012: Advances in Artificial Intelligence. Lecture Notes in Computer Science*, vol. 7629, 2012.

[61] P. Viola and W. M. Wells, "Alignment by maximisation of mutual information," *Proc. IEEE Int. Conf. Comput. Vision*, vol. 24, no. 2, pp. 16–23, June 1995.

[62] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multi-modality image registration based on information theory," in *Information Processing in Medical Imaging.* The Netherlands: Kluwer Academic Publishers, 195, pp. 263–274.

[63] R. K. Raman and L. R. Varshney, "Universal joint image clustering and registration using multivariate information measures," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 928–943, Oct. 2018.

[64] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[65] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 373–386, 1999.

[66] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[67] R. He and P. Narayana, "Improvement to global optimization of mutual information on retrospective registration of magnetic resonance images," in *Proc. 2nd Joint 24th Annu. Conf. Annu. Fall Meet. Biomed. Eng. Soc*, 2002.

[68] B. Farsaii and A. Sablauer, "Global cost optimization in image registration using simulated annealing," in *Proc. SPIE Conf. Math. Model. Estimation Tech. Comput. Vision*, vol. 3457, 1998, pp. 117–125.

[69] G. K. Matsopoulos, N. A. Mouravliansky, K. K. Delibasis, and K. S. Nikita, "Automatic retinal image registration scheme using global optimization techniques," *IEEE Trans. Inf. Technol. Biomed.*, vol. 3, no. 3, pp. 47–60, Mar. 1999.

[70] Q. Li and I. Sato, "Multimodality image registration by particle swarm optimization of mutual information," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Lecture Notes in Computer Science*, vol. 4682, 2007.

[71] J.-M. Guyader, W. Huizinga, V. Fortunati, D. H. Poot, M. van Kranenburg, J. F. Veenland, M. M. Paulides, W. J. Niessen, and S. Klein, "Total correlation-based groupwise image registration for quantitative MRI," in *2016 IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016.

[72] J. Kern, M. Pattichis, and S. Stearns, "Registration of image cubes using multivariate mutual information," in *37th Asilomar Conf. Signals, Syst., Comput.*, 2003.

[73] L. Zollei, "A unified information theoretic framework for pair- and group-wise registration of medical images," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, Jan. 2006.

[74] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Networks*, 1995, pp. 1942–1948.

[75] H. Li, B. S. Manjunath, and S. K. Mitra, "A contour-based approach to multisensor image registration," *IEEE Trans. Image Process.*, vol. 4, no. 3, pp. 320–334, Mar. 1995.

[76] Z. Li, X. Yang, and L. Wu, "Image registration based on Hough transform and phase correlation," in *Int. Conf. Neural Networks and Signal Proc.*, 2003.

[77] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vision*, vol. 2, 1999, pp. 1150–1157.

[78] M. Brown and S. Susstrunk, "Multi-spectral SIFT for scene category recognition," in *Comput. Vision Pattern Recognit.*, 2011.

[79] T. Hossain et al., "Improved symmetric-sift for multi-modal image registration," in *2011 Int. Conf. Digital Image Computing: Techniques Appl.*, 2011.

[80] Mahesh and M. V. Subramanyam, "Automatic feature based image registration using SIFT algorithm," in *Third Int. Conf. on Computing, Communication and Networking Technologies*, vol. 3, 2012.

[81] S. Paul, U. K. Durgam, and U. C. Pati, "Automatic feature based image registration using SIFT algorithm," in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Advances in Intelligent Systems and Computing*, vol. 518, 2017.

[82] L. Baboulaz and P. L. Dragotti, "Exact feature extraction using finite rate of innovation principles with an application to image super-resolution," *IEEE Trans. Signal Process.*, vol. 18, no. 2, pp. 281–298, Feb. 2009.

[83] B. D. He, A. Wein, L. R. Varshney, J. Kusuma, A. G. Richardson, and L. Srinivasan, "Generalized analog thresholding for spike acquisition at ultra-low sampling rates," *J. Neurophysiol.*, vol. 114, no. 1, pp. 746–760, July 2015.

[84] J. Kusuma, "Economical sampling of parametric signals," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2006.

[85] Columbia Vision and Graphics Center, "Multispectral image database," 2018. [Online]. Available: http://www.cs.columbia.edu/CAVE/databases/multispectral/

[86] I. A. Latiff and M. O. Tokhi, "Fast convergence strategy for particle swarm optimization using spread factor," in *2009 IEEE Congress on Evolutionary Computation*, 2009, pp. 2693–2700.

[87] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *2006 European Conf. on Computer Vision*, 2006, pp. 430–443.

[88] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[89] G. V. Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," *arXiv:1406.1222 [cs.LG]*, Oct. 2014.