COMPRESSIVE SAMPLING FOR PHENOTYPE CLASSIFICATION

DISSERTATION

Eric L. Brooks, Maj, USAF

AFIT-ENC-DS-18-S-001

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-DS-18-S-001

COMPRESSIVE SAMPLING FOR PHENOTYPE CLASSIFICATION

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctorate of Philosophy in Applied Mathematics

Eric L. Brooks, M.S.

Maj, USAF

August 30, 2018

AFIT-ENC-DS-18-S-001

COMPRESSIVE SAMPLING FOR PHENOTYPE CLASSIFICATION

DISSERTATION

Eric L. Brooks, M.S.
Maj, USAF

Committee Membership:

Lt Col R. A. Kappedal, PhD
Chair

Dr. C. M. Schubert-Kabban
Member

LTC D. R. Lewis, PhD
Member

Abstract

Phenotype classification has become an increasingly important genomic research method for disease identification and treatment. Phenotypes are the observable traits of the deoxyribonucleic acid (DNA) that contains the chemical and environmental effects of the genetic profile. Phenotype classification is the investigation into the genetic information concerned with locating biomarkers (features) in order to identify an observed effect. The primary challenge associated with phenotype classification is with analyzing the data due to the inherent high-dimensionality of DNA data. High-dimensionality refers to the exorbitant features space. As a result, phenotype classification faces challenges with feature selection, and consequently, classification accuracy.

This research developed methodology to alleviate these challenges while improving classification accuracy. The methodology mainly leverages concepts of compressive sampling, specifically, incoherence, L_1-minimization, and the restricted isometry property (RIP) to arrive at a process that identifies features most relevant to the phenotype. Additionally, this research presents a probabilistic acceptance of the RIP and uses it to qualify data frames constructed by the proposed methodology. Overall, this methodology is a viable approach to dimension reduction and feature selection, which improved phenotype classification accuracy.

Acknowledgements

I would like to first give honor to God, who orders my steps and guides my life. I would like to thank my wife and kids for their support and patience. I would like to thank my advisor Lt Col Kappedal and my committee, Dr Schubert-Kabban and LTC Lewis. Last but not least I want to thank my grandmother for her inspiration and strength, Thanks Mudd.

Eric L. Brooks

Table of Contents

Page

List of Figures

List of Tables

COMPRESSIVE SAMPLING FOR PHENOTYPE CLASSIFICATION

I. Introduction

In this section I introduce phenotypes and the phenotype classification problem; I discuss the goals of my research; and I outline the focus of this paper.

1.1 What is Phenotype Classification?

Often when we speak about genealogy, heredity, or deoxyribonucleic acid (DNA), we are referencing the human genome; a person's genotype. Genotypes can be described as an individuals genetic information. Phenotypes are the observable traits within DNA that comprise the responses to the chemical effects of the genome itself or, the interaction between the genome and the environment [1]. Moreover, the genome hosts the entire genetic profile, while the biological and environmental changes to the genetic profile are captured within the phenotype.

Phenotype classification is the investigation into phenotypes concerned with identifying biomarkers, or classification features, for a specific observed effect. It has become an increasingly important genomic research method for disease identification and treatment; specifically in cancer research. Phenotype classification primarily contributes to cancer research by providing an expedient and accurate method to identify subjects at risk for cancer. In addition, accurate phenotype classification fosters timely diagnosis and proper treatment [2].

The primary challenge associated with phenotype classification is feature selection due to the inherent high-dimensionality feature space of DNA data. DNA feature space contains 3 billion base-pairs and upward of 25 thousand genes [3]. Identifying

features for classification in such an exorbitant space is challenging given only fractions of a genome sequence, or specific genes, are relevant to the phenotype [4]. Additionally, the number of observations available for analyses is significantly disproportional in comparison to the size of the feature space, which creates an underdetermined system. An underdetermined system is a system in which there are more features than observations. Parameter estimates on features in underdetermined system cannot be well calculated [5]. For this reason, feature selection and classification accuracy suffer.

1.2 Research Objective

The goal of my research was to develop and demonstrate methodology that addresses the phenotype classification dimensionality challenges; and to evaluate the features identified by the proposed methodology as biomarkers for disease. The novel methodology proposed in this dissertation is primarily an application of compressive sampling theory to DNA data. I show in this dissertation that the methodology based on compressive sampling and K-means clustering for feature evaluation facilitates efficient dimensionality reduction and feature selection for phenotype classification. Ultimately, I seek to reduce the dimensionality problem; to improve classification accuracy; and to show this methodology as a means to identify biomarkers for a phenotype. Lastly, I use two cancer datasets to demonstrate the proposed methodology.

1.3 Research Focus

This dissertation is comprised of an introductory section, followed by five additional chapters. Chapter 2 focuses on phenotype data representation and the principles of compressive sampling including: L_1 minimization, sparsity, incoherence, and the restricted isometry property (RIP). In chapter 3, I present the proposed novel methodology of compressive sampling for dimensionality reduction and pheno-

type classification, which is called correlation-informed incoherent framing. Lastly, I present the proof for calculating the probability of the RIP (p-RIP), with simulation results to substantiate the proof. In Chapter 4, I present the analysis and results of applying the methodology to cancer data. In Chapter 5, I present the clustering analysis for biomarker evaluation of the features identified by the proposed methodology. Finally, in Chapter 6, I conclude with a synopsis of the research, recommendations, and I suggest areas for future work.

II. Background

In this section I describe gene expression data and I provide a background on phenotype classification; the fundamental properties of compressive sampling; and K-means clustering.

2.1 Gene Data

Genomes contain the deoxyribonucleic acid (DNA) information of all living organisms. DNA strands are composed of the following chemicals (nucleotides): adenine, thymine, guanine, and cytosine. These base nucleotides are represented, respectively, as 'A', 'T', 'G', and 'C' in DNA sequences [6]. DNA can be visualized as two strands in a double helix configuration. The strands are opposites of each other according to its chemical bind; 'A' binds to 'T' and 'C' binds to 'G' [6]. Therefore, the arrangement of one strand reveals the arrangement of the other. As a result, the binding of a pair of nucleotides is referred to as a "base-pair" (bp).

DNA sequencing is the process to determine the bp arrangement in a genome strand. In 1990, the Human Genome Project (HGP) was commissioned by the United States Department of Energy and the National Institute of Health (NIH) to sequence the human genome [3]. The HGP completed this task in 2003 and revealed that the human genome is comprised of approximately 3 billion bps [6]. The evolution of genome sequencing technology and methods can be categorized into two generations: 1) 1st Generation Sanger Sequencing and DNA Microarrays, and 2) Next Generation Sequencing (NGS). First-generation sequencing revolves around the polymerase chain reaction (PCR) method. PCR process exploits the dichotomy of bps and synthesizes fragments of DNA strands from a chemical reaction in the presence of known nucleotides introduced to the process. PCR is a gradual process and faces

challenges when attempting to sequence an entire genome [7]. The NGS process invokes the same chemical reaction as the PCR process, but it replicates fragments of the genome by labeling the known nucleotides with a chemical label that produces a fluorescent signature when excited [8]. Those signatures transcribe to bp assignment. Through this method, NGS expedites the sequencing process to the effect of processing millions of fragments simultaneously. Consequently, millions of genome replications are produced in mass. This abundance of replications allows for more cross-referencing over the segments of genome. This, in-turn, increases the fidelity of the transcriptions which ultimately increases DNA sequence accuracy [9].

Illumina, a biotechnology company, reports the accuracy of the Sanger process at 99.4% while NGS yields a sequence accuracy of 99.9% [10]. Therefore, NGS sequences are preferable candidates for research and statistical analysis. I elected to use NGS sequence data for this research due its replication process and its accuracy advantage over Sanger sequence data.

2.2 Phenotype Classification

There are primarily two approaches to phenotype classification; alignment-based and alignment-free classification. Alignment-based classification relies on inter-tumor heterogeneity, which is the genetic and phenotypic variation observed between samples with similar tumors. It assumes tumors exhibit intra-tumor homogeneity, which refers to the biological variation within tumors [11]. However, tumors are not necessarily intra-tumor homogeneous, even with malignant tumors from the same organ [12]. Additionally, alignment-based classification depends on an extensive reference database which requires an exorbitant amount of computation time [13]. Because it is comparison in nature, alignment-based classification requires a repository of reference sequences and "sufficient" coverage. Coverage refers to the number of reads

used across a specific section of the genome in order to substantiate the transcription or to confirm the variation between reads [14]. Equation 1 is the coverage depth calculation,

$$C = L \cdot N/G \tag{1}$$

where 'C' is coverage depth; 'G' is the genome length; 'L' is read length; 'N' is the number of reads [14]. Consequently, alignment-based methods require more time and processing [14]. Sequence-based classification is probabilistic in nature and is faster to implement [13]. Alignment-free methods are single-cell approaches that rely on feature selection for classification, opposed to variation detection [11]. Single cell methods are ideal for genomes and transcriptomes [15]. The transcriptome is the collection of RNA in a cell, which provides insight to gene functionality of the cell and cell properties [16]. Macaulay et al. found that a collection of single cell genomes or transcriptomes can be used to show the correlation between genomic variation and phenotype; specifically in cancer cells [17]. Exploring cells individually allows for the quantification and, potentially, characterization of gene-specific heterogeneity for a particular condition [18]. Instead of phenotyping and drawing statistical inference from bulk data, alignment-free methods can be used with single-cell measurements; like DNA and RNA sequences or gene expression data [17, 19].

The national Cancer Institute keeps a ledger of 200 cancer types, though the number of subtypes is unknown. In a recent study, Song et al. found 10 molecular subtypes from an analysis of 10,000 malignant breast tumors [11]. A comparable study conducted by Sørlie et al., classified 456 breast cancer complementary DNAs (cDNAs), which are DNA clones, into five tumor subtypes using gene expression data. These results suggest that gene expression data can be used to classify tumors based on molecular composition [20, 21]. Conceptually, these findings implicitly advocate that with appropriate representation, single-cell data can be used for cancer phenotype

classification absent a reference genome.

2.3 Compressive Sampling

Compressive sampling is a sampling theory which maintains that sparse high dimensional signals can be recovered using a sufficient frame of vectors (basis/dictionary). Notationally, matrices are in bold and vectors are presented with a superscript arrow.

Definition 2.3.1 A **basis** for a vector space $\mathbf{V}$ is a set of linearly independent vectors that span $\mathbf{V}$ [22].

If signal $\vec{\beta}$ is s-sparse, in which it has at most s nonzero vectors: $\|\vec{\beta}\|_0 \leq s$, then it can be exactly recovered by a measurement $\vec{Y} \in \mathbb{R}^m$ taken by sensing matrix $\mathbf{X} \in \mathbb{R}^{m \times N}$, where $m \ll N$ [23, 24, 25]. More concisely stated, $\mathbf{X} : \mathbb{R}^N \mapsto \mathbb{R}^n$, and

$$\vec{Y} = \mathbf{X}\vec{\beta} \tag{2}$$

where, $\vec{\beta}$ has s nonzero elements, $\vec{\beta}$, and $\mathbf{X}$ adheres to the restricted isometry property (RIP).

Restricted Isometry Property.

The fundamental principle of compressive sampling is the restricted isometry property (RIP). It states that for a sensing matrix, $\mathbf{X}$, the RIP guarantees recovery of a sparse signal using L_1 minimization [23, 26]. RIP is formally defined as follows:

Definition 2.3.2 The RIP for a given sparsity, s, and isometry constant $\delta_s, \in (0, 1)$ of matrix $\mathbf{X} \in \mathbb{R}^{m \mathbf{x} N}$, is the smallest value such that:

$$(1-\delta_s)\|\vec{\beta}\|_2^2 \leq \|\mathbf{X}\vec{\beta}\|_2^2 \leq (1+\delta_s)\|\vec{\beta}\|_2^2 \tag{3}$$

is satisfied for every s-sparse vector $\vec{\beta} \in \mathbb{R}^N$ [23].

Clearly, if the RIP condition holds true for $\delta_s = 0$, $\mathbf{X}$ exactly preserves $\vec{\beta}$. If the RIP conditions hold true for $\delta_s < 1$, $\mathbf{X}$ is said to satisfy RIP of order s. Simply stated, s vectors in the column space of $\mathbf{X}$ are sufficient to recovery $\vec{\beta}$ [23]. Verifying all s combinations of columns comply with the RIP is NP-Hard [27]. Equation 4 is presented as a condition on the number of observations, m, to ensure recovery with high probability if $\delta_s \leq (s-1)\mu(\mathbf{X})$, where $\mu(\mathbf{X})$ is defined later as a coherence calculation on a matrix $\mathbf{X}$.

$$m \geq 2s \log\left(\frac{N}{s}\right) \quad [25,\ 28], \tag{4}$$

Equation 5 is presented as the sparsity condition on an $m \times n$ random matrix, for high probability to adhere to the RIP is given as

$$s \approx \frac{m}{\log^k n} \ for\ k \geq 1 \quad [25,\ 28]. \tag{5}$$

Both Equations 4 and 5 certify RIP with high probability based on the s and m. Later, I present a proof to calculate probability of RIP (p-RIP) if these conditions are not met. Adherence to the RIP is not a requirement for signal recovery, but, recovery is guaranteed if RIP is satisfied. The RIP is important because it quantifies the extent that $\mathbf{X}$ changes $\vec{\beta}$. In pursuit of identifying features for a phenotype, the RIP provides a means to evaluate the quality of the features remaining after dimensionality reduction. It provides a lower bound to the size of the reduction in order to guarantee recovery. I posit that the guarantee of recovery positions those

features as potential biomarkers for the phenotype.

L_1 minimization.

Signal recovery aims to find the signal's sparse coefficient vector. L_0 minimization can recover a s-sparse signal exactly with high probability using $M = s + 1$ measurements, however, the zero-norm,$\|\cdot\|_0$, is nonconvex and approximating the minimum is NP-complete in computational time [24]. The L_1 minimization problem is a convex optimization problem and can be solved via linear programming methods and in linear computational time [24]. Therefore, the one-norm can be used to approximate the zero-norm solution. The p-norm unit ball for $p \in \{\frac{1}{2}, 1, 2, 11\}$, Figures 1-4, are presented here to demonstrate this idea. Figure 1 illustrates the unit ball for $p = \frac{1}{2}$, but serves as a visual of the p-norm unit ball for $p < 1$. As p approaches 0, for $p < 1$, the p-norm unit ball converges to the origin. If L is an affine line in two-dimensional space the p-norm unit ball for $p < 1$ intersects the L at point t. Figure 2 shows that the one-norm unit ball intersects L at the same location where the zero-norm unit ball would intersect L. Figures 3 and 4 show that as p increases for $p > 1$ the intersection between the unit ball and L occurs elsewhere.

Therefore, the L_1 norm recovers the same solution for $p < 1$. This phenomena can be extended to higher (finite) dimensionality without loss of generality and is used for regularization in compressive sampling [24]. In the context of compressive sampling, L_1 minimization ensures that a sparse signal of length N can be recovered exactly. The original sparse recovery problem is given as [24]:

$$\begin{aligned} &\min \|\vec{\beta}\|_0 \\ &\text{subject to} \\ &\mathbf{X}\vec{\beta} = \vec{Y}. \end{aligned} \tag{6}$$

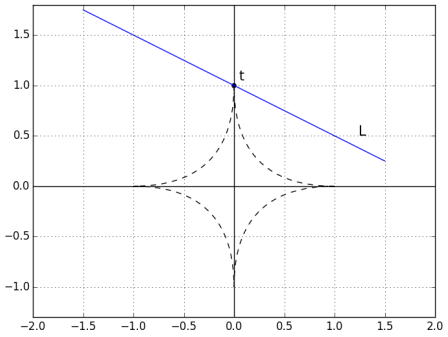


Figure 1. $\frac{1}{2}$-Norm Unit Ball.

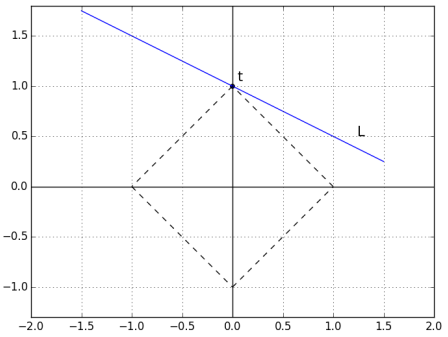


Figure 2. 1-norm Unit Ball.

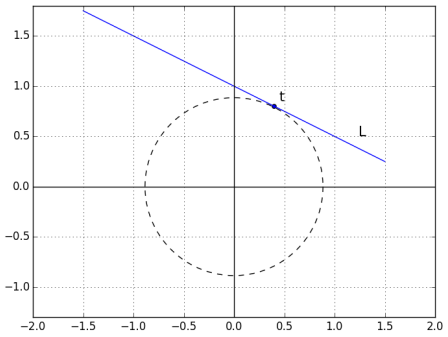


Figure 3. 2-Norm Unit Ball.

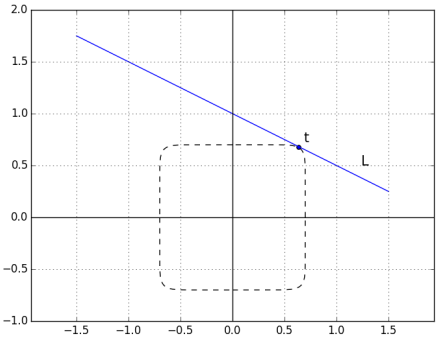


Figure 4. 11-Norm Unit Ball

The L_0 solution to the recovery problem can be approximated with the L_1 solution to the recovery problem [24]:

$$\begin{aligned} &\min \|\vec{\beta}\|_1 \\ &\text{subject to} \\ &\mathbf{X}\vec{\beta} = \vec{Y}. \end{aligned} \tag{7}$$

The signal recovery problem from noisy data, with bound error, becomes [24]:

$$\begin{aligned} &\min \|\vec{\beta}\|_1 \\ &\text{subject to} \\ &\|\mathbf{X}\vec{\beta} - \vec{Y}\|_2 \leq \epsilon. \end{aligned} \tag{8}$$

The regularized version of the bounded error problem becomes [29]:

$$\min \|\mathbf{X}\vec{\beta} - \vec{Y}\|_2 + \alpha\|\vec{\beta}\|_1, \tag{9}$$

where α is the regularization parameter.

Sparsity.

Conceptually, sparsity suggests that a signal can be expressed, more concisely, by a linear combination of a sufficient basis. This principle can be better realized when considering image compression. An image can be recreated, without unrecognizable distortion, from preserving pixels with "large" wavelet coefficients [23]. The recognizable image is captured with the large wavelet coefficients, and the smaller, less significant, coefficients can be "gained" to zero; which creates sparse representation of the original signal. Consequently, sparsity enables signal recovery. This research leverages the phenomena of sparse signal recovery; posits genome sequences can be regarded as a signal; and pursues the sufficient basis of features for phenotype data.

Incoherence.

Coherence assesses the quality of a measurement matrix by measuring for the largest correlation between all the columns of the matrix; coherence speaks to the linear dependence between the columns of the matrix. For this dissertation, incoherence declares pairwise independence amongst all columns; or that an acceptable measure of coherence has been met between all columns. Two coherence calculations along with a coherence criterion are given below in order to demonstrate this principle:

Definition 2.3.3 A matrix $\mathbf{A}$ with normalized columns, $\|a_i\|_2 = 1$ and $i, j \in N$, the

worst-case coherence $\mu(\mathbf{A})$ is:

$$\mu(\mathbf{A}) := \max_{i,j:i\neq j} |\langle a_i, a_j\rangle| \quad [30,\ 31]. \tag{10}$$

Definition 2.3.4 A matrix $\mathbf{A}$ with normalized columns, $\|a_i\|_2 = 1$ and $i, j \in N$, the **average coherence** $\nu(\mathbf{A})$ is:

$$\nu(\mathbf{A}) := \frac{1}{n-1} \max_i \Big| \sum_{j:j\neq i} \langle a_i, a_j\rangle \Big| \quad [30,\ 31]. \tag{11}$$

Definition 2.3.5 An $m \times n$ unit norm frame satisfies the **Strong Coherence Property** if

$$\mu(\mathbf{A}) \leq \frac{1}{164 \log n} \text{ and } \nu(\mathbf{A}) \leq \frac{\mu(\mathbf{A})}{\sqrt{m}} \quad [30,\ 31]. \tag{12}$$

Columns with coherence close to zero, $\mu(\mathbf{A}) \approx 0\ \forall a_i, a_j,\ i \neq j$, implies the columns are near orthogonal [25]. A set of vectors with zero coherence represents a completely mutually orthogonal set. Orthogonality is vital to the idea of sparse recovery for two specific reasons. First, the idea of recovery implies that the mapping operator is invertible. Since linear independence is a requirement for a matrix to be invertible, the assumption on the sensing matrix is that it conforms to independence. In the absence of a completely mutually orthogonal set, a coherence criteria (Equation 12) can be used to assess independence or to preserve the assumption of independence. The second reason is essentially the fact that a sparse solution can be achieved for recovery. This implies that the solution set of vectors is a basis for $\mathbb{R}^N$. For these reasons, incoherence of the sensing matrix is a critical component to compressive sampling. Collectively, the principles associated with compressive sampling make compressive sampling appealing as a methodology for dimensionality reduction in a high dimensional sample space.

2.4 K-means

K-means clustering is a centroid-based method that grants membership into a cluster or group based on proximity to a center or convex hull of the cluster [5]. It is initiated by randomly assigning the data into a predetermined number of clusters, K. Centroids for each cluster is calculated. Reassignment of the data is then conducted based on proximity to the calculated centroids. This process is repeated until a stopping criterion is met or the centroids (clusters) have no longer changed; or the changed is smaller than a predetermined threshold. K-means clustering is a method that can be used to determine the number of clusters that may exist in a dataset.

2.5 Levenshtein Distance

Levenshtein distance (LD), also referred to as edit distance, is a methodology used as a distance calculation between words or strings of letters. It is used for spell checking, language classification, and other similar text mining applications. Its algorithm is based on three equally weighted operations: insertion, deletion, and substitution [32]. Essentially, LD is the minimum cost to transform one string to match another. Consider the following strings in order to demonstrate LD operations: g_1 - tcaa, g_2 - tcaga, g_3 - tcag, g_4 - caa. Table 1 shows the LD for several operations, to include combining operations.

Table 1. Levenshtein Distance Demonstration

Operation	**Transformation**	**LD**
Insertion: **g**	$g_2 \rightarrow g_1$	1
Substitution: **g**-to-**a**	$g_3 \rightarrow g_1$	1
Deletion: **t**	$g_1 \rightarrow g_4$	1
Insertion/Substitution: **t** & **a**-to-**g**	$g_4 \rightarrow g_3$	2
Insertion: **t** & **g**	$g_4 \rightarrow g_2$	2

2.6 Summary

This research treats the phenotype classification problem as a signal recovery problem and employs compressive sampling principles to address the challenges associated with high dimensionality DNA data. CIF is an application of compressive sampling principles that uses features correlated with the phenotype to build an incoherent sensing matrix for the phenotype. Adherence to the RIP guarantees that the features of the sensing matrix can recovery the signal. This is important because the guarantee of recovery infers that the features of the sensing matrix can be regarded as a basis for the phenotype. The probability of RIP (p-RIP) will be introduced in Section 3 as means to determine the statistical likelihood of adherence to the RIP for a random sensing matrix. Features identified as bases will be evaluated as potential biomarkers using K-means clustering with LD distance as the measure of dissimilarity.

III. Methodology

In this section I introduce the CIF process. Additionally, I present the proof for probabilistically satisfying the RIP and the simulation results to support the proof.

3.1 Correlation-Informed Incoherence Framing

I now introduce the correlation-informed incoherent framing (CIF) methodology. The CIF is an approach to constructing sensing matrices, $\mathbf{X}$, that relies on the idea that random assignment and compliance to a coherence criteria produces a basis for the phenotypes. The objective for CIF is to produce a linearly independent $m \times n$, matrix where $m > n$. The framing process is geared toward constructing this matrix from a set of candidate vectors (features), $\{\vec{v}_i\}_{i=1}^{N}$. It is initiated from an empty set, then a single candidate vector from $\{\vec{v}_i\}_{i=1}^{N}$ is randomly selected for entry. The unit-norm of that vector is then entered into the empty set. Of note, the unit norms are used in the framing process because worst-case coherence, Equation 10, is used as the entry criteria. Next, another candidate vector is randomly selected for entry consideration. Worst-case coherence between the new set with the candidate vector is then calculated and compared to the entry criteria. If this set meets the criteria, the candidate vector gains membership. If the set fails to meet the criteria, the candidate vector is dismissed from the new set and removed from further consideration. This process is repeated until the candidate set is completely exhausted or a stopping criterion is reached. Using worst-case coherence as a single criteria grants membership to features with potentially little effect to the response vector. Therefore, in order to construct an incoherent frame with comparable classification accuracy, correlation coefficients were used to subset the parent matrix according to the relationship between features and the response. The correlation coefficients between each $\{\vec{v}_i\}_{i=1}^{N}$

and the response vector were calculated and was used as a filter to the inform the candidacy process. The vectors with the largest correlation coefficient magnitudes, were selected as the candidate set. The number of vectors to use as the candidate set was arbitrary, however, the size of the candidate set was selected to maintain the integrity of the random process. For instance if the $N = 700,000$, taking the largest 1% of the correlation coefficients results in a candidate set of $7,000$. Where as, if $N = 7,000$, taking the largest 1% of the correlation coefficients results in a candidate set of 70. Obviously, a set of $7,000$ vectors opposed to 70 vectors provides for more possible combinations when choosing s vectors at a time. Once the correlation-informed candidates set is generated, the incoherence framing process routine is conducted on that set. Algorithm 1 describes the CIF process.

Algorithm 1 Correlation-Informed Incoherence Framing

Input: Measurement Matrix, $\mathbf{X}$, categorical response vector, $\vec{Y}$.
Output: Incoherent Matrix, $\tilde{\mathbf{X}}$
$Ncol \leftarrow$ length of columns in $\mathbf{X}$
for $i = 1, \cdots, Ncol$ **do**
    ρ_i between X_i and Y
Initialize: Set $\mathbf{V} := X_i : |\rho_i| \in$ top 1%
Set Coherence Criteria: $crit$
Randomly pull a vector from $\mathbf{V}$: v_k
$\tilde{X} \leftarrow v_k / \|v_k\|_2$
Remove v_k from $\mathbf{V}$
$num =$ length $\mathbf{V}$
for $j = 1 \cdots num$ **do**
    Randomly pull a vector from $\mathbf{V}$: v_k
    $\tilde{v}_k = v_k / \|v_k\|_2$
    $\mu(X) := \max_i |\langle \tilde{x}_i, \tilde{v}_k \rangle|$
    if $\mu(X) < crit$ **then**
        $\tilde{X} \leftarrow \tilde{v}_k$
    Remove v_k from $\mathbf{V}$

After the incoherent sensing matrix is constructed, L_1 regularization logistic re-

gression is conducted for optimization of the parameter estimates and to evaluate classification performance via classification accuracy. Classification accuracy is the percentage of correctly identified observations.

3.2 Probability of the Restricted Isometry Property

Earlier, Equations 4 and 5 were presented as conditions on m and s of a sensing matrix for high probability of adherence to the RIP. Later, I present a proof for calculating probability of the RIP (p-RIP) for a random sensing matrix, if those conditions are not satisfied. Tables 2 and 3 were created using Equation 4 in order to show the number of observations required for five sparsities and corresponding strong coherence based on $N = 690,553$ and $N = 6,906$, and 107 observations. These values of N were chosen because they correspond to the breast cancer dataset and its CIF candidate set presented later. Table 2 indicates that for $s = 18$, $m \geq 165$ is required for high probability of RIP, with a corresponding $\mu(\mathbf{X}) \leq 0.005$. The 107 observations in the breast cancer dataset do not meet the required m, therefore, adherence to RIP cannot be stipulated to for this case. Table 3 similarly shows that for $N = 6,906$ an $s = 18$ requires an $m \geq 93$ with a $\mu(\mathbf{X}) \leq 0.005$ for high probability of RIP. This case meets the requirement on m, but, as shown later $\mu_c < 0.05$ produced sensing matrices with an average classification accuracy less than 90% for all $s < 20$. Of note, μ_c is used as the notation for a coherence criteria opposed to a coherence calculation. A $\mu_c = 0.05$ used to construct the sensing matrix produced the highest average classification accuracy for the breast cancer dataset with a sparsity of $s = 18$. Similarly, Tables 4 and 5 show the number of observations required for five sparsities and the corresponding strong coherence property based on a dataset with $N = 16,382$ and $N = 1,638$ respectively, and 65 observations. These values of N were chosen because they correspond to the small cell lung cancer dataset and its CIF candidate

set presented later. Tables 4 and 5 indicate that these data consist of a sufficient number of observations by sparsity for high probability of recovery. However, the corresponding $\mu(\mathbf{X})$ produced sensing matrices with an average classification accuracy less than 85% for all $s < 4$.

Table 2. RIP conditions based on the number of observations using Equation 4; considering all the column vectors.

N = 690,553	**s**				
	2	**10**	**18**	**20**	**25**
$m \geq$	22	96	165	181	222
$\mu(\mathbf{X})$	0.020	0.006	0.005	0.005	0.004

Table 3. RIP conditions based on the number of observations using Equation 4; considering top 1% correlation coefficient magnitudes.

n = 6,906	**s**				
	2	**10**	**18**	**20**	**25**
$m \geq$	14	56	93	101	122
$\mu(\mathbf{X})$	0.020	0.006	0.005	0.005	0.004

Table 4. RIP conditions based on the number of observations using Equation 4; considering all the column vectors.

N = 16,382	**s**				
	2	**3**	**4**	**5**	**6**
$m \geq$	18	26	33	40	48
$\mu(\mathbf{X})$	0.020	0.013	0.010	0.009	0.008

Table 5. RIP conditions based on the number of observations using Equation 4; considering top 10% correlation coefficient magnitudes.

N = 1,638	**s**				
	2	**3**	**4**	**5**	**6**
$m \geq$	14	20	25	30	36
$\mu(\mathbf{X})$	0.020	0.013	0.010	0.009	0.008

These observations provide insight to the trade space between s, $\mu(\mathbf{X})$, and classification accuracy for N; which is essentially the composition of the sensing matrix.

The probability of RIP (p-RIP) is now introduced to evaluate a sensing matrix when the composition does not adhere to the conditions required for high probability of RIP.

Theorem 3.2.1 Suppose $\mathbf{A} \in \mathbb{R}^{m \times N}$ is a random matrix. Applying CIF to $\mathbf{A}$, using worse-case coherence, transforms $\mathbf{A}$ into $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ where $n << N$ such that the probability of RIP (p-RIP) for $\tilde{\mathbf{A}}$ of order s is given as:

$$Pr\Big[(1-\delta_s) \leq \|\tilde{\mathbf{A}}\vec{\beta}\|_2^2 \leq (1+\delta_s)\Big]. \tag{13}$$

Proof. Take $\mathbf{A}$ as a random matrix with columns defined as $a_{*,j}$ for $j = 1, 2, \cdots, N$. Applying CIF to $\mathbf{A}$, using worse-case coherence, transforms $\mathbf{A}$ into $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ where $n << N$. The columns of $\tilde{\mathbf{A}}$ are defined as:

$$\tilde{a}_{*,j} = \frac{a_{*,j}}{\|a_{*,j}\|_2}, \;\; j = 1, \cdots, n \tag{14}$$

where,

$$\|a_{*,j}\|_2 = \Big(\sum_{i=1}^{n} a_{i,j}^2\Big)^{\frac{1}{2}}. \tag{15}$$

Let $\|\vec{\beta}\|_2^2 = 1$ to establish the most conservative bounds on $\|\tilde{\mathbf{A}}\vec{\beta}\|_2^2$. Then Equation 3 can now be written as:

$$(1-\delta_s) \leq \|\tilde{\mathbf{A}}\vec{\beta}\|_2^2 \leq (1+\delta_s). \tag{16}$$

Therefore, the probability that $\tilde{\mathbf{A}}$ satisfies the RIP condition of order s is:

$$Pr\Big[(1-\delta_s) \leq \|\tilde{\mathbf{A}}\vec{\beta}\|_2^2 \leq (1+\delta_s)\Big].$$

□

Furthermore, following s-sparse vector $\|\vec{\beta}\|_2^2 = 1$, the entries of $\vec{\beta}$ are $1/\sqrt{s}$. As a result, $\vec{\beta}$ scales the entries of the row vector of $\tilde{\mathbf{A}}$. Therefore, $\tilde{\mathbf{A}}\vec{\beta}$, is an s-sparse vector in which each entry represents an s-summed random variable from the rows of $\tilde{\mathbf{A}}$. In order to calculate the probability of satisfying RIP, Equation 13, the distribution of $\tilde{\mathbf{A}}$ and subsequently, $\tilde{\mathbf{A}}\vec{\beta}$ has to be determined. Next, I explore the distribution of the incoherent sensing matrix constructed from a Bernoulli and a Standard Normal random matrix.

Bernoulli Random Matrix.

Theorem 3.2.2 Suppose $\mathbf{X} \in \mathbb{R}^{m \times N}$ is a Bernoulli random matrix. If $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times n}$ is constructed from the CIF process, using worst-case coherence, then:

$$\log \|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 \sim \mathrm{N}\Big(0, \frac{n-x}{nx} + \frac{m-y}{my}\Big)$$

Proof. Take $\mathbf{X}$ as a Bernoulli random matrix such that the columns, $x_{*,j}$, have a probably of occurrence p_j, $j = 1, 2, \cdots, m$, $x_{*,j} \sim \mathrm{Ber}(p_j)$. Assuming $x_{*,1} \cong x_{*,2} \cong, \cdots, \cong x_{*,j}$, $j = 1, 2, \cdots, N$ and the columns are independent and identically distributed, then $\mathbf{X} \sim \mathrm{Ber}(p)$. Therefore, $\forall x_{ij} \in \mathbf{X} \sim \mathrm{Ber}(p)$. The columns of $\tilde{\mathbf{X}}$ are defined as:

$$\tilde{x}_{*,j} = \frac{x_{*,j}}{\|x_{*,j}\|_2}, \;\; j = 1, \cdots, n.$$

Let $V = \|x_{*,j}\|_2$ then,

$$\tilde{x}_{*,j} = \frac{x_{*,j}}{V}.$$

Vector $\tilde{\mathbf{X}}\vec{\beta}$ can be written as

$$\tilde{\mathbf{X}}\vec{\beta} = \sum_{j=1}^{n} \tilde{x}_{ij} b_j = \sum_{j=1}^{n} \frac{x_{i,j}}{V} b_j = \frac{\sum_{j=1}^{n} x_{i,j} b_j}{V},$$

where b_j is the j^{th} element of $\vec{\beta}$. Setting $\|\vec{\beta}\|_2^2 = 1$ implies $b_j = \frac{1}{\sqrt{s}}\ \forall j$. Since $\vec{\beta}$ is s-sparse, s of the entries are $\frac{1}{\sqrt{s}}$ and $n - s$ entries are zero. Therefore,

$$\sum_{j=1}^{n} x_{ij} b_j = \frac{1}{\sqrt{s}} \sum_{j=1}^{s} x_{i,j}$$

and

$$\tilde{\mathbf{X}}\vec{\beta} = \frac{1}{\sqrt{s}} \frac{\sum_{j=1}^{s} x_{i,j}}{V}.$$

Let $W_i = \sum_{j=1}^{s} x_{i,j}$, and $W_1 \cong W_2 \cong, \cdots, \cong W_i$, $i = 1, 2, \cdots, m$ then,

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 = \sum_{i=1}^{m} \Big(\frac{1}{\sqrt{s}} \frac{W_i}{V}\Big)^2 = \frac{m}{s} \frac{W^2}{V^2}.$$

Therefore, $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ is a ratio between two binomial (bin) random variables: $W^2 \sim \text{bin}(s, p)$ and $V^2 \sim \text{bin}(m, p)$. As a results,

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 = \frac{m}{s} \frac{W^2}{V^2} \equiv \frac{\text{bin}(s, p)/s}{\text{bin}(m, p)/m}.$$

In 1978, Katz et al. found that the natural logarithm of the ratio of two binomial random variables divided by their corresponding number of trials is distributed Normal [33]. For example, take

$$Z = \frac{U/n}{Q/r},$$

where $U \sim \text{bin}(n, p_1)$ and $Q \sim \text{bin}(r, p_2)$, then,

$$\log(Z) \sim N\Big(\log \frac{p_1}{p_2}, \frac{n-u}{nu} + \frac{r-q}{rq}\Big) \text{ [33].}$$

Consequently,

$$\log \|\mathbf{X}\vec{\beta}\|_2^2 = \log\left(\frac{m}{s}\frac{W^2}{V^2}\right) \sim \mathrm{N}\Big(\log\frac{p}{p}, \frac{s-w}{sw} + \frac{m-v}{mv}\Big).$$

Therefore,

$$\log \|\mathbf{X}\vec{\beta}\|_2^2 \sim \mathrm{N}\Big(0, \frac{s-w}{sw} + \frac{m-v}{mv}\Big). \tag{17}$$

□

Gaussian Random Matrix.

Theorem 3.2.3 Suppose $\mathbf{X} \in \mathbb{R}^{m\times N}$ is a Standard Normal, $\mathrm{N}(0,1)$, random matrix. If $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times n}$ is constructed from the CIF process, using worst-case coherence, then:

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 \sim F_{m,m}.$$

Proof. Take $\mathbf{X}$ as a Standard Normal, $\mathrm{N}(0,1)$, random matrix. Then, $\forall x_{ij} \in \mathbf{X} \sim \mathrm{N}(0,1)$. Let $V = \|x_{*,j}\|_2$, which is a Chi distribution with m degrees of freedom: $V \sim \chi_m$, then,

$$\tilde{x}_{i,j} = \frac{x_{i,j}}{V}.$$

Vector $\tilde{\mathbf{X}}\vec{\beta}$ can be written as

$$\tilde{\mathbf{X}}\vec{\beta} = \sum_{j=1}^{n} \tilde{x}_{ij} b_j = \sum_{j=1}^{n} \frac{x_{i,j}}{V} b_j = \frac{\sum_{j=1}^{n} x_{i,j} b_j}{V}.$$

Let $(\tilde{xb})_i$ be define as the elements of $\tilde{\mathbf{X}}\vec{\beta}$, such that

$$(\tilde{xb})_i = \sum_{j=1}^{n} \tilde{x}_{i,j} b_j = \sum_{j=1}^{n} \frac{x_{i,j}}{V} b_j.$$

where b_j is the j^{th} element of $\vec{\beta}$. Setting $\|\vec{\beta}\|_2^2 = 1$ implies $b_j = \frac{1}{\sqrt{s}}\ \forall j$. Since $\vec{\beta}$ is s-sparse, s of the entries are $\frac{1}{\sqrt{s}}$ and $n - s$ entries are zero. Therefore,

$$(\tilde{xb})_i = \sum_{j=1}^{s} \frac{1}{\sqrt{s}} \frac{x_{i,j}}{V} = \frac{1}{\sqrt{s}} \frac{\sum_{j=1}^{s} x_{i,j}}{V}$$

where $\sum_{j=1}^{s} x_{i,j} \sim \mathrm{N}(0, s)$. Let $W_i = \frac{\sum_{j=1}^{s} x_{i,j}}{\sqrt{s}} \sim \mathrm{N}(0, 1)$, then

$$(\tilde{xb})_i = \frac{W_i}{V.}$$

Now, $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ can be written as

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 = \sum_{i=1}^{m} \left(\frac{W_i}{V}\right)^2 = \frac{\sum_{i=1}^{m} W_i^2}{V^2}$$

Note, that $V^2 \sim \chi_m^2$ and $W_i \sim \mathrm{N}(0, 1), \forall i$, which makes $\sum_{i=1}^{m} W_i^2 \sim \chi_m^2$, therefore,

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 = \frac{\sum_{i=1}^{m} W_i^2}{V^2} = \frac{\sum_{i=1}^{m} W_i^2/m}{V^2/m}.$$

Moreover,

$$\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2 \sim F_{m,m}. \tag{18}$$

□

3.3 p-RIP Simulation

Bernoulli Random Matrix.

The composition of a Bernoulli random matrix is a sparse binary representation in itself, because its elements are zero or one. This necessitated a need to simulate multiple p settings to validate the Theorem 3.2.2. In the case of Bernoulli random matrices, p regulates the extent of the sparsity of the matrix; how many elements are zero or one. Varying p allows for a more thorough investigation into Theorem 3.2.2 and the relationship between p-RIP and matrix composition. Theorem 3.2.2 hypothesizes that the framing process transforms a Bernoulli random matrix into a Normal random matrix. The simulation for the Bernoulli random matrix considered multiple settings for $p \in \{0.004, 0.006, 0.008, 0.010\}$ and $s \in \{4, 10, 16, 18\}$. The simulation generated $\tilde{\mathbf{X}}$ matrices for each values of p and $\vec{\beta}$ vectors of length 50 at each value of s. The dimensions of the random matrix were held constant with 107 rows and 50 columns. The $\mathbf{X}$ matrix was constructed by columns of size 107, using a Bernoulli random number generator to create its elements. Next, the $\vec{\beta}$ vector was generated randomly assigning the value of $\frac{1}{\sqrt{s}}$ for s of its 50 entries. Each combination of p and s was bootstrapped 500 times for constructing $\mathbf{X}$ and $\vec{\beta}$. Then $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ was calculated, which generated a random sample of size 500. This process was simulated 2,000 for each combination.

Figures 24-28 in Appendix A are the histograms from one draw for each combination of p and s. These figures show that as s increases for each p, the samples appear to converge toward a Normal distribution. Figures 30-34 in Appendix B are the quantile plots for the same draws. These plots agree that as s increases, the draws appear to converge toward a Normal distribution for each p. Additionally, p appears to affect the rate at which the draws converge. Each sample was tested for goodness-of-fit, to a Normal distribution, using the Kolmogorov-Smirnov test. This

resulted in 2,000 goodness-of-fit tests for each combination of p and s. Table 6 shows the percentages out of 2,000 draws that did not lack fit to a Normal distribution. Using a level of significance of 0.05, 95% of the draws should not lack fit in order to conclude in favor of the data being Normally distributed.

Table 6. Percentage of 2,000 trials for which the goodness-of-fit tests showed non-lack of fit to a Normal Distribution.

	s			
p	4	10	16	18
0.004	0.0	0.0	0.0	0.12
0.006	0.0	0.0	0.48	0.92
0.007	0.0	0.0	0.88	0.99
0.008	0.0	0.0	0.98	1.0
0.010	0.0	0.0	1.0	1.0

The results in Table 6 agree with the qualitative assessment from above which suggested that as p or s increases, simulated data becomes more Normally distributed. As depicted, for $s = 16$ and $p \geq 0.008$, and for $s \geq 18$ and $p \geq 0.007$ the simulated data reasonably follows a Normal distribution. The concern with this relationship is how rapidly does the decrease in overall sparsity of $\mathbf{X}$ translate into a Normally distributed $\tilde{\mathbf{X}}$. Showing that Theorem 3.2.2 holds for certain cases is required in order to leverage Theorem 3.2.1 to calculate p-RIP.

Tables 7-11 reflect the results of using the simulated data to show the p-RIP according to Theorem 3.2.2 for $p \in \{0.004, 0.006, 0.008, 0.010\}$ and for $s \in \{4, 10, 16, 18\}$. Tables 7-11 show that, in general, as p increases the chances of p-RIP slightly decreases. This is likely due to the effect p has on composition of $\mathbf{X}$. As mentioned previously, p regulates the sparsity of $\mathbf{X}$. As p increases, $\mathbf{X}$ is less of a sparse-binary representation. Tables 7-11 show that for $s = 16, 18$ a sparse solution is obtainable at high p-RIP. Sufficient signal recovery requires a high probability for $\delta_s < 0.33$ [34]. This provides evidence of reasonable expectancy that CIF can construct a sensing

matrix, from a Bernoulli random matrix, that can achieve p-RIP.

Table 7. Simulated p-RIP for a Ber(0.004) Random Matrix

			δ_s		
sparsity	0.0	0.1	0.2	0.3	0.4
4	0.387 ± 0.009	0.763 ± 0.032	0.947 ± 0.019	0.992 ± 0.005	0.999 ± 0.001
10	0.208 ± 0.01	0.523 ± 0.019	0.802 ± 0.02	0.942 ± 0.012	0.987 ± 0.004
16	0.086 ± 0.008	0.313 ± 0.015	0.624 ± 0.02	0.853 ± 0.016	0.958 ± 0.008
18	0.06 ± 0.007	0.252 ± 0.013	0.555 ± 0.019	0.811 ± 0.017	0.941 ± 0.009
			δ_s		
sparsity	0.5	0.6	0.7	0.8	0.9
4	0.999 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
10	0.987 ± 0.004	0.998 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
16	0.958 ± 0.008	0.991 ± 0.003	0.998 ± 0.001	1.0 ± 0.0	1.0 ± 0.0
18	0.941 ± 0.009	0.986 ± 0.004	0.997 ± 0.001	1.0 ± 0.0	1.0 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

Gaussian Random Matrix.

Theorem 3.2.3 demonstrates that the framing process transforms a Standard Normal random matrix into an F random matrix. The Gaussian simulation was demonstrated using the following sparsities: $s \in \{2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18\}$. The dimensions of the random matrix were held constant with 65 rows and 50 columns. The $\mathbf{X}$ matrix constructed by columns of size 65, using a Standard Normal random number generator to create its elements. The $\vec{\beta}$ vector was generated, randomly assigning the value of $\frac{1}{\sqrt{s}}$ for s of its 50 entries. $\mathbf{X}$ and $\vec{\beta}$ was bootstrapped 500 times

Table 8. Simulated p-RIP for a Ber(p = 0.006) Random Matrix

			δ_s		
sparsity	0.0	0.1	0.2	0.3	0.4
4	0.376 ± 0.009	0.751 ± 0.029	0.942 ± 0.018	0.991 ± 0.005	0.999 ± 0.001
10	0.187 ± 0.01	0.493 ± 0.018	0.782 ± 0.02	0.934 ± 0.012	0.985 ± 0.004
16	0.068 ± 0.007	0.272 ± 0.014	0.579 ± 0.019	0.826 ± 0.017	0.948 ± 0.009
18	0.045 ± 0.006	0.211 ± 0.013	0.503 ± 0.018	0.775 ± 0.018	0.926 ± 0.011

			δ_s		
sparsity	0.5	0.6	0.7	0.8	0.9
4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
10	0.997 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
16	0.988 ± 0.003	0.998 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
18	0.982 ± 0.004	0.996 ± 0.001	0.999 ± 0.001	1.0 ± 0.0	1.0 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

for every s. Then $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ was calculated, which generated a random sample of size 500. This process was simulated 2,000 for each s.

The graphs in Figure 16 in Appendix A are the histograms from one draw for each s of the process. The proof for Theorem 3.2.3 indicates that the Standard Normal transformation to an F distribution which relies on m. Therefore, the draws generated from different s settings should not affect the transformation. Of note, the curve of an F distribution with equal degrees of freedom, $\nu_1 = \nu_2$, resembles a right-skewed Normal distribution curve. The histograms in Figure 16 provide visual evidence to support both of these premises. The graphs in Figure 23 in Appendix B are the quantile plots for the same draws. These plots agree that the draws appear to be

Table 9. Simulated p-RIP for a Ber(0.007) Random Matrix

			δ_s		
sparsity	0.0	0.1	0.2	0.3	0.4
4	0.371 ± 0.009	0.745 ± 0.03	0.939 ± 0.019	0.99 ± 0.005	0.999 ± 0.001
10	0.176 ± 0.01	0.477 ± 0.018	0.77 ± 0.02	0.929 ± 0.012	0.984 ± 0.005
16	0.06 ± 0.007	0.251 ± 0.013	0.554 ± 0.019	0.811 ± 0.017	0.941 ± 0.01
18	0.038 ± 0.005	0.189 ± 0.012	0.475 ± 0.018	0.755 ± 0.018	0.917 ± 0.012

			δ_s		
sparsity	0.5	0.6	0.7	0.8	0.9
4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
10	0.997 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
16	0.986 ± 0.004	0.997 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
18	0.979 ± 0.005	0.996 ± 0.001	0.999 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

$F_{65,65}$ distributed. Samples for $s = 2, 3, 4, 5$ were tested for goodness-of-fit to an $F_{65,65}$ distribution using the Kolmogorov-Smirnov test. These s settings were tested because they correspond to the small cell cancer data set presented later. This resulted in 2,000 goodness-of-fit tests for each $s = 2, 3, 4, 5$. Table 12 shows the percentages, out of 2,000 draws, that did not lack fit to an $F_{65,65}$ distribution. Using an level of significance of 0.05, 95% of the draws should not lack fit in order to conclude in favor of the data to be $F_{65,65}$ distributed. The results in Table 12 agree with the qualitative assessment for $s = 4, 5$, however, the data for $s = 2, 3$ does lack fit an $F_{65,65}$ distribution. The random process appears to have an effect from s on the transformation. Showing that Theorem 3.2.3 holds, for some values of s, is required

Table 10. Simulated p-RIP for a Ber(0.008) Random Matrix

			δ_s		
sparsity	0.0	0.1	0.2	0.3	0.4
4	0.365 ± 0.009	0.738 ± 0.029	0.936 ± 0.019	0.99 ± 0.006	0.999 ± 0.001
10	0.166 ± 0.01	0.461 ± 0.017	0.757 ± 0.02	0.923 ± 0.013	0.982 ± 0.005
16	0.052 ± 0.006	0.23 ± 0.013	0.529 ± 0.018	0.793 ± 0.017	0.934 ± 0.01
18	0.032 ± 0.005	0.17 ± 0.012	0.447 ± 0.018	0.733 ± 0.019	0.906 ± 0.012

			δ_s		
sparsity	0.5	0.6	0.7	0.8	0.9
4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
10	0.997 ± 0.001	1.0 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
16	0.984 ± 0.004	0.997 ± 0.001	1.0 ± 0.001	1.0 ± 0.0	1.0 ± 0.0
18	0.975 ± 0.005	0.995 ± 0.002	0.999 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

in order to leverage Theorem 3.2.1 to calculate p-RIP.

Table 13 reflects the results of using the simulated data to show the probability of satisfying the RIP according to Theorem 3.2.3 for $m = 65$. Again, sufficient signal recovery requires a high probability for $\delta_s < 0.33$ [34]. Table 13 shows a probability greater than 78% of achieving sufficient recovery. This provides evidence of reasonable expectancy that CIF can construct a sensing matrix from a Standard Normal random matrix that can achieve p-RIP.

Table 11. Simulated p-RIP for a Ber(0.010) Random Matrix

			δ_s		
sparsity	0.0	0.1	0.2	0.3	0.4
4	0.353 ± 0.009	0.726 ± 0.028	0.93 ± 0.019	0.988 ± 0.006	0.999 ± 0.001
10	0.145 ± 0.009	0.428 ± 0.016	0.731 ± 0.02	0.911 ± 0.013	0.979 ± 0.005
16	0.038 ± 0.005	0.19 ± 0.012	0.476 ± 0.018	0.756 ± 0.018	0.917 ± 0.011
18	0.022 ± 0.004	0.133 ± 0.011	0.388 ± 0.016	0.683 ± 0.018	0.881 ± 0.013

			δ_s		
sparsity	0.5	0.6	0.7	0.8	0.9
4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
10	0.996 ± 0.002	0.999 ± 0.001	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
16	0.979 ± 0.005	0.996 ± 0.001	0.999 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
18	0.967 ± 0.006	0.993 ± 0.002	0.999 ± 0.001	1.0 ± 0.0	1.0 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

Table 12. Percentage of 2,000 trials for which the goodness-of-fit tests showed non-lack of fit to an $F_{65,65}$. Distribution

s			
2	3	4	5
0.45	0.88	0.95	0.96

Table 13. Simulated p-RIP for an $F_{65,65}$ Random Matrix

				δ_s		
ν_1	ν_2	0.0	0.1	0.2	0.3	0.4
65	65	0.0 ± 0.0	0.313 ± 0.0	0.582 ± 0.0	0.777 ± 0.0	0.891 ± 0.0
				δ_s		
ν_1	ν_2	0.5	0.6	0.7	0.8	0.9
65	65	0.945 ± 0.0	0.97 ± 0.0	0.983 ± 0.0	0.99 ± 0.0	0.995 ± 0.0

* Entries are the p-RIP average ± standard deviation of the 2,000 draws.

IV. CIF Results

This section presents the results of applying the CIF method on the breast cancer and small cell lung cancer dataset.

4.1 Bernoulli Example: Breast Cancer

The breast cancer data for this project was obtained from the National Center for Biotechnology Information [35]. It consists of 107 next-generation sequence (NGS) files from benign and malignant tumors. NGS sequencing is conducted by different organizations and captured in a variety of formats. Common to all NGS files is the sampling method. Essentially, base-pair subsets of varying lengths are sampled millions of times along the genome of interest. Each of these samples are called a read, and NGS data files contains millions of reads. Figure 5 is an example of two reads from a data file of 36 base-pairs.

CCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGC

CTGCCAGTAGCATATGGGTGTCTCAAAGCCAAAGCC

Figure 5. Illumina NGS Reads

Reads such as these were tokenized to create k-length strings of base-pair called k-mers. The tokenization process consisted of taking k-length subsets of the reads by file. All files become observations, and all k-mers produced from each file become features. The first k base-pairs become the first k-mer feature; the second k base-pairs would become the second k-mer feature this continues until no more k-length subsets can be produced. When all the observation are aligned, this process produced

a sparse binary matrix. The entries of the row vectors are '1' if a k-mer was produced from that file; else it was assigned a '0'. Figure 6 is an illustration of this process.

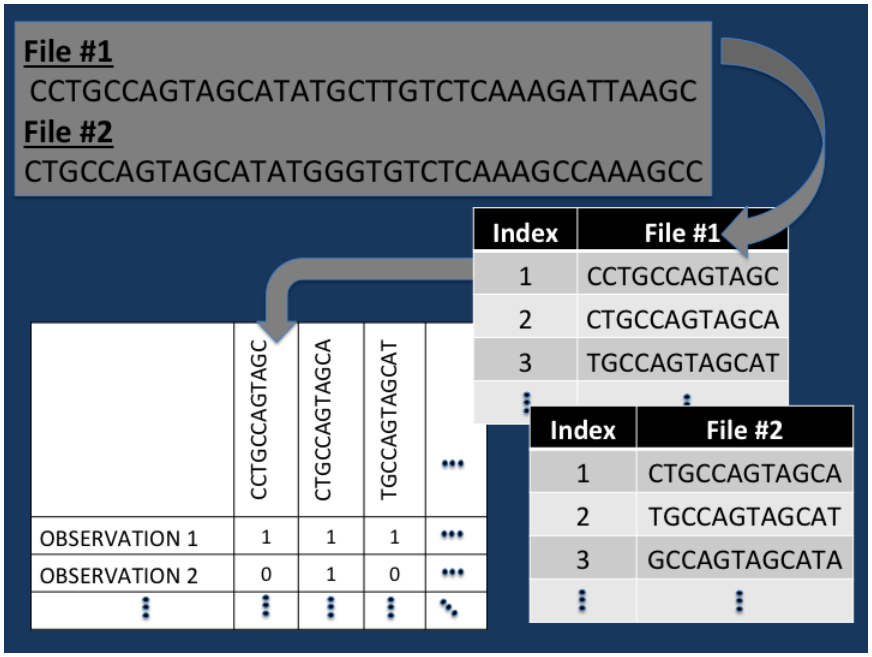


Figure 6. Sequence Data Tokenization Illustration.

A read of length N produces $N - k + 1$ k-mers with a step size of 1 along the read. These k-mers become the features for phenotype classification. Using k-mers in this capacity relies on the following assumptions:

Assumption 4.1.1 Organs affected by cancer exhibit common variations in base-pairs along the genome sequence. These variations can be captured with "sufficient" k-mer tokenization.

Under this assumption, exploring the entire genome, or rather, using all the reads in a NGS data file is superfluous. This assumption suggests that genome classification can be conducted using k-mers constructed from a fraction of the sequences in each NGS data file. This assumption leads to Assumption 4.1.2.

Assumption 4.1.2 There exist a sufficient number of reads, ξ, to generate a cardinality of k-mers that represents the phenotype.

If there exist a k-mer candidate set that can be used to define a phenotype; then there exist a $\xi << \Xi$ from which a candidate set of k-mers can be generated where Ξ represents the number of reads in each data file. The number of possible unique k-mers, given the four possible base-pairs, is 2^{2k}. The number of k-mers in a sequence read is $N - k + 1 \approx N$ for $k << N$; where N is the number of base-pairs in the genome sequence. The ratio between unique k-mers to number of k-mers created from a genome sequence is given by:

$$\frac{2^{2k}}{N - k + 1}. \tag{19}$$

For $k << N$, a genome sequence can be tokenized to provide a cardinality that, under Assumption 4.1.1, maintains that k-mers are a viable option for feature selection. The next step was to determine k, and an appropriate number of ξ to process in order to produce a sufficient cardinality.

The "right-size" k varies with the type of genome being sequenced as well as with number of groups to classify amongst. A study conducted in 2005 of 32 archaea ribosomal ribonucleic acid (rRNA) sequences found 12-mers successful in distinguishing between three genera. It assessed the similarity between sequences by calculating an out-of-place measure between k-mers [36]. RNA molecules are of particular importance for decoding genes [37], and for this reason $k = 12$ was used for the tokenization process. Using $\xi = 1,000$ and $k = 12$ produced a 2,675,000 12-mers from the 107 data files. This resulted in a dictionary of 690,553 unique 12-mers; which will be referred to as the "parent" matrix. In terms of the framing process, tokenization generated

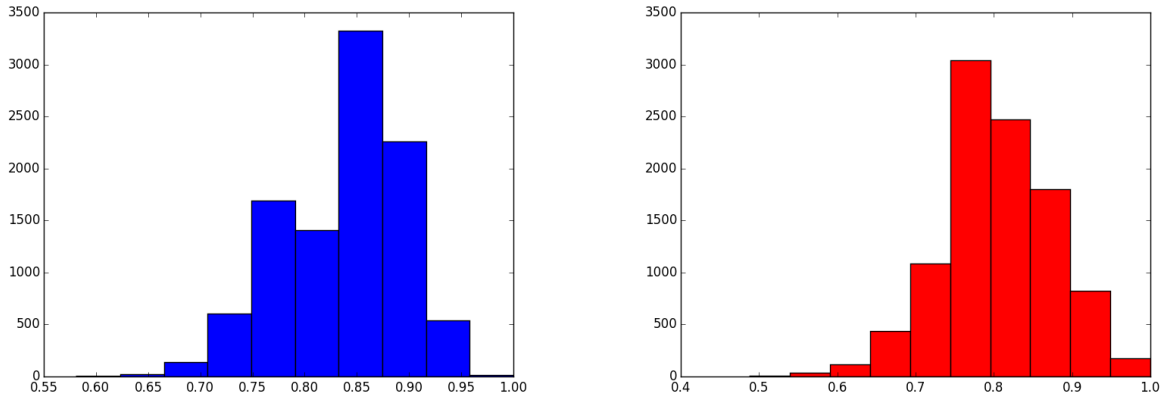
690,553 candidate vectors.

L_1 Regularized Logistic Regression.

Equation 9 is the objective function for L_1 regularized logistic regression. The hyper-parameter, α, controls the regularization of the L_1-norm loss function. Tuning α reveals the trade-space between model sparsity and classification performance. The data were randomly split into a training and test set with 60 and 47 observations respectively. Next, an exponential grid search was used to tune α on the training set which decreased the number of features in order to assess classification accuracy as a function of sparsity. The search grid was implemented with 50 increments which corresponded to 50 α's producing 50 sparsities.

The features were randomly selected, at each α, and logistic regression was performed using those features. This process was bootstrapped 10,000 times for each α. Average classification accuracy computed over 10,000 runs was used as the evaluation criteria. The α that yielded the highest classification accuracy on the test set was chosen as the optimal setting. As a point of reference, using $\mu_c = 0.30$ to construct an incoherent matrix from the parent set of 690,553 features reduced the the dimensionality down to 81 features. L_1 regularization further reduced the dimensionality of the parent set to $s = 49$, (Parent_{49}). Additionally, logistic regression was bootstrapped 10,000 times, with a 60/47 split, on the original parent set of 690,553 features (Parent_{OF}) and the average classification accuracy was recorded as a rudimentary baseline. These results served as a "best" accuracy for this problem and provided a means for comparison as displayed on Table 14. Table 14 also shows the results from a CIF-constructed dataframe using $\mu_c = 0.30$. This reduced the parent set to 41 features and then down to $s = 35$ (CIF_{35}) after L_1 regularization. Comparing the results on Table 14 highlights the need for CIF. As displayed, Parent_{OF}

produced an average classification accuracy of 84% while Parent_{49} produced and average accuracy of 52%, with 49 features. CIF_{35} produced and average accuracy of 89% with 35 features. Using CIF, with $\mu_c = 0.30$ produced higher average classification accuracy than the baseline and Parent_{49}, which used worst-case coherence as a single criteria. Figure 7 provides the histograms of the accuracy for Parent_{OF} and CIF_{35} runs.



(a) Parent_{OF}. (b) CIF_{35}.

Figure 7. Classification Accuracy Distribution.

The point estimates in Table 14 and visual inspection of histograms in Figure 7 indicates that CIF_{35} outperforms Parent_{OF}. A Wilcoxon signed-rank test was conducted to determine if there exist a statistical difference between the medians of the distribution of average accuracies for Parent_{OF} and CIF_{35}. The hypotheses were as follows:

$$H_0 : \text{median}_{Parent_{OF}} - \text{median}_{CIF_{35}} \geq 0$$

$$H_a : \text{median}_{Parent_{OF}} - \text{median}_{CIF_{35}} < 0$$

This test revealed that the CIF_{35} median average accuracy was greater than Parent_{OF} with a p-value = 1.73E-08. Consequently, I concluded that the CIF process

produced a reduced dataframe with classification accuracy better than the "best" dataframe for this example.

Table 14. Accuracy descriptive statistics: Breast Cancer

					99% CI	
	Median	Std.Dev	Min	Max	LB	UB
Parent_{OF}	0.837	0.055	0.581	1.000	0.836	0.839
Parent_{49}	0.535	0.060	0.256	0.721	0.522	0.526
CIF_{35}	0.907	0.055	0.605	1.000	0.891	0.894

Further investigation revealed the variance inflation factor (VIF) on the features for the CIF_{35} frame ranged from [1.22, 54.7]. This is an indication of multicollinearity, which presents a problem on parameter estimates. Therefore, I investigated frames for worst-case criterion: $\mu_c \in \{0.20, 0.10, 0.05, s0.025, 0.01\}$ in order to fix the issue of multicollinearity. Table 15 shows the result of the framing process of these five coherence criteria. Levying worst-case coherence criteria on the selection process significantly reduced the dimensionality of the datafame from 690,553 to the n values in the table. Figure 8 is a comparison of average classification accuracy by average sparsity for the frames created using these criterion. Ultimately, sensing matrices constructed using $\mu_c = 0.05$ produced higher average classification accuracy than the other criteria. The best case was with an average $s = 16$ and an average accuracy of 0.946. Additionally, the $\mu_c = 0.05$ constructed frames corrected the multicollinearity issues with VIF = 1 on all the features. Again, CIF constructed frames drastically reduce the original dataframe and proved to be a sparse solution to the phenotype classification problem. Next, p-RIP was used to identify the frames that adheres to p-RIP. Again, p-RIP is not necessary for recovery; however, it provides a method to identify the smallest set of columns (minimum number of features) needed to guarantee sparse signal recover.

Table 15. CIF Results: Breast Cancer

μ_c	n
0.01	16
0.025	19
0.05	19
0.1	18
0.2	20

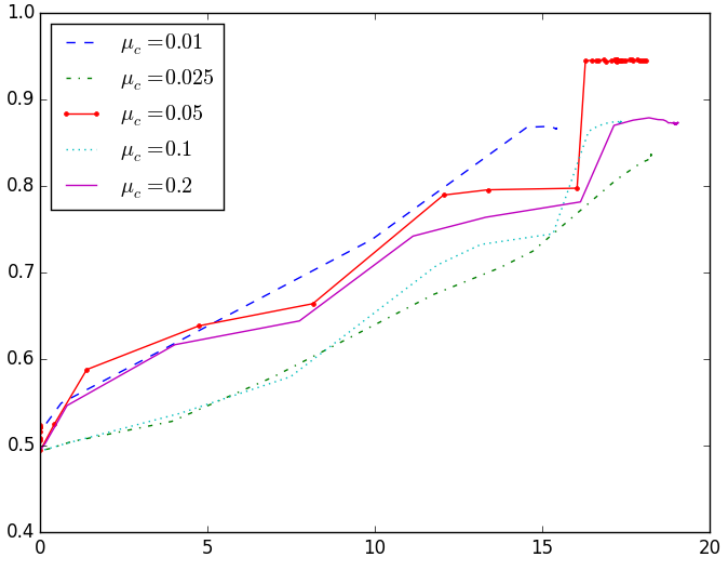


Figure 8. Coherence threshold comparison: average accuracies by sparsities for N = 10,000. The y-axis is performance accuracy and the x-axis is the sparsity.

p-RIP.

Figure 9 shows the probability of meeting the criteria for frame sufficiency by sparsity for the phenotype data for four worst-case coherence criterion. Four cases were chosen for investigation purposes, but also for exploring the trade space. For example, if $\mu_c = 0.05$ yielded a frame that did not adhere to p-RIP, and that criteria was more important that accuracy, then a different μ_c can be considered. It is important to understand the relationship between coherence, accuracy, and p-RIP. These plots highlight the influence that sparsity of the sensing matrix has on p-RIP in these

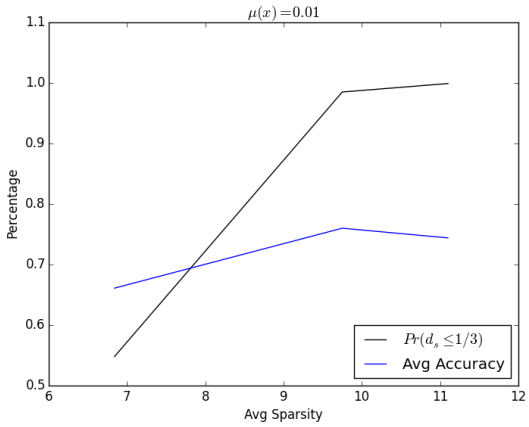
cases. The plots indicate that $\Pr(d_s \leq 0.333)$ decreases as sparsity of $\vec{\beta}$ increases. Recall ($d_s \leq 0.333$) is required for sufficient recovery. It is important to clarify that sparsity in the plots refers to $\vec{\beta}$, which corresponds to the number of features in the dataframe. Moreover, it provides a snapshot of the trade space between coherence, sparsity, classification accuracy, and p-RIP. Figure 8 indicates that a worst-case coherence criteria of $\mu_c = 0.05$ yields the best results in terms of classification accuracy for the phenotype data. The best results were with $s > 15$. The plot for $\mu_c = 0.05$ in Figure 9 mirrors this findings and indicates sufficient recovery for $s > 15$. Even though the other worst-case coherence criteria show dataframes with high p-RIP, the corresponding accuracies are below 80%.

The phenotype breast cancer dataframe is distributed $\text{Ber}(p \approx 0.007)$. Comparing the plots in Figure 9 to Table 9 indicate that p-RIP is higher, in general, for the actual data. It is possible that the framing process, and its pursuit for independent columns, forged dataframes that yielded smaller means and standard deviations. As a result, the p-RIP is higher where the data is vaster.

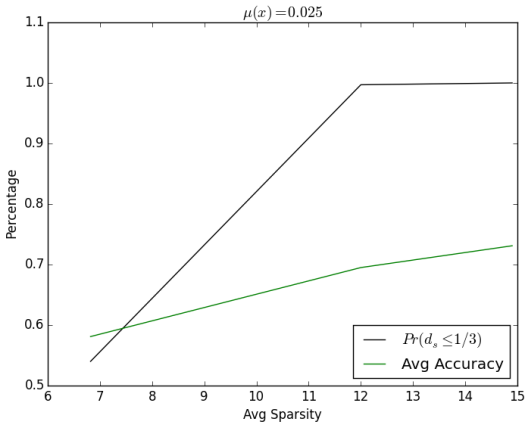
According to Figure 8, frames constructed with an average $s \geq 16$ performed well for the phenotype study across all worst-case criteria. Showing that those frames also have a high probability of sufficient recovery, allows me to certify those features as a basis for the phenotype. Negotiating the trade-space between coherence, sparsity, and accuracy prescribes a sampling convention to identify bases for the phenotype.

4.2 Gaussian Example: Small Cell Lung Cancer

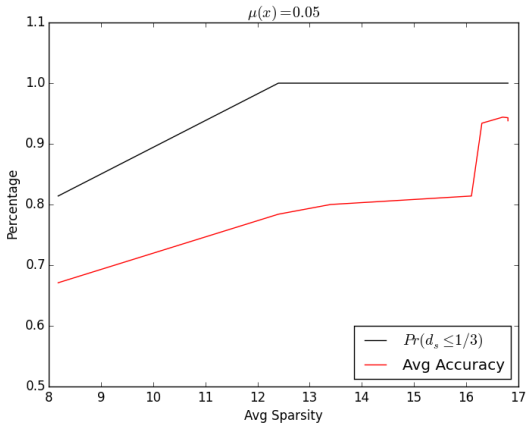
The small cell lung cancer (SCLC) dataset was obtained from the National Center for Biotechnology Information [38]. It consists of gene expression values from 65 samples across 16,382 genes. SCLC is a subtype of lung cancer, highly related to smoking, that has very distinguishable biological features [39]. This dataset has 23



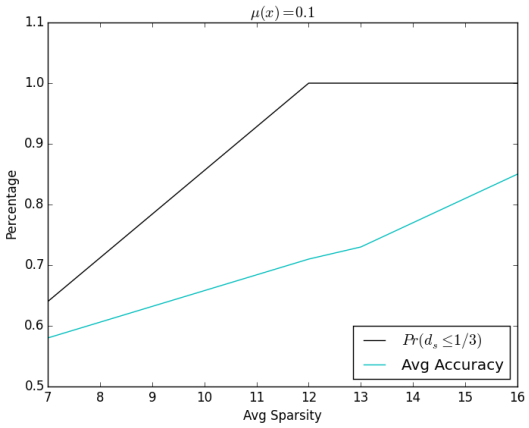
(a) $\mu_c = 0.01$.



(b) $\mu_c = 0.025$.



(c) $\mu_c = 0.05$.



(d) $\mu_c = 0.1$.

Figure 9. Classification Accuracy and p-RIP Comparison by Sparsity.

cases of SCLC and 42 normal tissues samples.

CIF Results.

The top 10% of the correlation magnitudes were used for this data in order to produce a candidate set large enough to maintain the integrity of a random process. Using 10% reduced the candidate set to 1,638 features. Table 16 shows the results of applying CIF to this candidate set for six coherence criterion. Again, worst-case coherence criteria significantly reduced the dimensionality of the dataset.

Table 16. CIF Results: SCLC

μ_c	n
0.01	4
0.025	1
0.05	4
0.1	4
0.2	5
0.3	9

L_1 Regularized Logistic Regression.

The SCSL data was split into training and testing sets with 33 and 32 observations respectively. I chose this split in order to avoid producing a test set with only 8 observations from the SCLC class. The CIF process and L_1 regularized logistic regression was otherwise implemented in the same manner as for the breast cancer data. L_1 regularization reduced the dimensionality of the parent frame to $s = 26$, (Parent_{26}). Table 17 reports the best classification accuracy for Parent_{26} at 0.81. Logistic regression was bootstrapped 10,000 times, with a 33/32, on this ParentOF, and the average classification accuracy was recorded as a rudimentary baseline. Table 17 shows the classification accuracy for Parent_{OF}, Parent36, and CIF_9. CIF_9 was constructed using $\mu_c = 0.30$. Wilcoxon signed-rank was conducted in the same manner as before between Parent_{OF} and CIF_9 using the bootstrapped accuracies. The tests revealed that the average accuracy for CIF_9 was greater than the Parent_{OF}, with a p-value approximately zero (p-value $<$ 1E-16).

Table 17. Accuracy descriptive statistics: SCLC

					99% CI	
	Median	Std.Dev	Min	Max	LB	UB
Parent_{OF}	0.942	0.041	0.667	1.000	0.778	1.000
Parent_{26}	0.810	0.092	0.364	0.970	0.424	0.970
CIF_9	1.000	0.006	0.910	1.000	0.939	1.000

p-RIP.

The F distribution is parameterized by two degrees of freedom; one for the numerator and one for the denominator. Therefore p-RIP is strictly governed by the degrees of freedom. Moreover, since the framing process transforms the $\mathrm{N}(0,1)$ random matrix into an $F_{65,65}$ random variable, p-RIP is directly related to the number of observations. Table 13 shows p-RIP for an $F_{65,65}$ distribution. Not shown is p-RIP for $\Pr(\delta_s \leq 1/3) = 0.82$, which is the best possible for $m = 65$. Maximum p-RIP for $m = 100$ is 0.90, and for $m = 150$ is 0.95. Though, sparsity does not affect p-RIP, however, it may affect the goodness-of-fit. Earlier I showed that the RIP conditions for sparsity and number of observations were satisfied for frames constructed using worst-case coherence $\mu_c = 0.05, 0.1$ with an average accuracy ≥ 0.95. Satisfying the RIP conditions based on observations and sparsity is even more pertinent given an $F_{65,65}$ random variable can only certify p-RIP for sufficient recovery to 82%.

V. K-means Clustering

Previously, I posited that CIF and p-RIP provide a means to certify feature sets as a bases for the phenotype. My object for this section is to build a dictionary of certified features, and explore the similarities between them in order to obtain a better understanding of the overall relationship between 12-mers and the phenotypes. This cluster analysis pertains to the breast cancer dataset.

5.1 Cluster Analysis

The purpose of the cluster analysis was to investigate if grouping 12-mers identified as features by CIF can be revealed as features for the phenotype. Clustering can be used for dimensionality reduction for supervised learning, under the assumption that members of a cluster have like or similar effect on the response. If this holds true, any feature in a cluster can be used to represent that cluster; which is how dimensionality reduction occurs with clustering. One of the objects of this research was to develop a method to identify features for phenotype classification. Cluster analysis was conducted to evaluate the features as potential biomarkers for the phenotype by determining how "unique" the features. Uniqueness is an assessment of the dissimilarity amongst the feature of the bases. In other words, if clustering amongst 12-mers yield comparable results to the CIF process, then the CIF process identifies clusters; centroids of clusters at best. If clustering does not lead to comparable or better classification performance, then the CIF process identifies more unique features.

To evaluate the features, I decided on two approaches for K-means clustering. The first approach was to establish a baseline by relaxing the worst-case coherence criteria to generate a candidate set of features that were not purposefully correlated to the phenotype. Worst-case coherence of 0.30 reduced the parent set to 35 features.

Thirty-five features is too small of a feature set to perform a cluster analysis, especially when considering up to 18 clusters. The best bases were sets with upward of 18 features, therefore, the cluster analysis had to include K=18. Therefore, worst-case criteria was relaxed to 0.8, which yielded a candidate set of 978 12-mers. The second approach involved bootstrapping the CIF process to construct 1,000 incoherent frames at $\mu_c = 0.05$. Then, L_1 regularized logistic regression was bootstrapped 1,000 times using the same grid search to tune α. Figure 10 is the accuracy histogram of the best α for each of the 1,000 incoherent frames. The average accuracy for these frames was 0.916. This approach yielded 5,232 different 12-mers. Figure 10 shows that some of the sets yielded classification accuracy of 100%. In fact, there were nine sets with 100% accuracy, which contained 201 different 12-mers. Therefore, I extended the second approach to investigate the features of these nine sets as centroids for clustering.

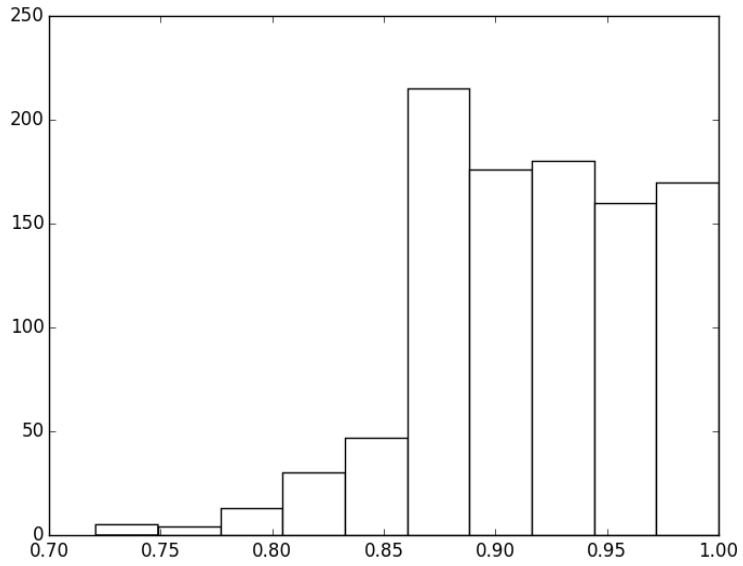


Figure 10. Classification Accuracy Simulations 1,000 simulation of the best CIF frames.

The Levenshtein distance between any two pairs of 12-mers is an integer value on

the interval of [1,12]. Cluster analysis may suffer if the groups are not well defined or if they are cluttered together. In addition to using the pre-computed Levenshtein distance (PRE) directly, two additional applications of Levenshtein distance were used to measure similarity between 12-mers; one is using the geometric mean (GM) the other is a coordinate system (Coord). These methods were implemented to exaggerate the dissimilarity between 12-mers to increase the range of dissimilarity. Both rely on establishing a datum on a four-dimensional coordinate system. Given the base-pairs are comprised of 'a', 't', 'c', and 'g', I established the origin at 'aaaaaaaaaaaa', 'ttttttttttt','ccccccccccc', and 'ggggggggggg', which correspond to an a-axis, t-axis, c-axis, and a g-axis. Any base-pair deviation from one of these axes corresponds to a unit-distance away from the origin in the direction of that axis. The geometric mean was calculated according to Equation 20:

$$gm = \sqrt[4]{l_a \times l_c \times l_t \times l_g}, \tag{20}$$

where $l_a, l_t, l_c,$ and l_g are the LD's from the axis of reference. The second method treats these distances as a coordinate set. Therefore, each 12-mer is located in a four dimensional vector space.

Clusters Results.

K-Means clustering was conducted for $K = [2, ..., 20]$ using the PRE, GM, and Coord approaches. These values of K were considered to correspond with sparsity of the breast cancer analysis. Figure 8 shows good classification accuracy using upward of 20 features. Lastly, I used the 12-mers from the frames which yielded 100% accuracy as the predetermined centroids for clustering. The candidate set of 12-mers from $\mu_c = 0.8$ will be referred to as $\mathcal{A}$ and the candidate set from simulation will be referred to as $\mathcal{B}$. Appendix C is the cluster plot for the GM approach on $\mathcal{A}$ and $\mathcal{B}$. For

visualization the data is plotted as $x = y = gm$. Appendix D is 3-dimensional plots for the Coord approach. Of note, the coordinate system is four-dimensional, however, only three dimensions were plotted. The color scheme was reserved to indicate cluster membership opposed to being used the fourth dimension. The clustering analysis, however, took all measurements into consideration. The GM plots for $\mathcal{A}$ and $\mathcal{B}$ look similar despite $\mathcal{B}$ has over five times the data. This is because the plot of $x = y = gm$ stacks the data on top of each other. Though the clusters are easy to distinguish, membership size cannot be visualized. The coordinate system plots appear to be better defined. Clusters are very apparent for smaller K, but membership size is clearer in comparison to the GM plots. The Coord plots for $\mathcal{A}$ and $\mathcal{B}$ look similar. Visual inspection of these plots suggests that $\mathcal{A}$ and $\mathcal{B}$ group similarly.

The clustering approach to dimensionality reduction and classification maintains that the similarity between 12-mers can be used for grouping and the members of a group express the same mutation in the genome. In order to assess this assumption, I built a classifier by randomly sampling one 12-mer from each cluster. I used these features to perform logistic regression again with a 60/40 split on the data. This process was repeated 1,000 times and the results are displayed in Table 19. Overall, $\mathcal{B}$ produced 12-mers which yielded better classification accuracy. This is expected given that set created by repeating the CIF process 1,000 times and collecting the features. However, these results are not comparable to the CIF results; except for small K. Smaller K produced results similar to smaller s in Figure 8.

Appendix E are the plots for the nine sets with 100% accuracy. The color scheme in these plots are for the fourth dimension; g-axis. Moreover, identical colors are an indication of an identical distance in the fourth vector space. Appendix F are the cluster plots of $\mathcal{A}$ and $\mathcal{B}$ using the features of the nine sets as centroids of the clusters. Table 20 is the classification accuracy for frames constructed from the clus-

Table 18. Average Cluster Classification Accuracy and Standard Deviation: $\mathcal{A}$

	$\mathcal{A}$: N = 978		
# Clusters	PRE	GM	Coord
2	0.66 ± 0.067	0.66 ± 0.064	0.67 ± 0.068
3	0.67 ± 0.071	0.67 ± 0.064	0.67 ± 0.072
4	0.63 ± 0.072	0.64 ± 0.069	0.63 ± 0.071
5	0.69 ± 0.083	0.70 ± 0.083	0.68 ± 0.082
6	0.64 ± 0.067	0.64 ± 0.070	0.64 ± 0.069
7	0.65 ± 0.067	0.64 ± 0.066	0.64 ± 0.069
8	0.65 ± 0.067	0.65 ± 0.065	0.65 ± 0.066
9	0.65 ± 0.069	0.65 ± 0.068	0.65 ± 0.066
10	0.66 ± 0.069	0.66 ± 0.060	0.66 ± 0.069
11	0.66 ± 0.069	0.65 ± 0.060	0.66 ± 0.067
12	0.66 ± 0.071	0.66 ± 0.062	0.66 ± 0.067
13	0.67 ± 0.075	0.67 ± 0.068	0.67 ± 0.070
14	0.67 ± 0.071	0.68 ± 0.077	0.67 ± 0.071
15	0.67 ± 0.074	0.69 ± 0.076	0.67 ± 0.069
16	0.67 ± 0.072	0.69 ± 0.078	0.68 ± 0.075
17	0.68 ± 0.075	0.69 ± 0.078	0.68 ± 0.077
18	0.68 ± 0.076	0.69 ± 0.079	0.68 ± 0.075
19	0.68 ± 0.077	0.69 ± 0.079	0.68 ± 0.079
20	0.68 ± 0.075	0.69 ± 0.079	0.69 ± 0.080

ters. On average these results are better than the results in Table 19, but still under performs CIF constructed frames. Clustering did not improve classification accuracy, nor were the results comparable to the CIF results. Grouping results based on the assigned centroid, specifically clustering $\mathcal{B}$, suggest that the sets are more unique to the phenotype than the individual features. If the features were more unique to the phenotype I would expect the results closer to 0.916 classification accuracy, which was the average accuracy from bootstrapping using the best α.

Table 19. Average Cluster Classification Accuracy and Standard Deviation: $\mathcal{B}$

	$\mathcal{B}$: N=5,232		
# Clusters	PRE	GM	Coord
2	0.72 ± 0.079	0.71 ± 0.078	0.71 ± 0.082
3	0.73 ± 0.075	0.72 ± 0.076	0.72 ± 0.080
4	0.56 ± 0.091	0.56 ± 0.091	0.57 ± 0.089
5	0.78 ± 0.071	0.78 ± 0.070	0.77 ± 0.071
6	0.60 ± 0.087	0.60 ± 0.089	0.60 ± 0.091
7	0.62 ± 0.085	0.62 ± 0.093	0.62 ± 0.090
8	0.65 ± 0.085	0.64 ± 0.090	0.64 ± 0.088
9	0.66 ± 0.086	0.65 ± 0.088	0.65 ± 0.088
10	0.68 ± 0.086	0.67 ± 0.086	0.67 ± 0.084
11	0.69 ± 0.083	0.68 ± 0.081	0.69 ± 0.080
12	0.71 ± 0.080	0.70 ± 0.079	0.69 ± 0.081
13	0.74 ± 0.074	0.72 ± 0.080	0.73 ± 0.076
14	0.75 ± 0.074	0.74 ± 0.076	0.73 ± 0.075
15	0.75 ± 0.072	0.74 ± 0.076	0.74 ± 0.071
16	0.76 ± 0.072	0.75 ± 0.075	0.74 ± 0.075
17	0.77 ± 0.069	0.75 ± 0.074	0.75 ± 0.072
18	0.77 ± 0.070	0.76 ± 0.072	0.76 ± 0.072
19	0.78 ± 0.067	0.77 ± 0.068	0.76 ± 0.070
20	0.78 ± 0.070	0.77 ± 0.068	0.77 ± 0.073

Table 20. Average Cluster Classification Accuracy and Standard Deviation: Assigned Centroids

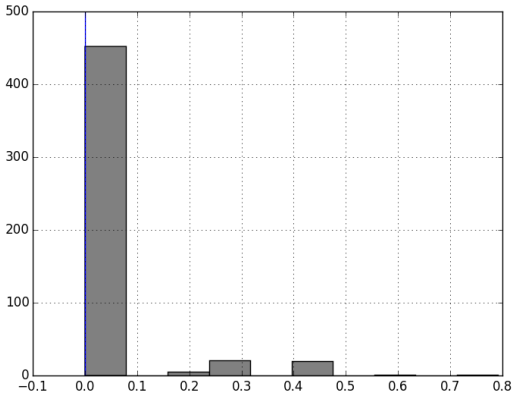
		$\mathcal{A}$: N = 978	$\mathcal{B}$: N=5,232
Set	# of Cluster	Coord	Coord
1	25	0.78±0.084	0.79±0.070
2	25	0.78±0.082	0.79±0.068
3	24	0.77±0.080	0.79±0.067
4	22	0.76±0.079	0.78±0.070
5	24	0.76±0.082	0.79±0.068
6	26	0.76±0.083	0.80±0.069
7	24	0.77±0.084	0.79±0.070
8	21	0.76±0.083	0.78±0.073
9	25	0.76±0.087	0.79±0.066

VI. Conclusions and Recommendations

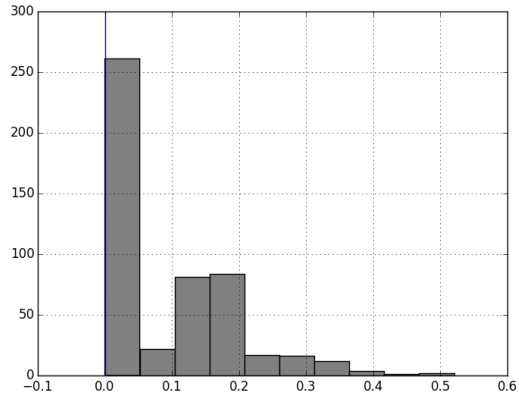
This research introduced CIF as a process developed for dimensionality reduction and feature selection using compressive sampling methods. My primary focus was to cast phenotype sequence data as a sparse binary matrix and leverage compressive sampling theory for dimensionality reduction and feature identification. I also evaluated the CIF process against a gene expression data set as well. I found that CIF produced incoherent frames with improved classification accuracy. Additionally, I showed that CIF produced sensing matrices with high probability of satisfying the RIP when applied to Bernoulli or Standard Normal matrices. Overall, I demonstrated CIF as viable approach to dimension reduction, feature selection, and subsequently phenotype classification.

The NGS sequence data used in this research derived from targeted sequencing and not whole genome data. Therefore, the inferences based on these results can only be extended to whole genome data. This was advantageous, because I posited that the CIF process identified bases for the phenotype. To validate this position I recommend evaluating the classification performance of these bases applied to a different breast cancer datasets as well as a dataset using whole genome sequences. For this research, I applied CIF to DNA sequence data by tokenizing sequences into a dictionary of 12-mers. All features were the same length. I recommend investigating CIF on a dictionary of k-mers of varying lengths to explore improving classification performance. Lastly, I showed CIF as a viable option for dimensionality reduction of Bernoulli and Gaussian sensing matrices; specifically for the two datasets analyzed in this research. I recommend combining sequence and gene expression candidate sets together to investigate whether or not this mixture model can further improve phenotype classification accuracy.

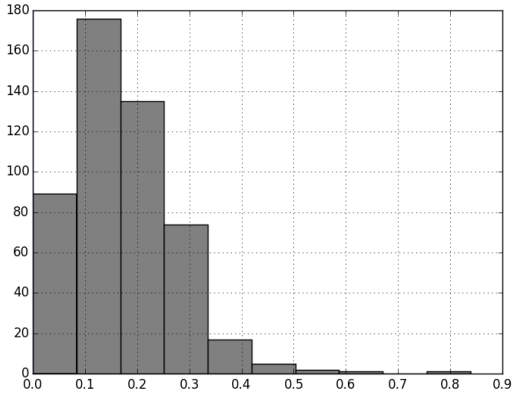
Appendix A. Histograms: Draws of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ Transformations.



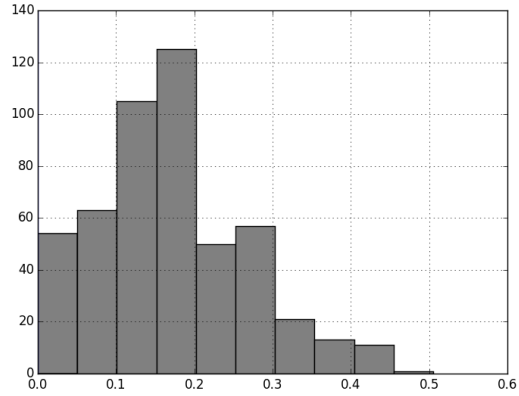
(a) $p = 0.004$; $s = 4$.



(b) $p = 0.004$; $s = 10$.

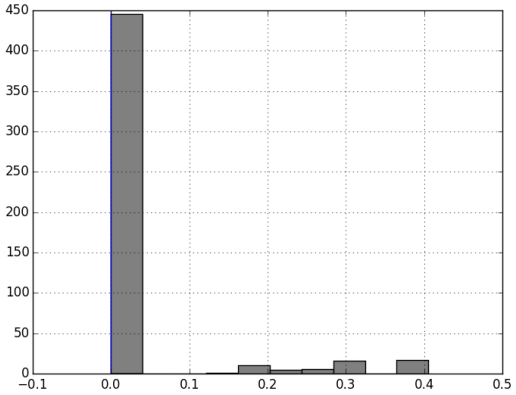


(c) $p = 0.004$; $s = 16$.

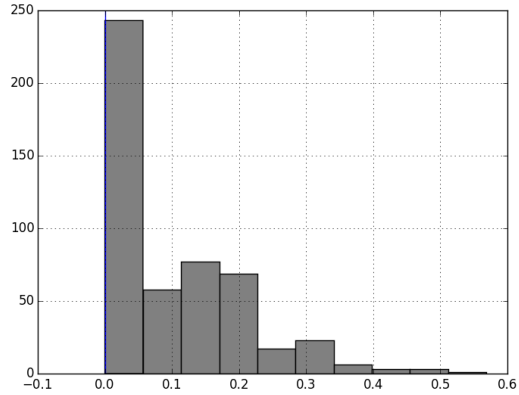


(d) $p = 0.004$; $s = 18$.

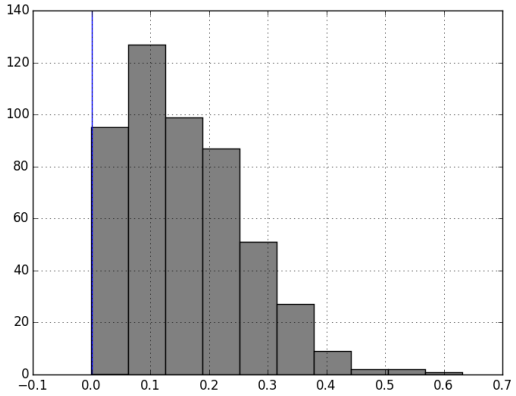
Figure 11. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.004)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



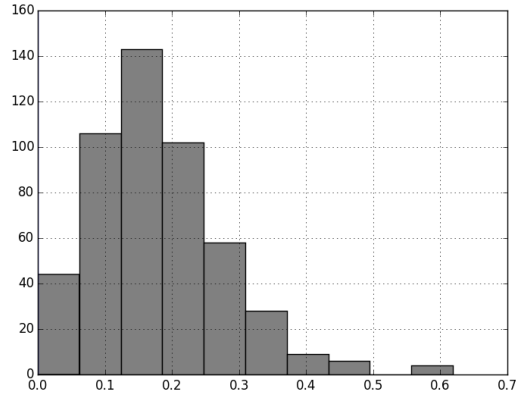
(a) $p = 0.006$; $s = 4$.



(b) $p = 0.006$; $s = 10$.

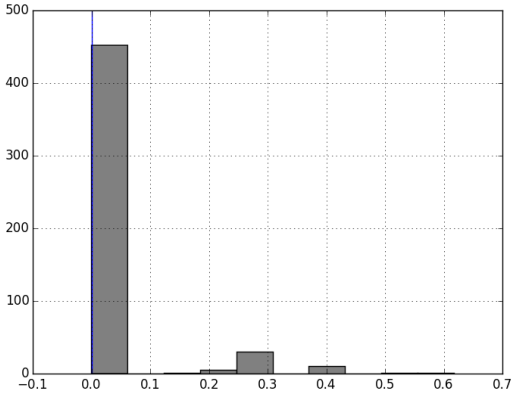


(c) $p = 0.006$; $s = 16$.

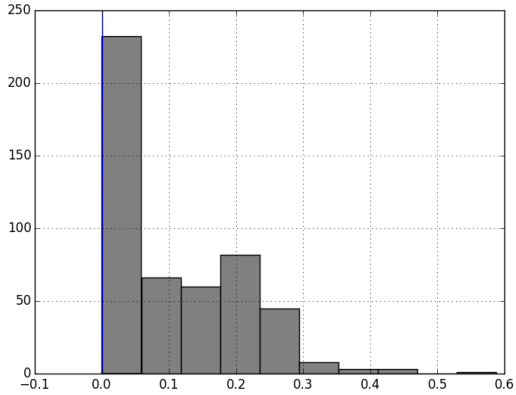


(d) $p = 0.006$; $s = 18$.

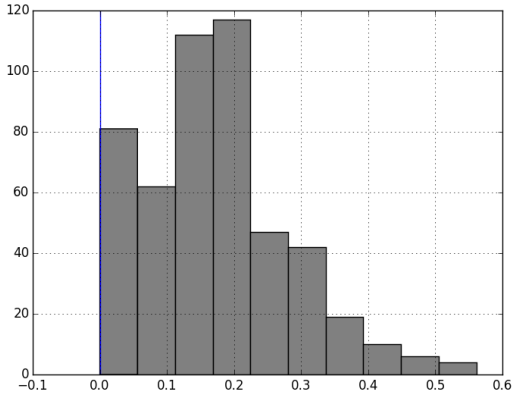
Figure 12. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.006)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



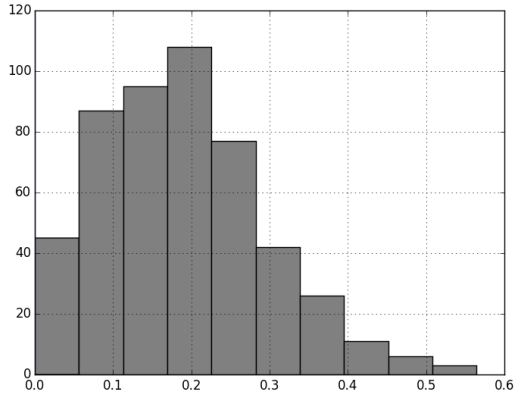
(a) $p = 0.007$; $s = 4$.



(b) $p = 0.007$; $s = 10$.

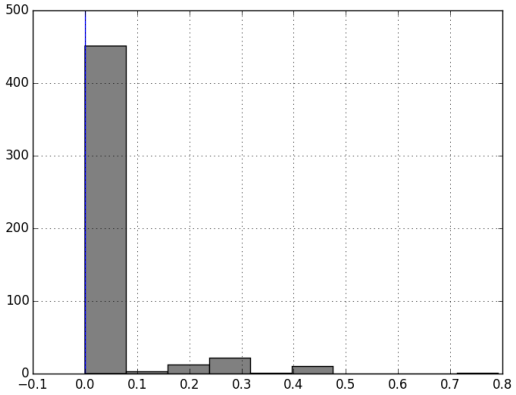


(c) $p = 0.007$; $s = 16$.

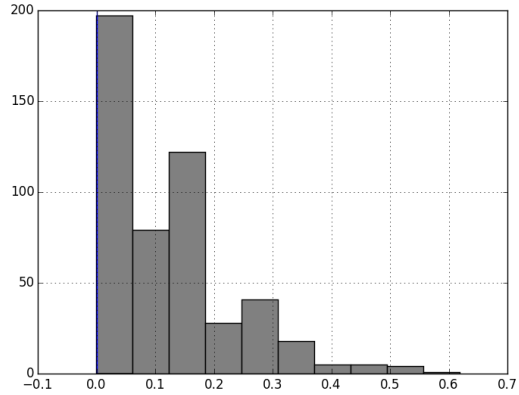


(d) $p = 0.007$; $s = 18$.

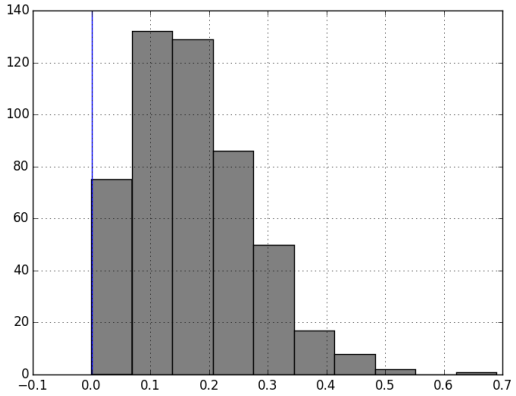
Figure 13. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.007)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



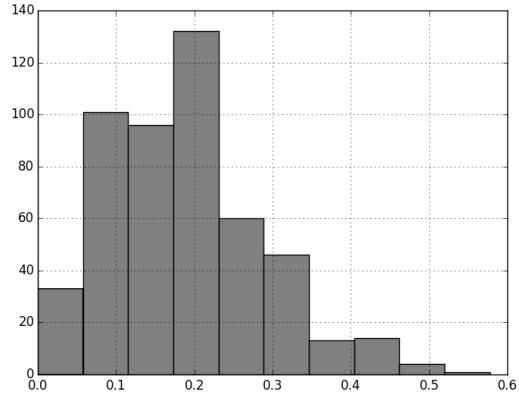
(a) $p = 0.008$; $s = 4$.



(b) $p = 0.008$; $s = 10$.

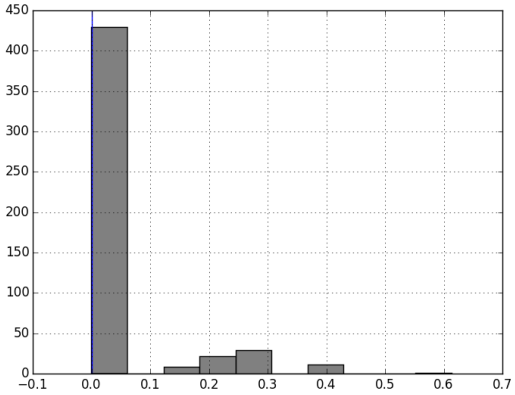


(c) $p = 0.008$; $s = 16$.

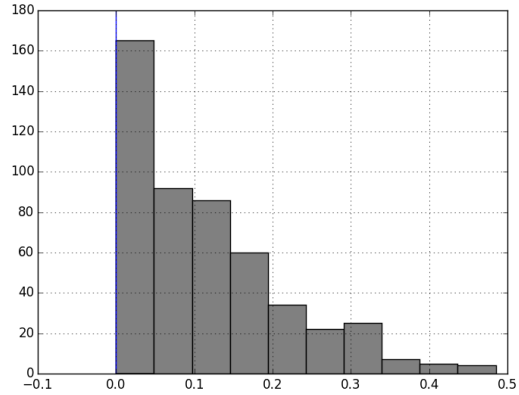


(d) $p = 0.008$; $s = 18$.

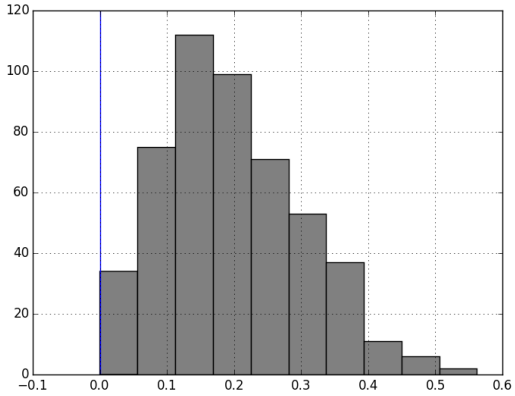
Figure 14. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.008)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



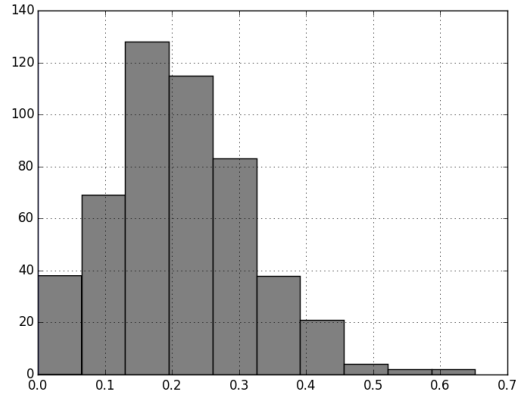
(a) $p = 0.01$; $s = 4$.



(b) $p = 0.01$; $s = 10$.

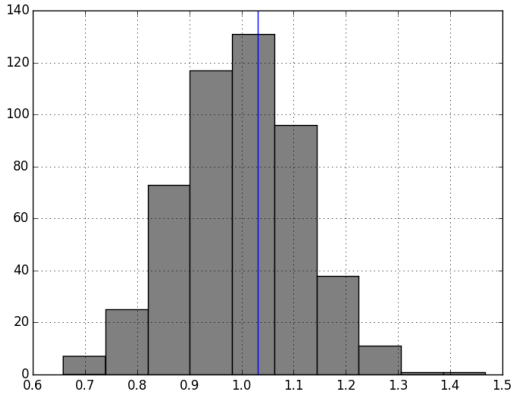


(c) $p = 0.01$; $s = 16$.

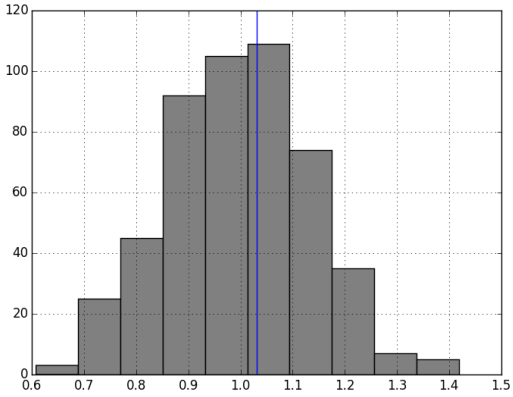


(d) $p = 0.01$; $s = 18$.

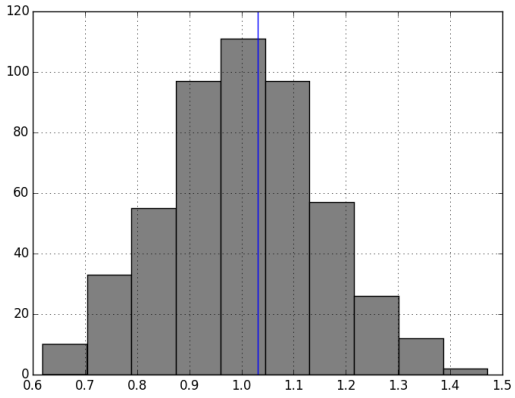
Figure 15. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.01)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



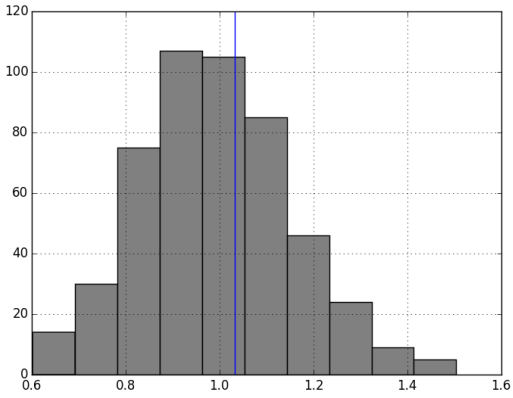
(a) $s = 2$.



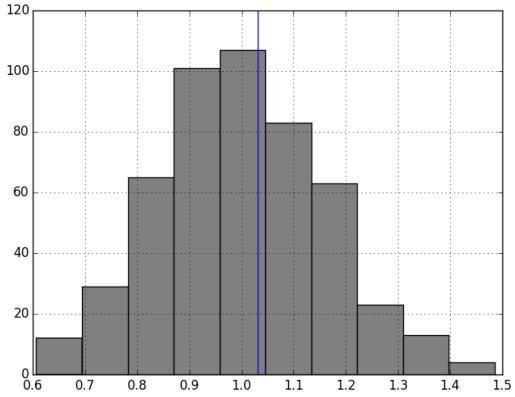
(b) $s = 3$.



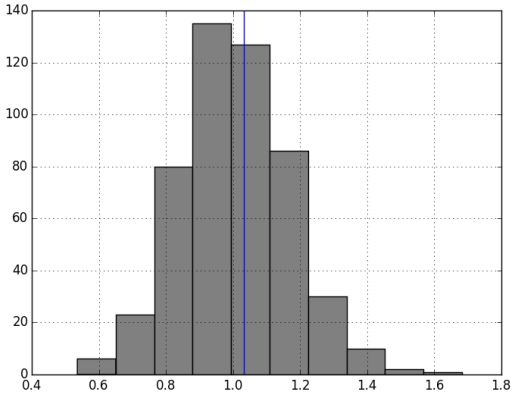
(c) $s = 4$.



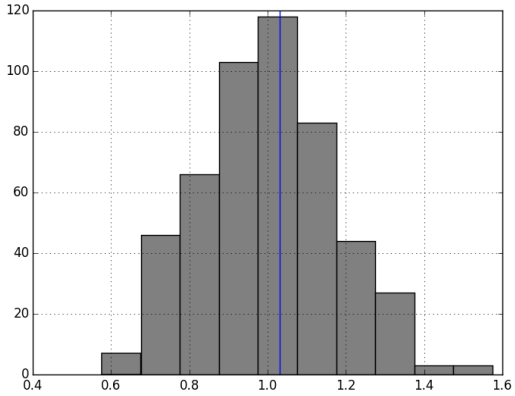
(d) $s = 5$.



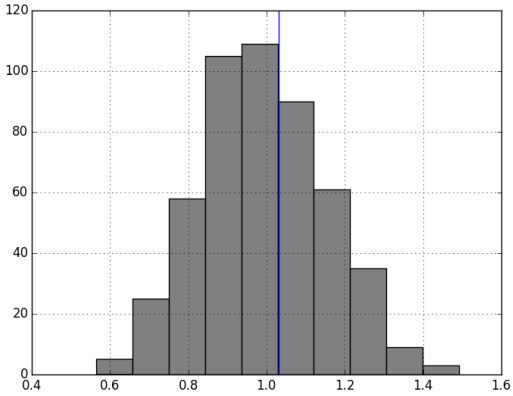
(e) $s = 6$.



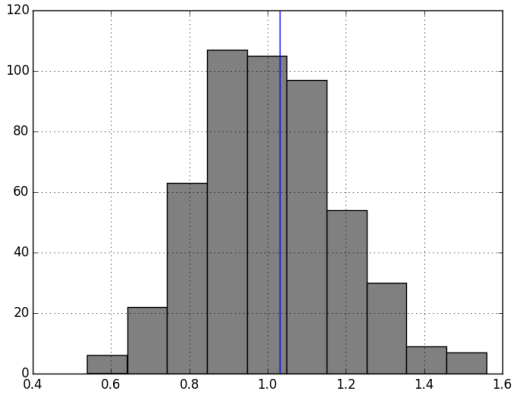
(f) $s = 8$.



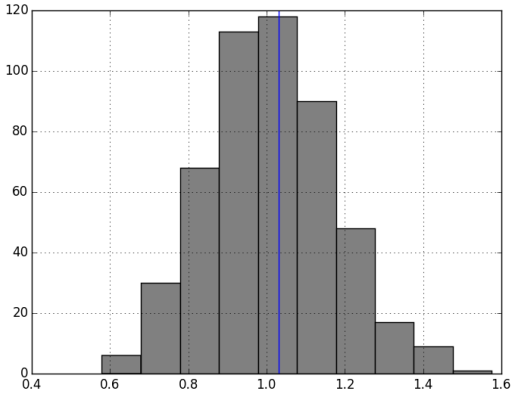
(g) $s = 10$.



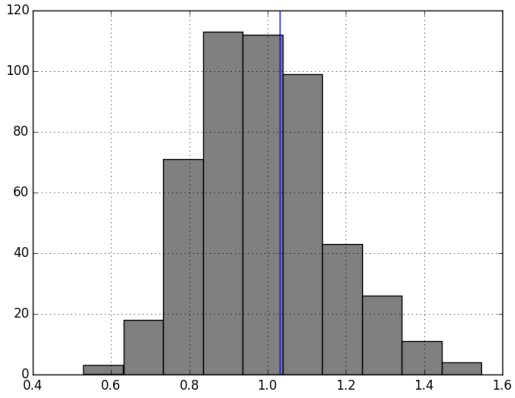
(h) $s = 12$.



(i) $s = 14$.



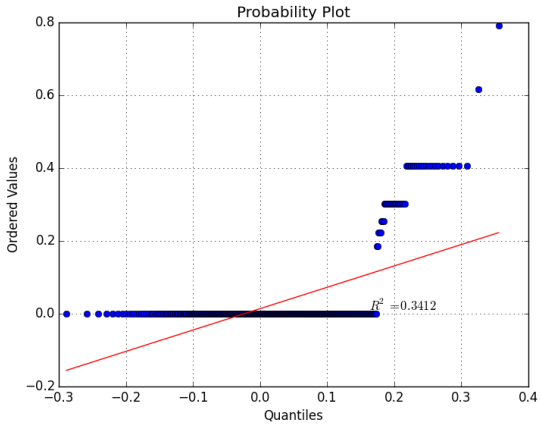
(j) $p = 0.01$; $s = 16$.



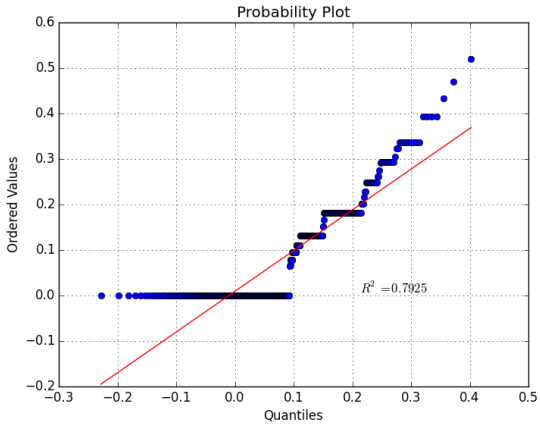
(k) $s = 18$.

Figure 16. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{N}(0,1)$ and $\vec{\beta}$ for $s \in \{2,3,4,5,6,8,10,12,14,16,18\}$.

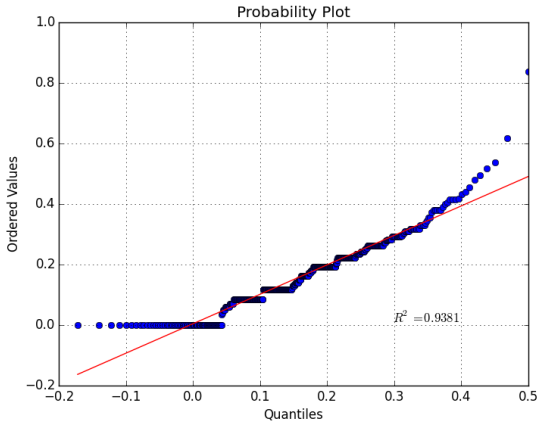
Appendix B. Quantile Plots: Draws of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ Transformations



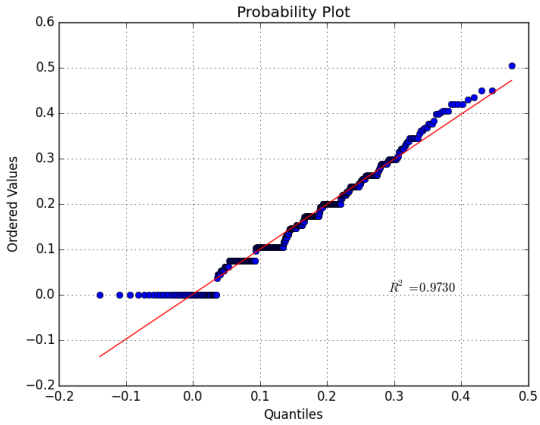
(a) $p = 0.004$; $s = 4$.



(b) $p = 0.004$; $s = 10$.

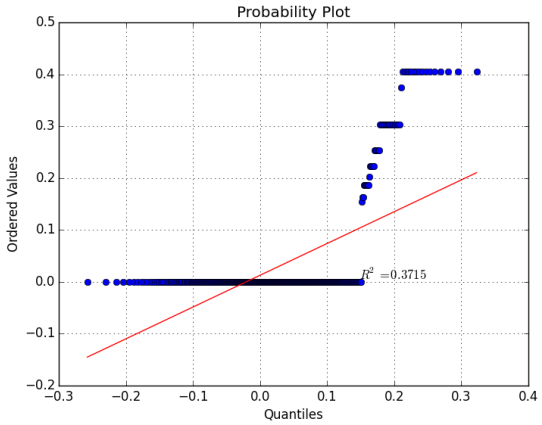


(c) $p = 0.004$; $s = 16$.

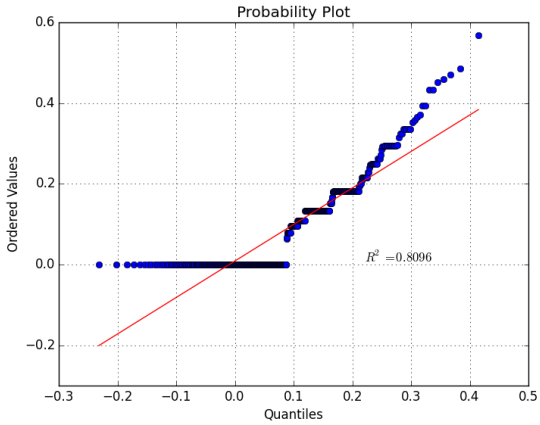


(d) $p = 0.004$; $s = 18$.

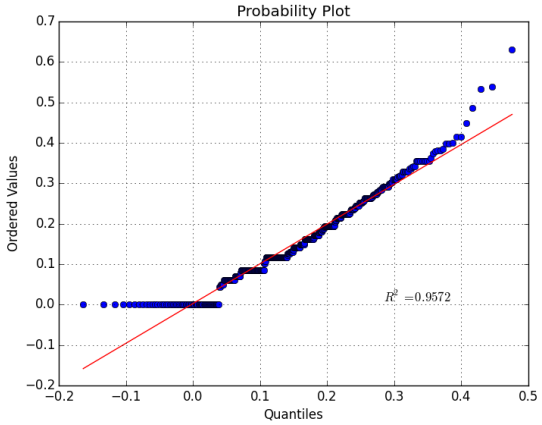
Figure 17. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \text{Ber}(0.004)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



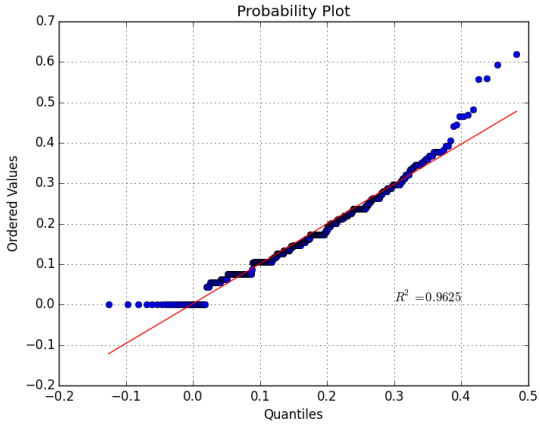
(a) $p = 0.006$; $s = 4$.



(b) $p = 0.006$; $s = 10$.

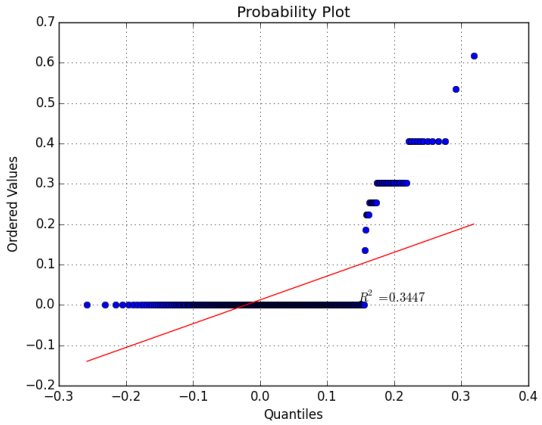


(c) $p = 0.006$; $s = 16$.

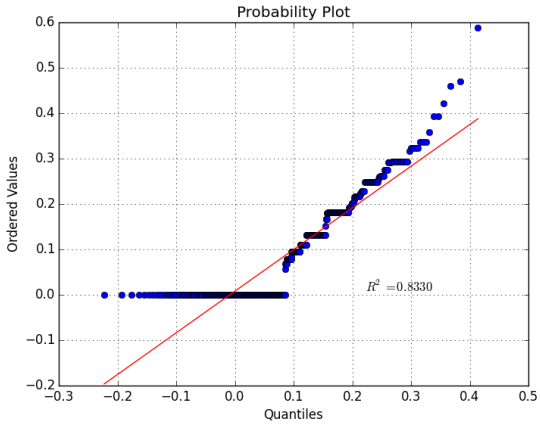


(d) $p = 0.006$; $s = 18$.

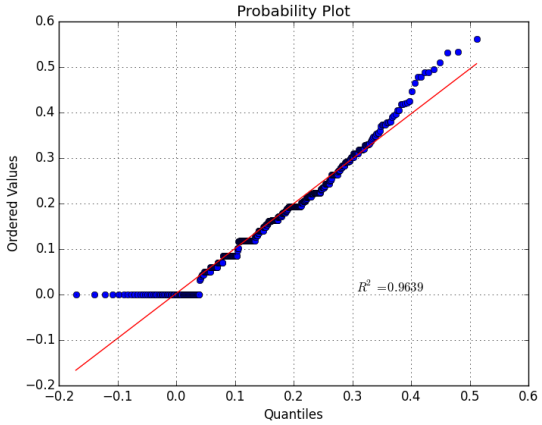
Figure 18. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \text{Ber}(0.006)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



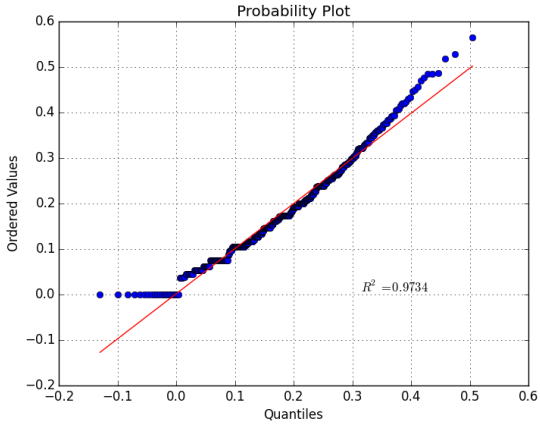
(a) $p = 0.007$; $s = 4$.



(b) $p = 0.007$; $s = 10$.

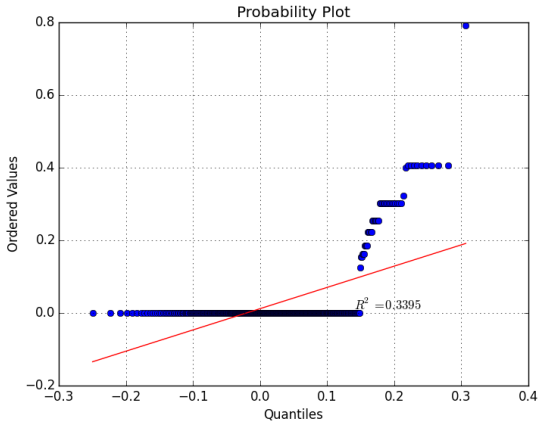


(c) $p = 0.007$; $s = 16$.

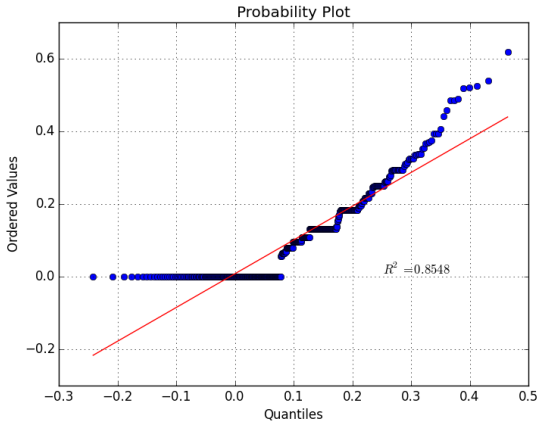


(d) $p = 0.007$; $s = 18$.

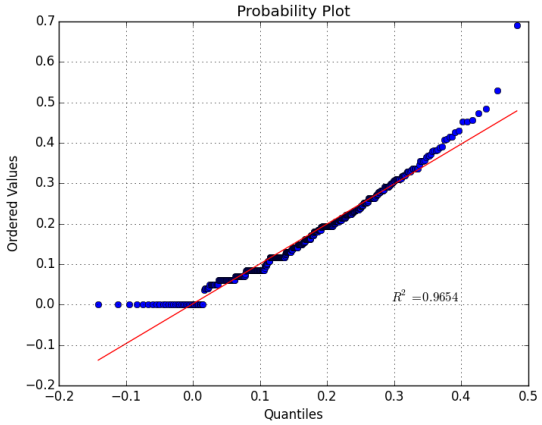
Figure 19. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \text{Ber}(0.007)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



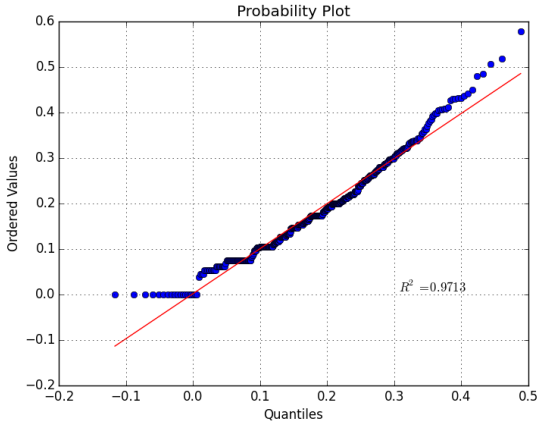
(a) $p = 0.008$; $s = 4$.



(b) $p = 0.008$; $s = 10$.

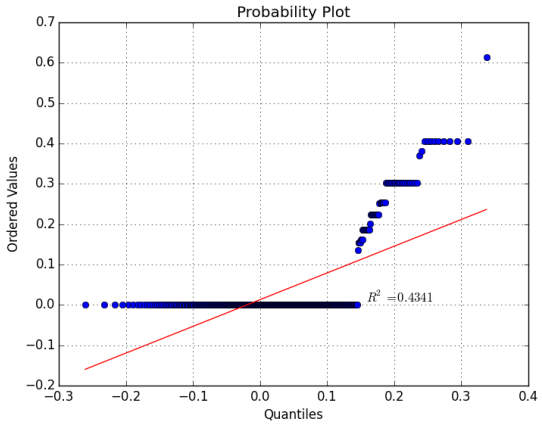


(c) $p = 0.008$; $s = 16$.

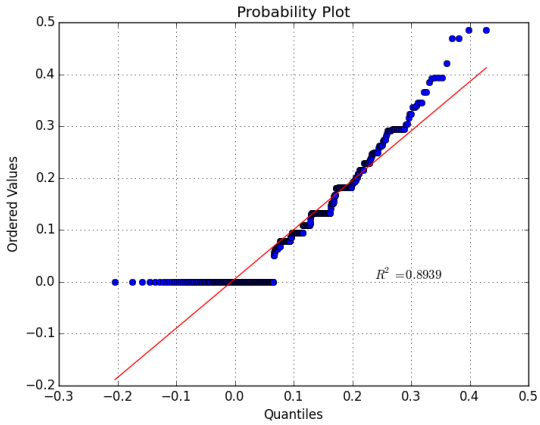


(d) $p = 0.008$; $s = 18$.

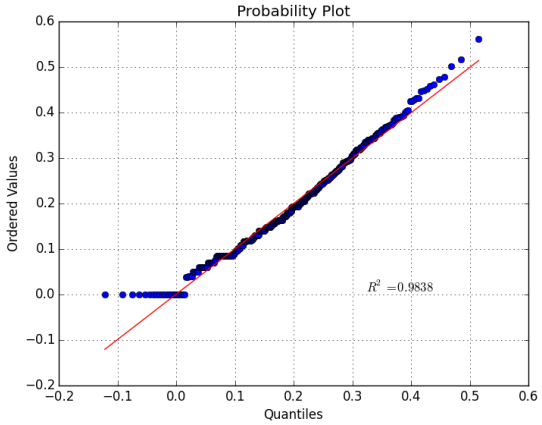
Figure 20. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \text{Ber}(0.008)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



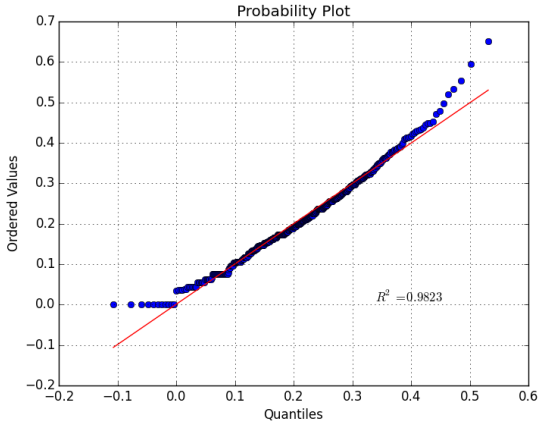
(a) $p = 0.01$; $s = 4$.



(b) $p = 0.01$; $s = 10$.

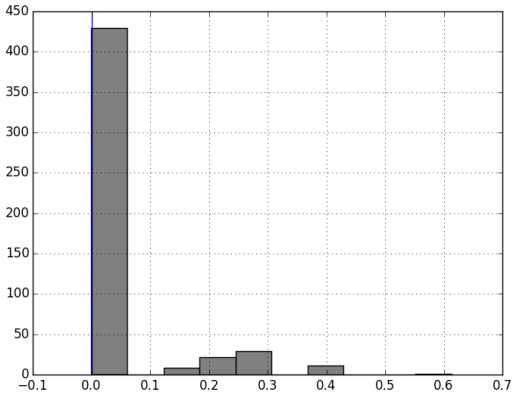


(c) $p = 0.01$; $s = 16$.

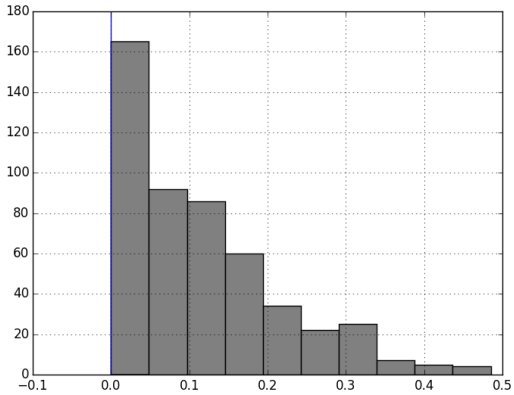


(d) $p = 0.01$; $s = 18$.

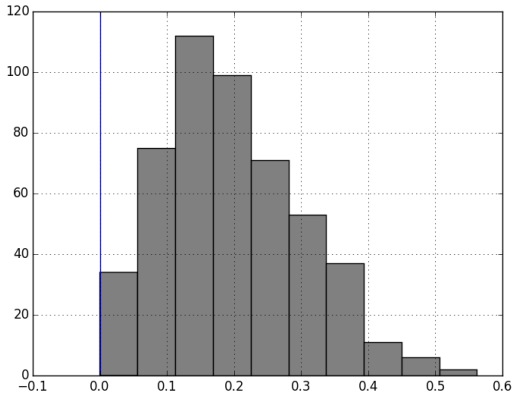
Figure 21. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \text{Ber}(0.01)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



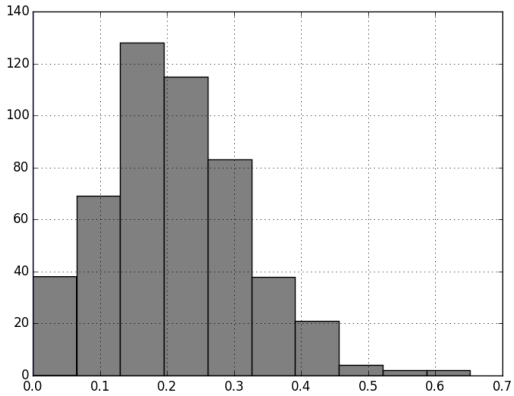
(a) $p = 0.01$; $s = 4$.



(b) $p = 0.01$; $s = 10$.

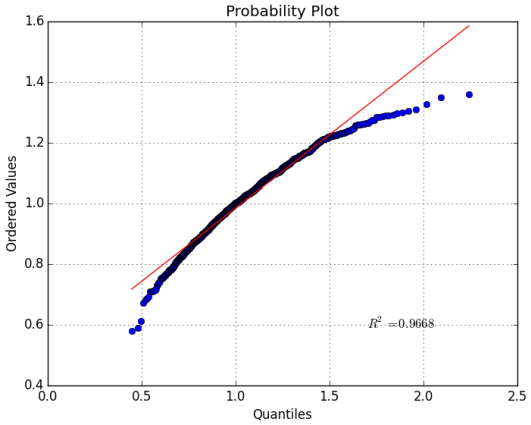


(c) $p = 0.01$; $s = 16$.

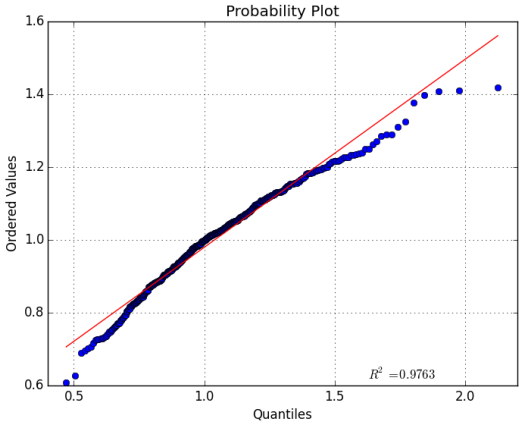


(d) $p = 0.01$; $s = 18$.

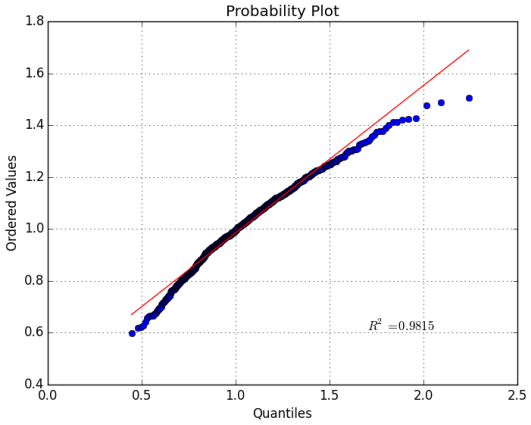
Figure 22. Histogram: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{Ber}(0.01)$ and $\vec{\beta}$ for $s \in \{4, 10, 16, 18\}$.



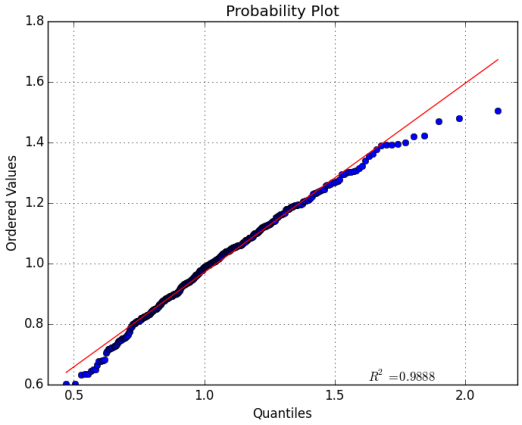
(a) $s = 2$.



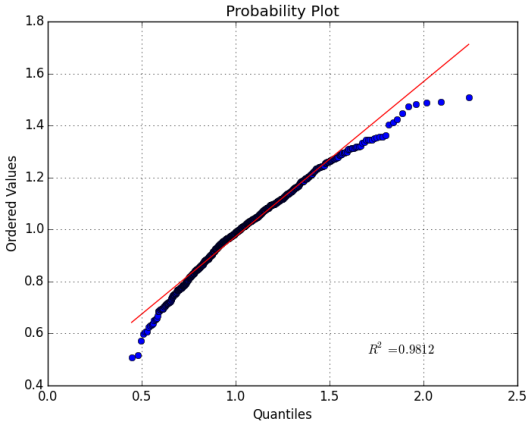
(b) $s = 3$.



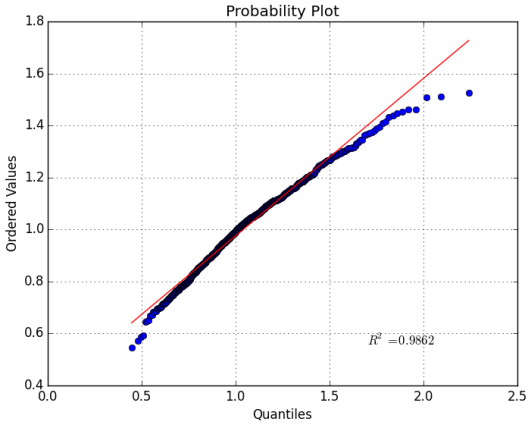
(c) $s = 4$.



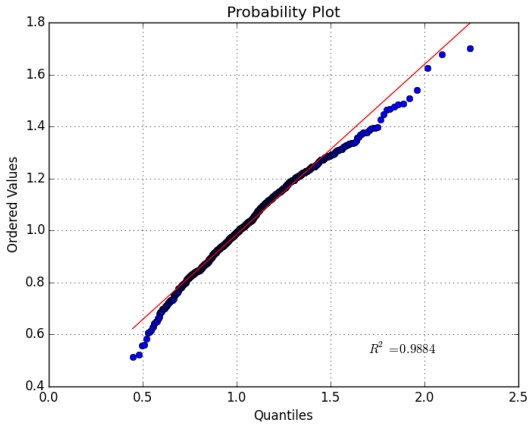
(d) $s = 5$.



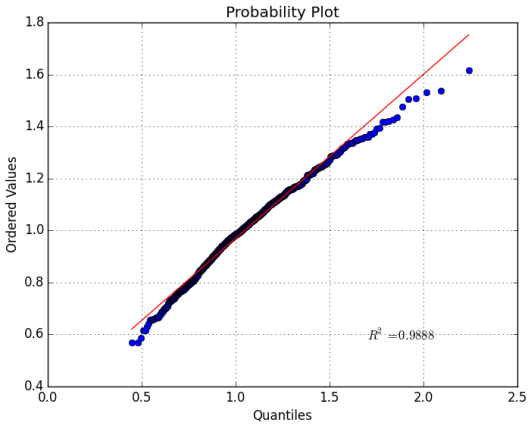
(e) $s = 6$.



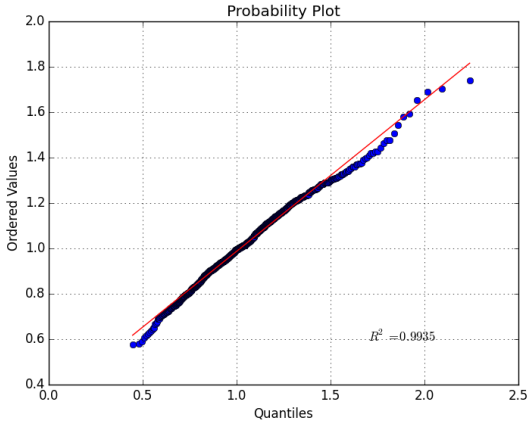
(f) $s = 8$.



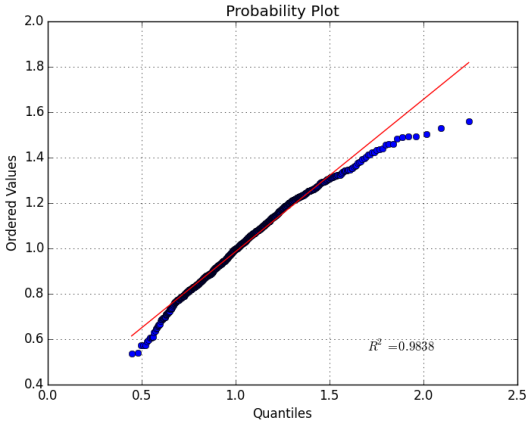
(g) $s = 10$.



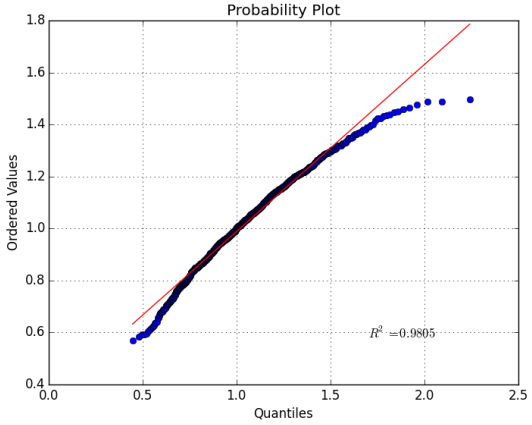
(h) $s = 12$.



(i) $s = 14$.



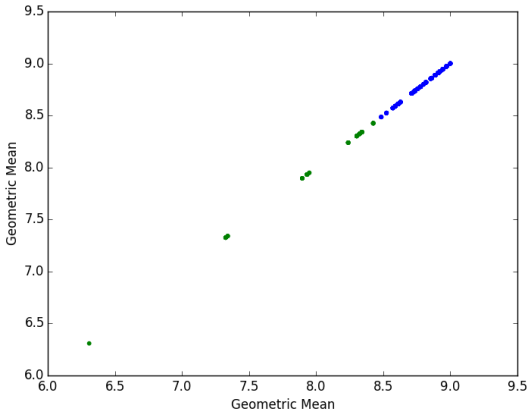
(j) $s = 16$.



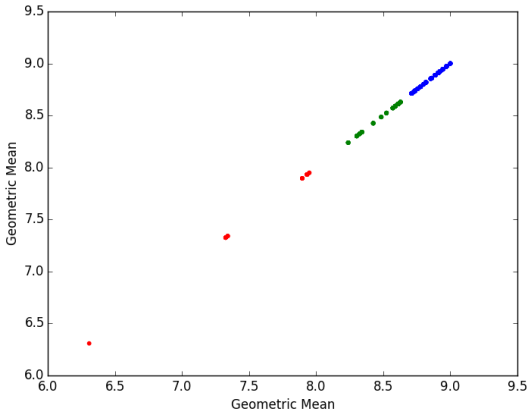
(k) $s = 18$.

Figure 23. Quantile Plots: Draw of $\|\tilde{\mathbf{X}}\vec{\beta}\|_2^2$ for $\mathbf{X} \sim \mathrm{N}(0,1)$ and $\vec{\beta}$ for $s \in \{2,3,4,5,6,8,10,12,14,16,18\}$.

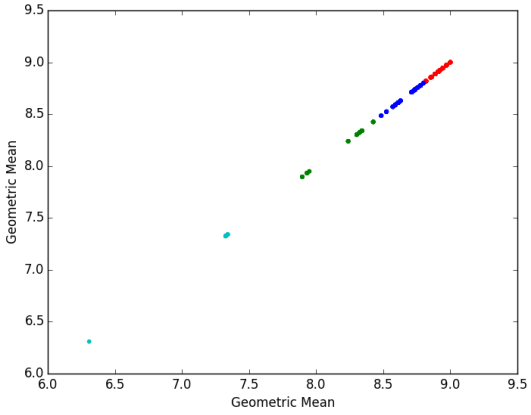
Appendix C. Geometric Mean Cluster Plots Using LD



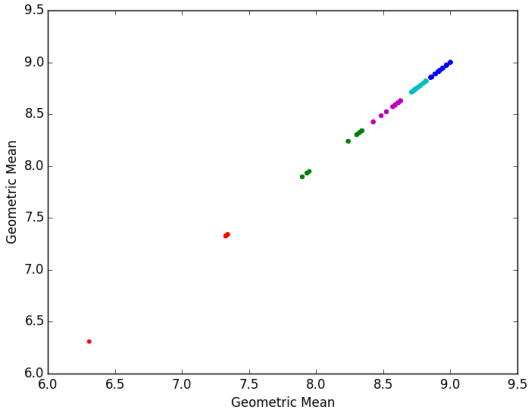
(a) K=2.



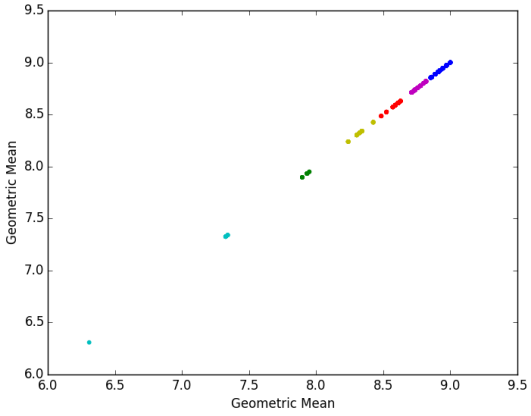
(b) K=3.



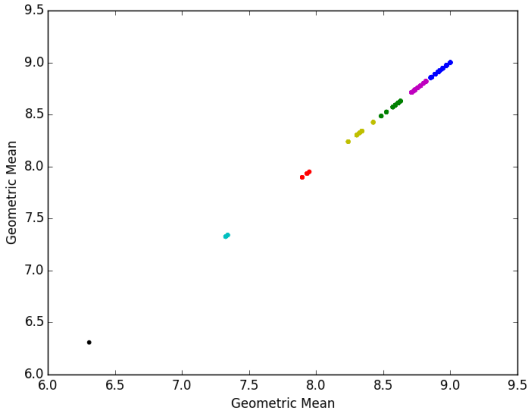
(c) K=4.



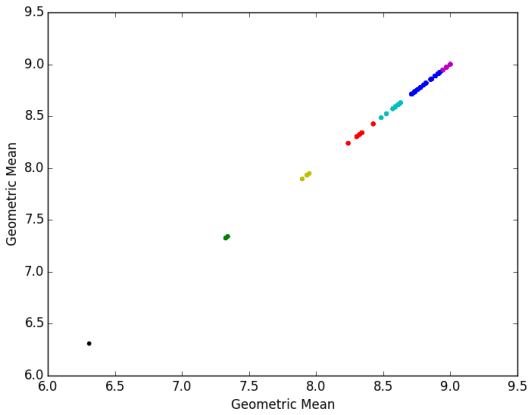
(d) K=5.



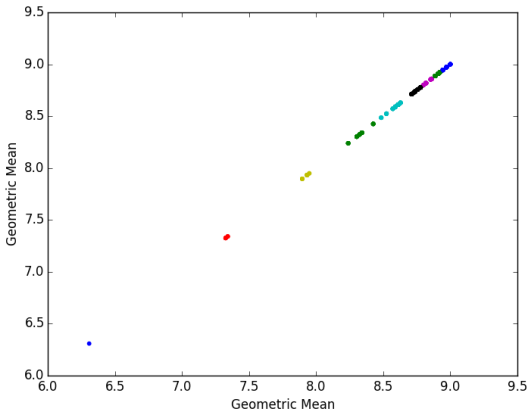
(e) K=6.



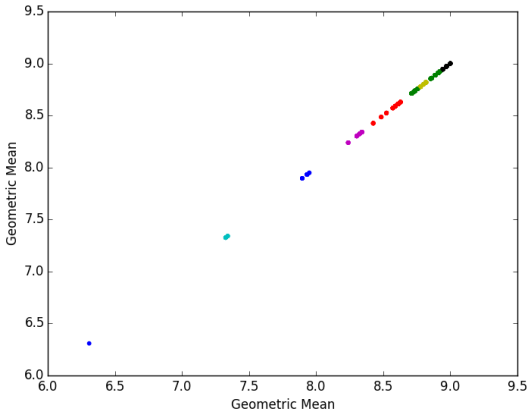
(f) K=7.



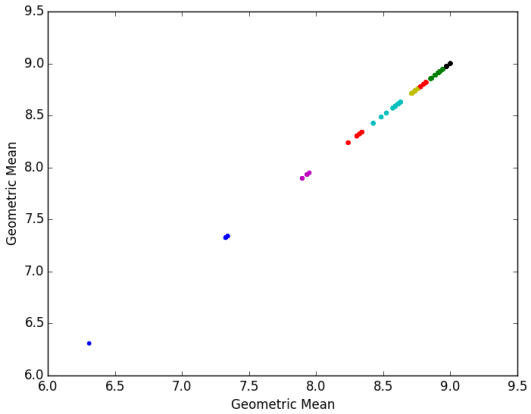
(g) K=8.



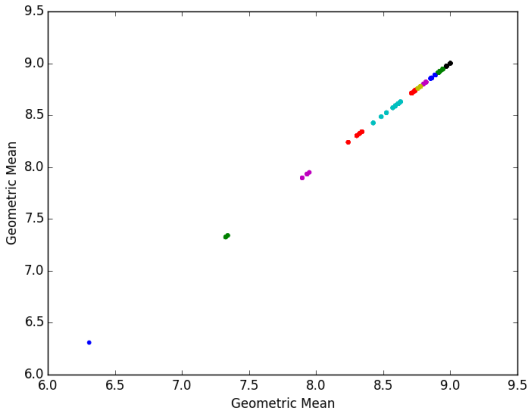
(h) K=9.



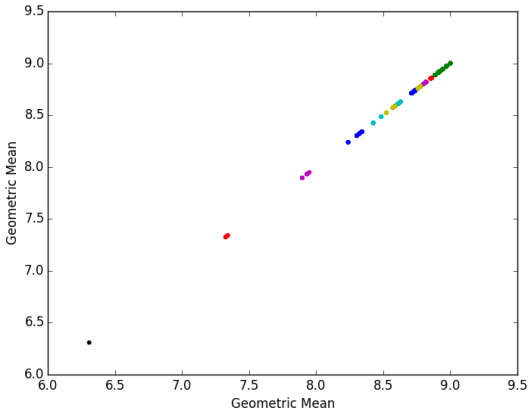
(i) K=10.



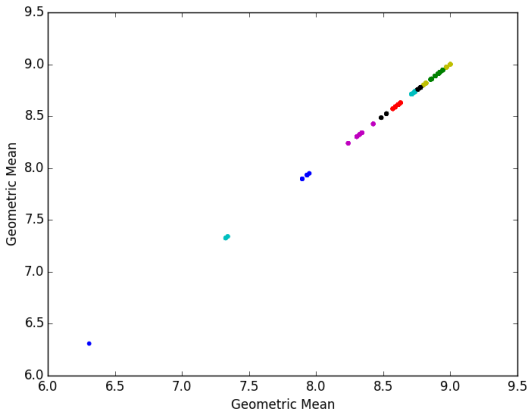
(j) K=11.



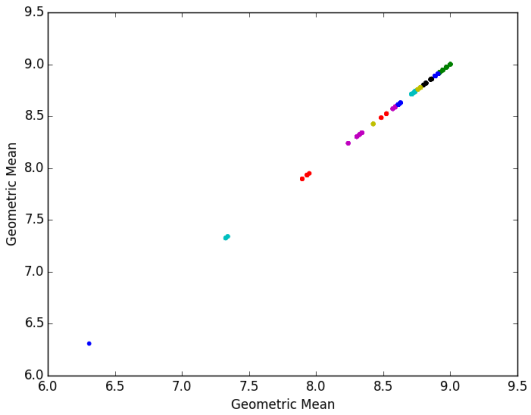
(k) K=12.



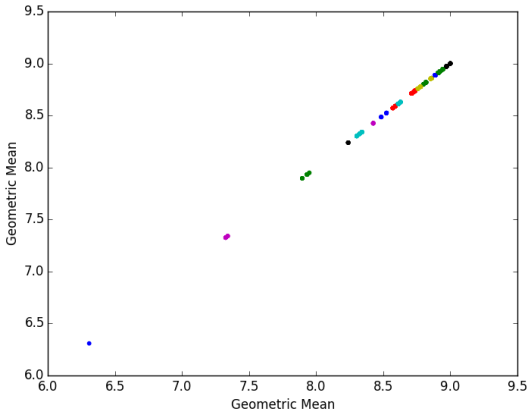
(l) K=13.



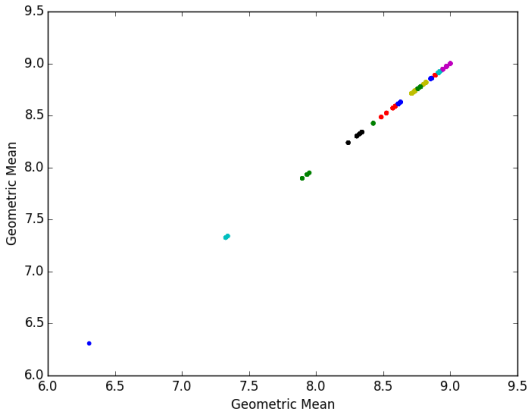
(m) K=14.



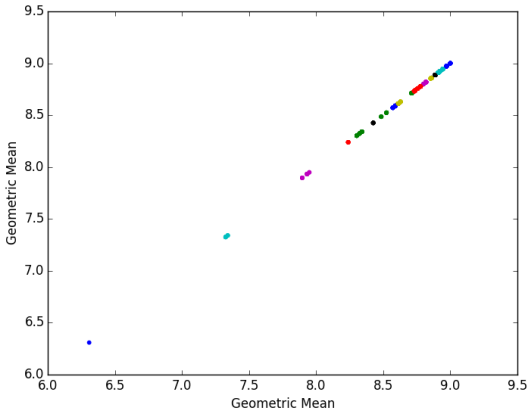
(n) K=15.



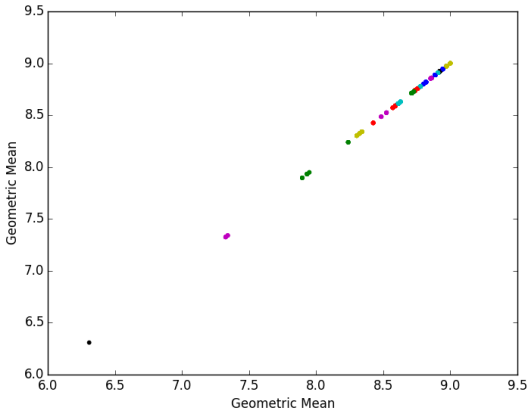
(o) K=16.



(p) K=17.

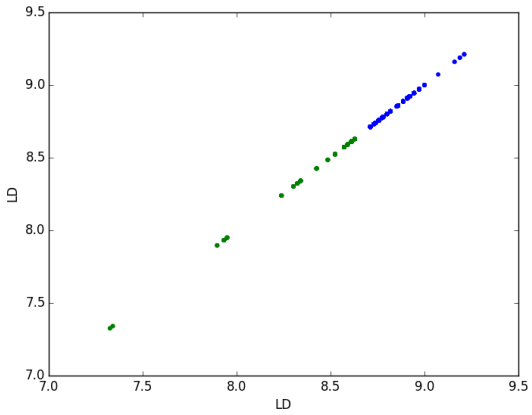


(q) K=18.

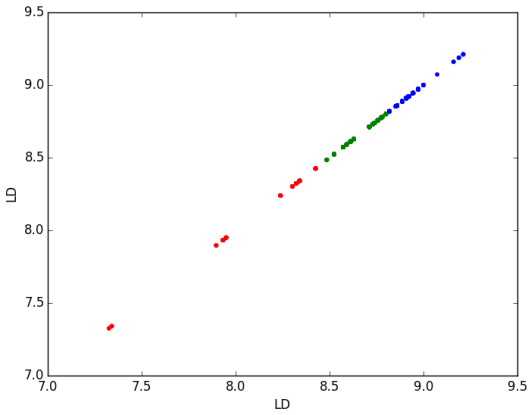


(r) K=19.

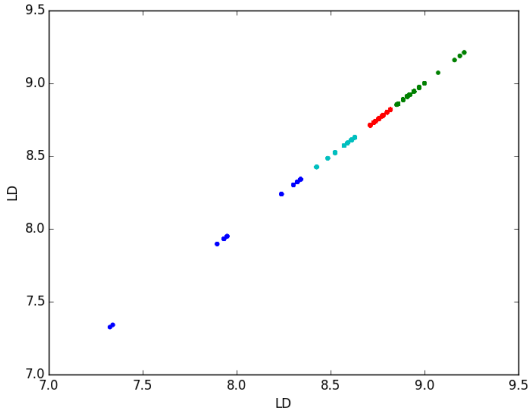
Figure 24. Cluster Plots of $\mathcal{A}$: Geometric Mean.



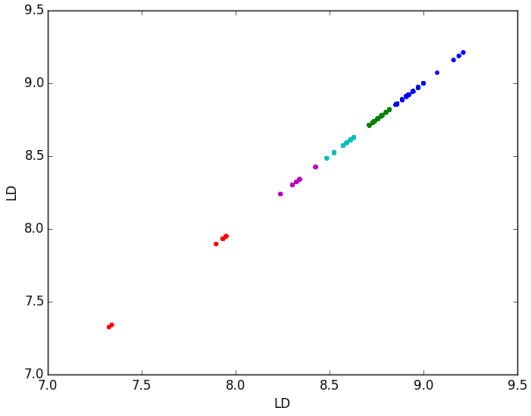
(a) K=2.



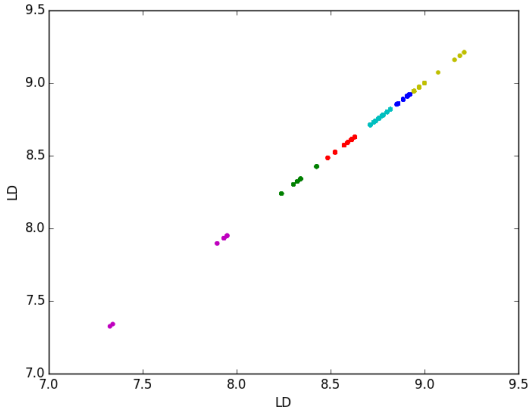
(b) K=3.



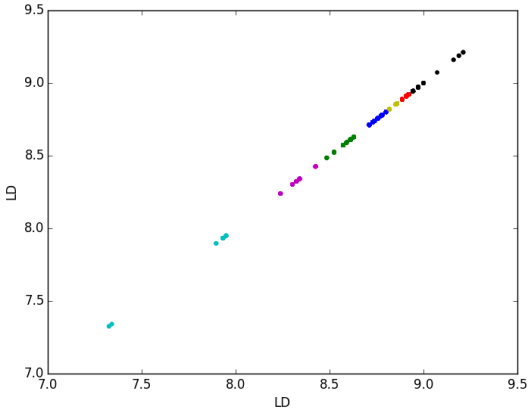
(c) K=4.



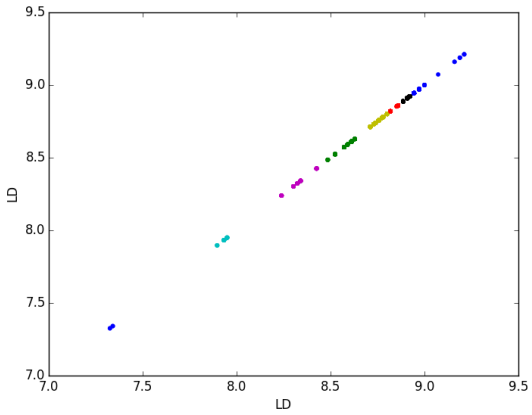
(d) K=5.



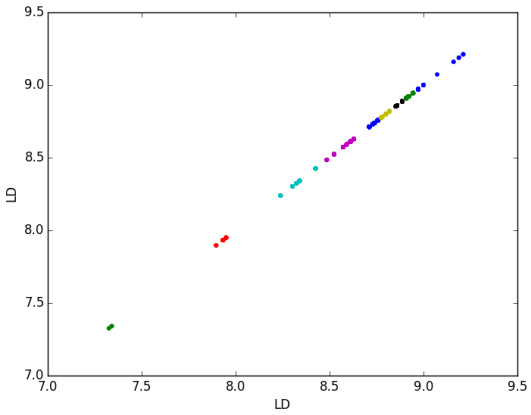
(e) K=6.



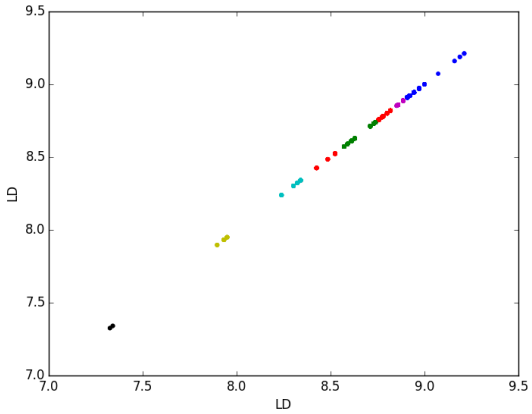
(f) K=7.



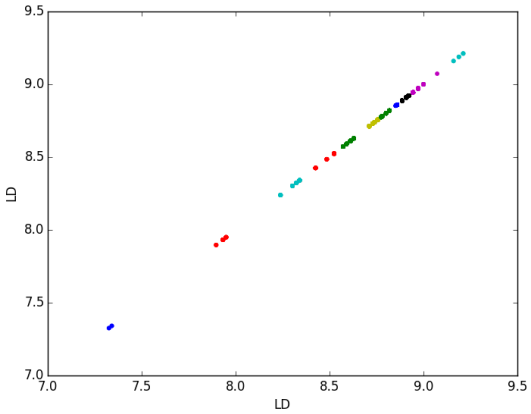
(g) K=8.



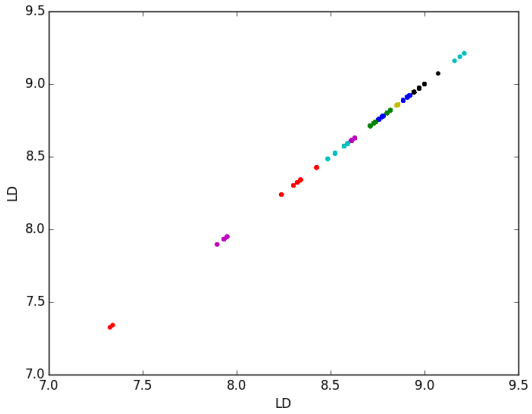
(h) K=9.



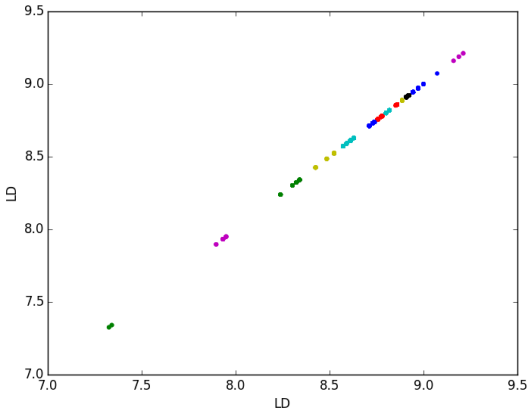
(i) K=10.



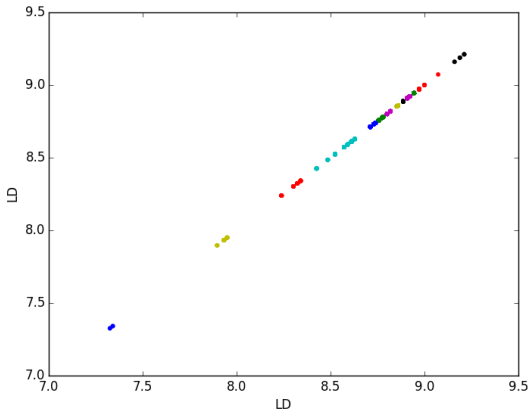
(j) K=11.



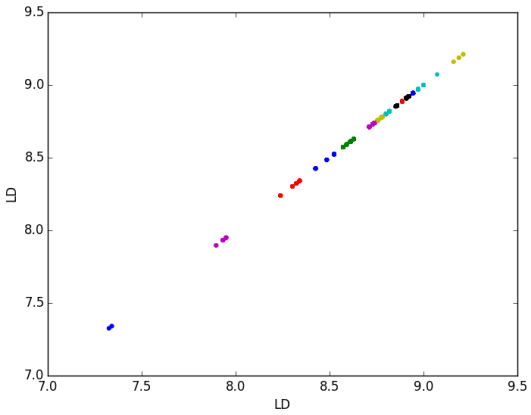
(k) K=12.



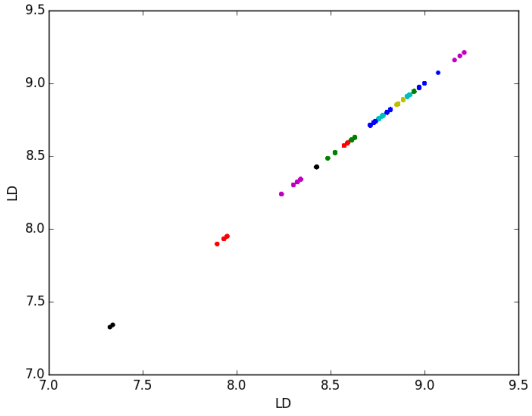
(l) K=13.



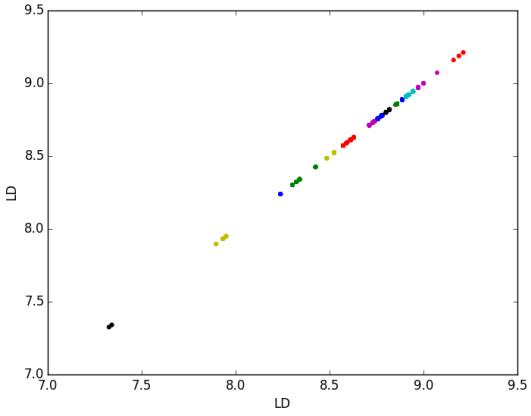
(m) K=14.



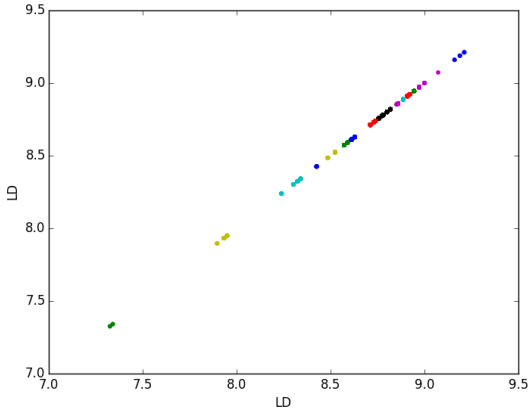
(n) K=15.



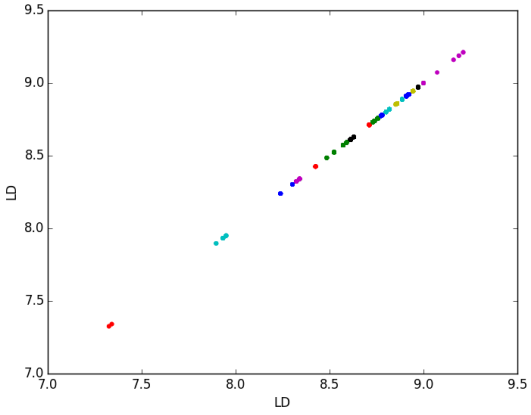
(o) K=16.



(p) K=17.



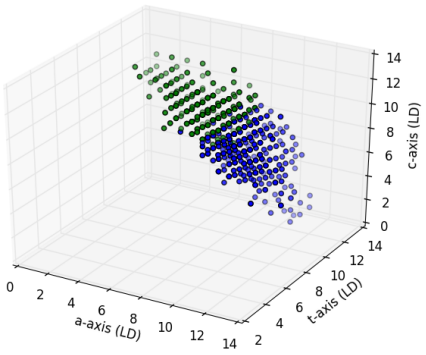
(q) K=18.



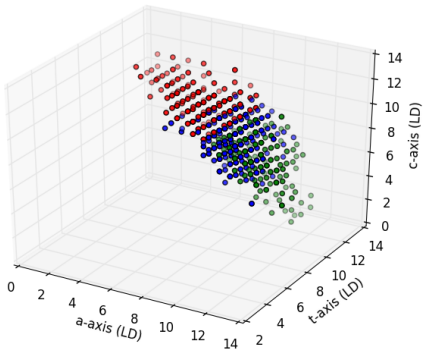
(r) K=19.

Figure 25. Cluster Plots of $\mathcal{B}$: Geometric Mean.

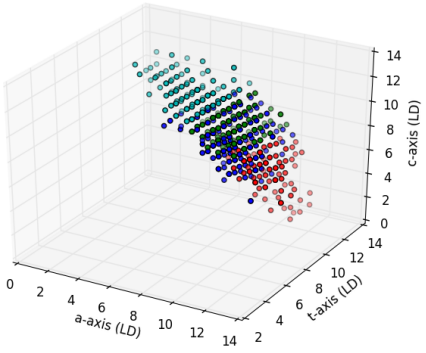
Appendix D. 3D Cluster Plots Using LD



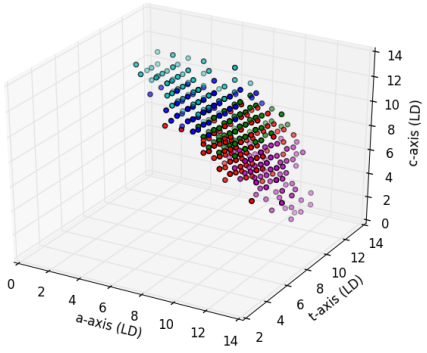
(a) K=2.



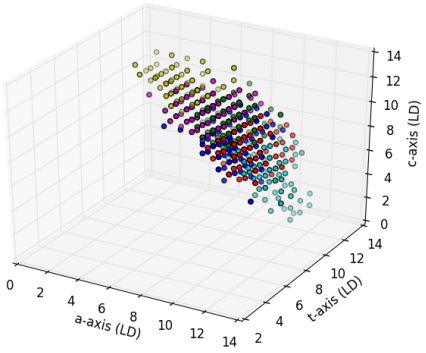
(b) K=3.



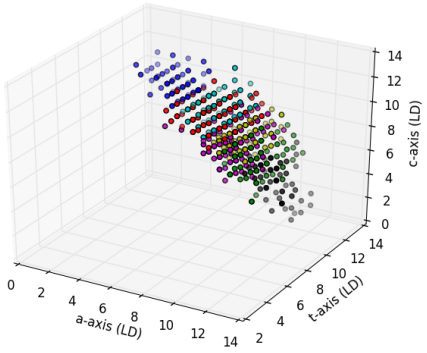
(c) K=4.



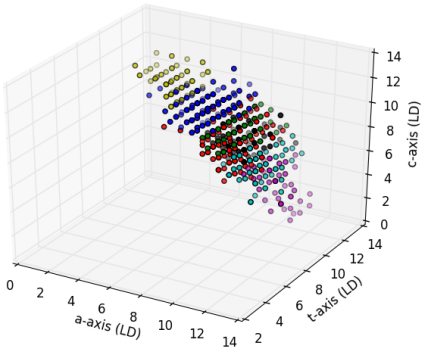
(d) K=5.



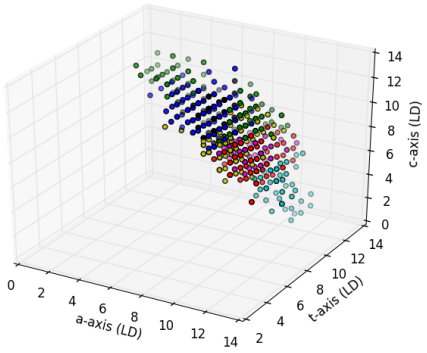
(e) K=6.



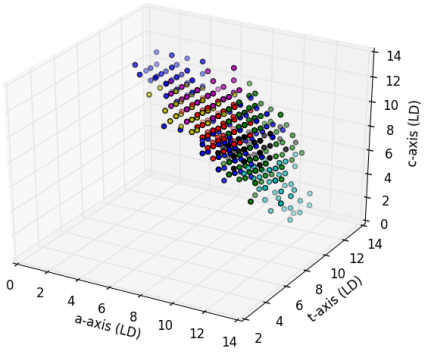
(f) K=7.



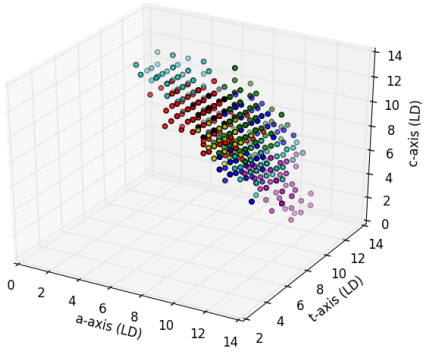
(g) K=8.



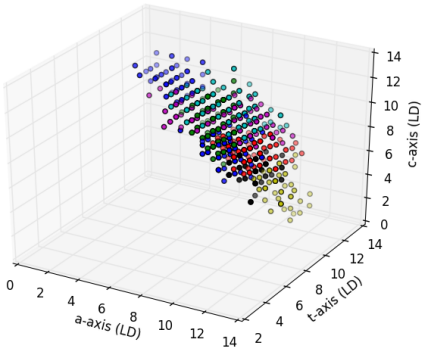
(h) K=9.



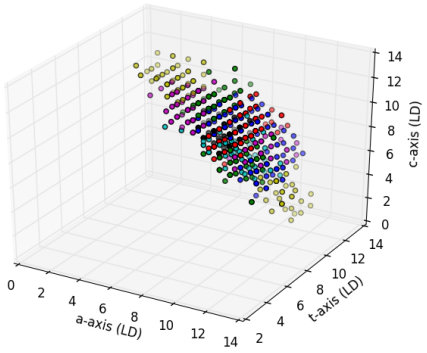
(i) K=10.



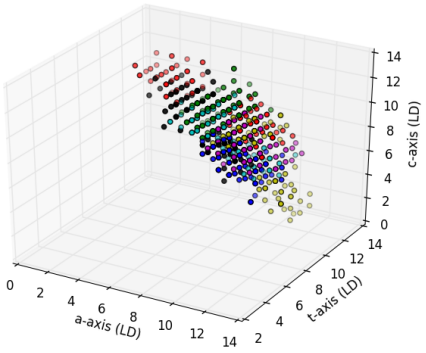
(j) K=11.



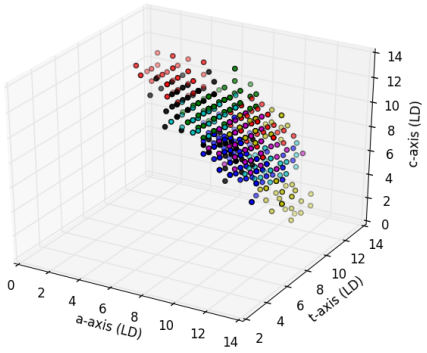
(k) K=12.



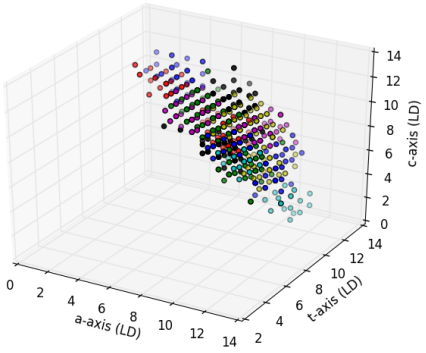
(l) K=13.



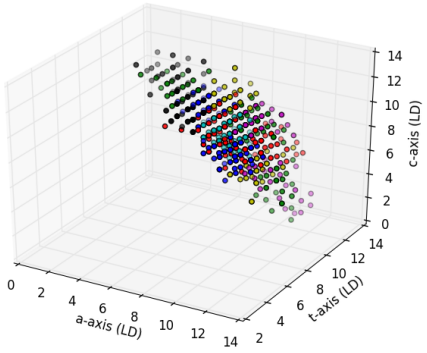
(m) K=14.



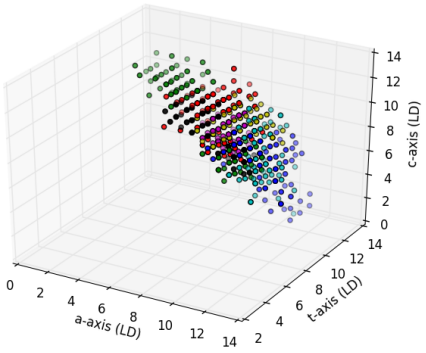
(n) K=15.



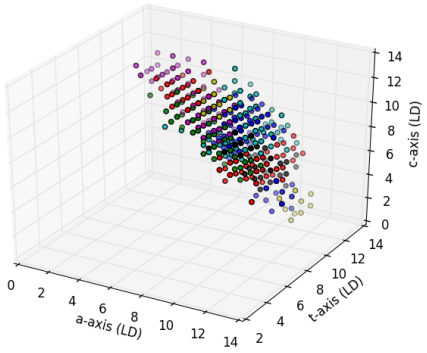
(o) K=16.



(p) K=17.

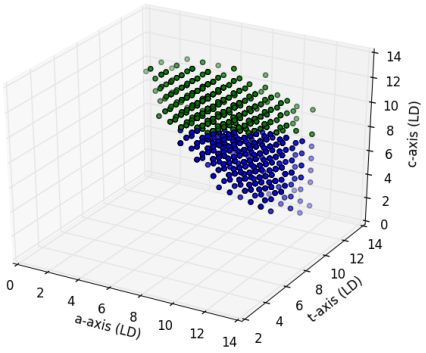


(q) K=18.

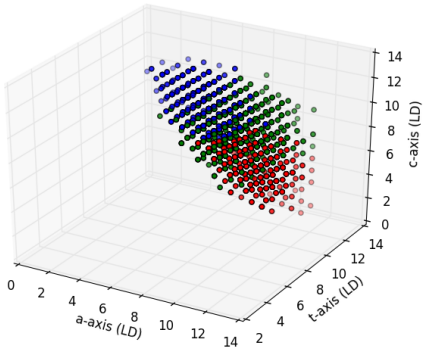


(r) K=19.

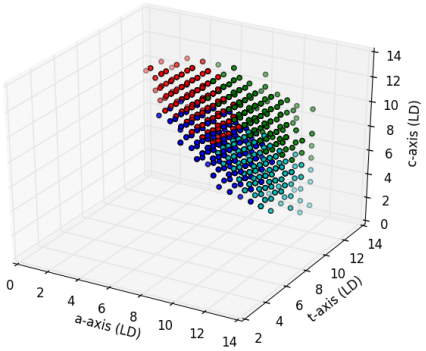
Figure 26. Cluster Plots of $\mathcal{A}$: Coordinates.



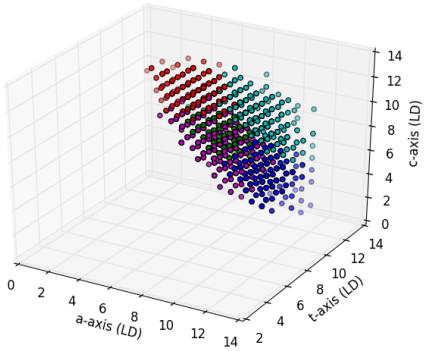
(a) K=2.



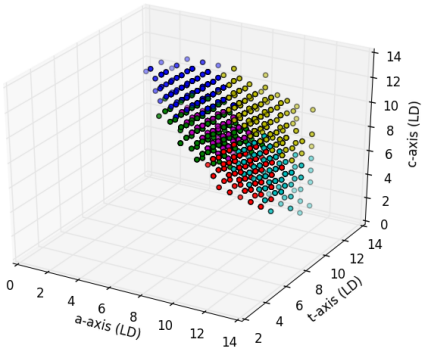
(b) K=3.



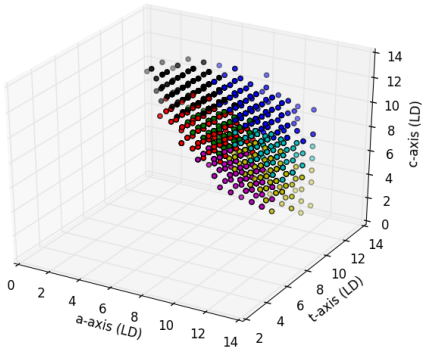
(c) K=4.



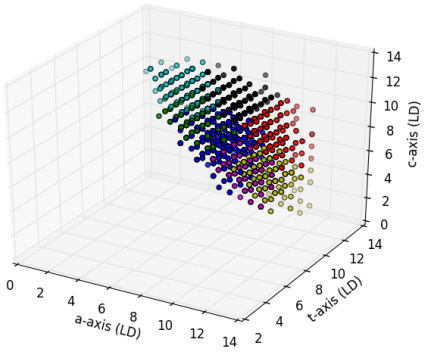
(d) K=5.



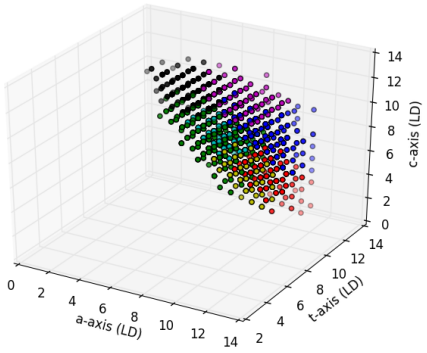
(e) K=6.



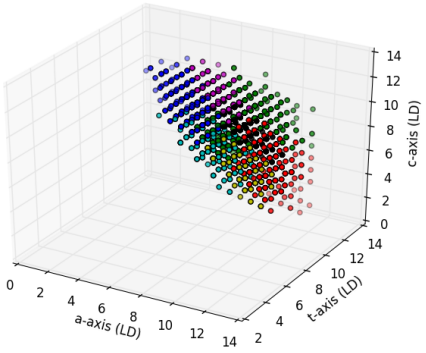
(f) K=7.



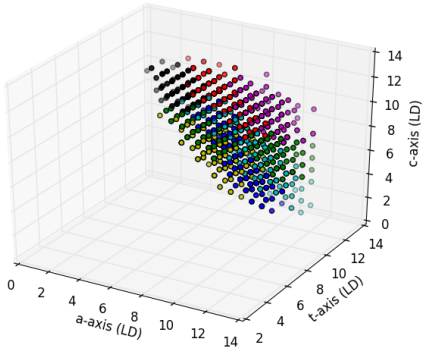
(g) K=8.



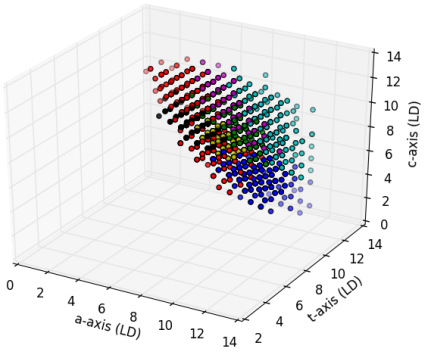
(h) K=9.



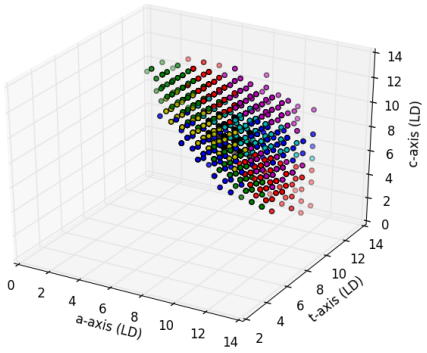
(i) K=10.



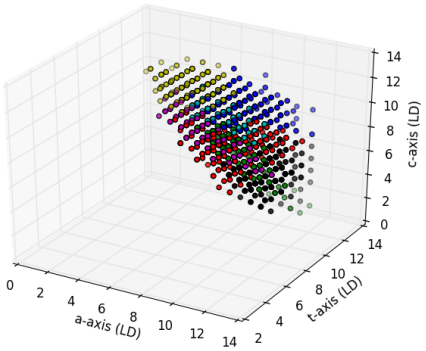
(j) K=11.



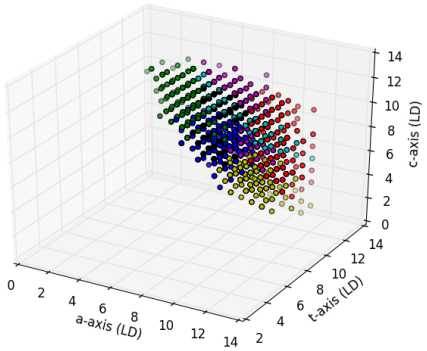
(k) K=12.



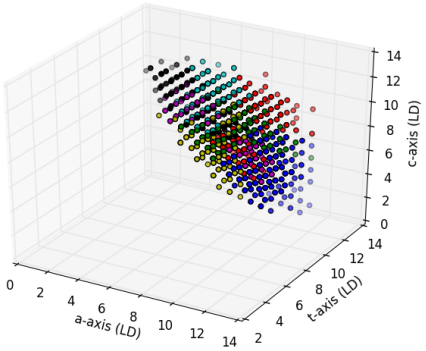
(l) K=13.



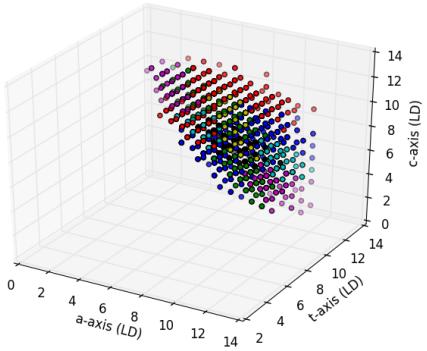
(m) K=14.



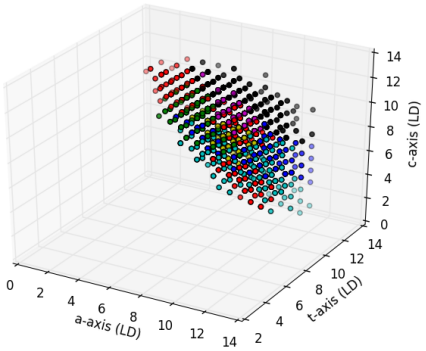
(n) K=15.



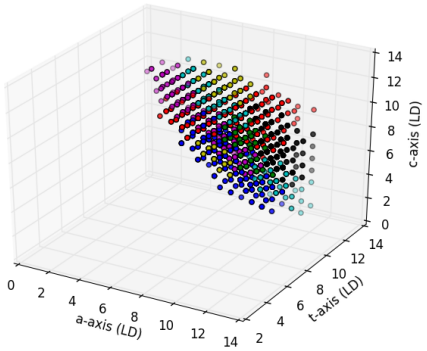
(o) K=16.



(p) K=17.



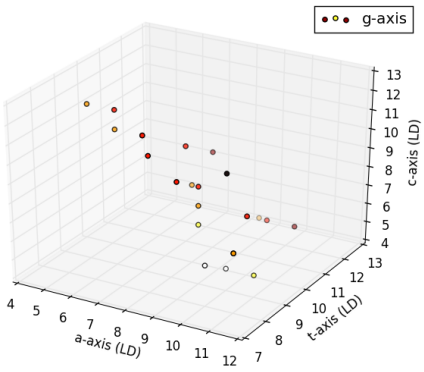
(q) K=18.



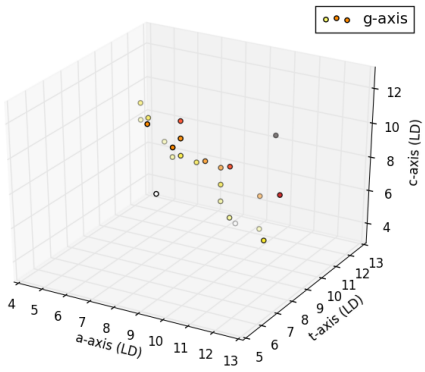
(r) K=19.

Figure 27. Cluster Plots of $\mathcal{B}$: Coordinates

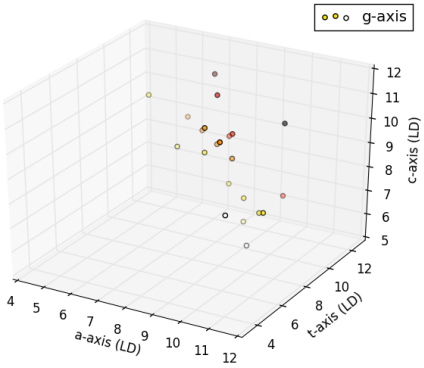
Appendix E. 3D Plots of Set Centroids



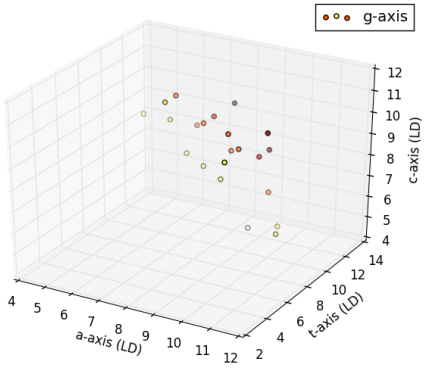
(a) Set 1, K=25.



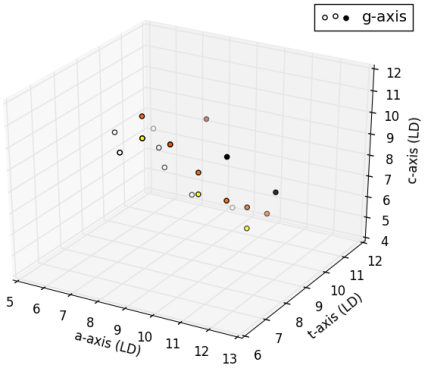
(b) Set 2, K=25.



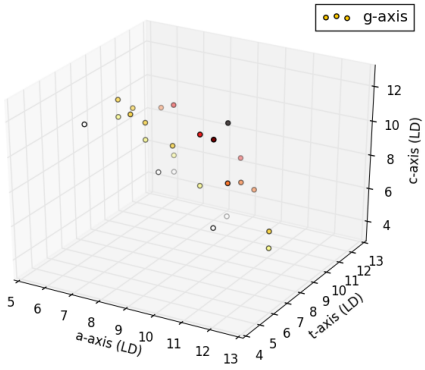
(c) Set 3, K=24.



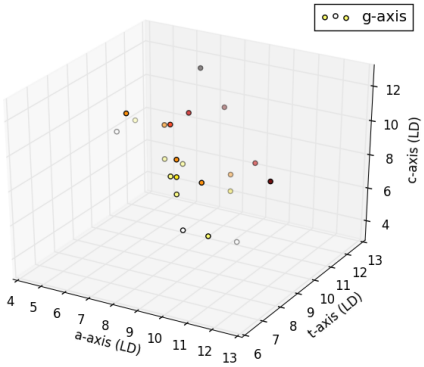
(d) Set 4, K=22.



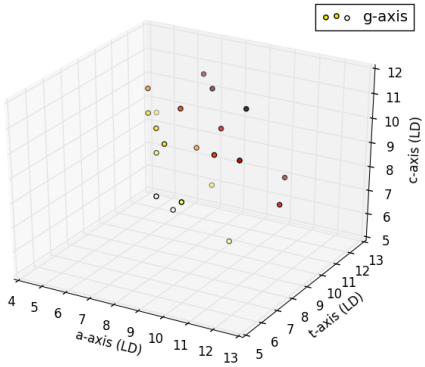
(e) Set 5, K=24.



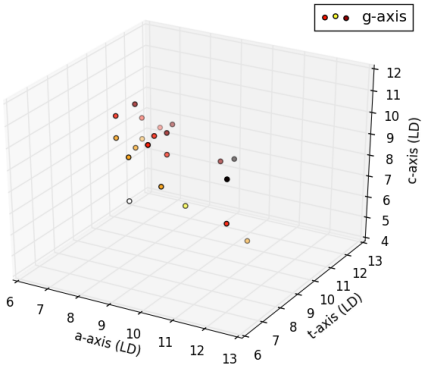
(f) Set 6, K=26.



(g) Set 7, K=24.



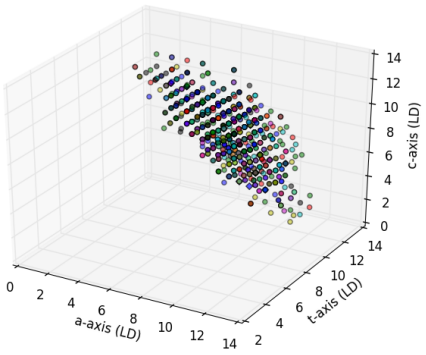
(h) Set 8, K=21.



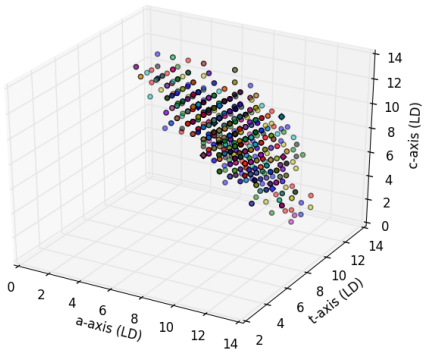
(I) Set, 9 K=25.

Figure 28. Set Centroid Plots.

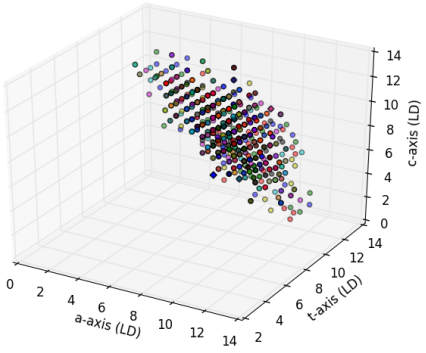
Appendix F. 3D Cluster Plots Using LD with Set Centroids



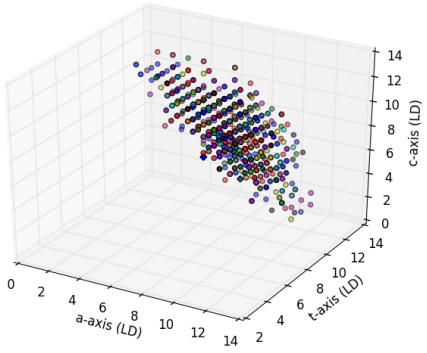
(a) Set 1, K=25.



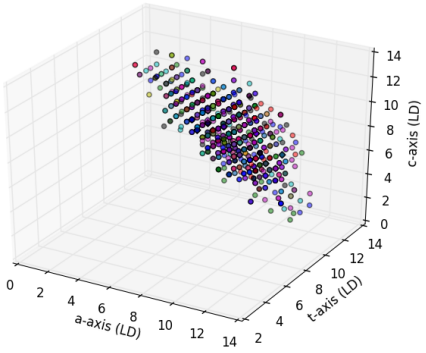
(b) Set 2, K=25.



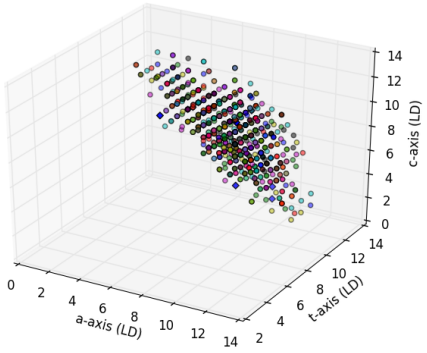
(c) Set 3, K=24.



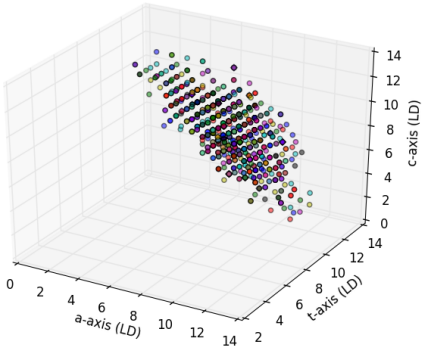
(d) Set 4, K=22.



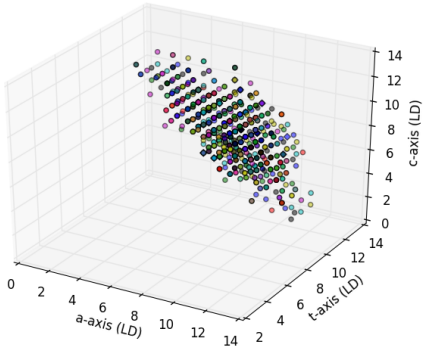
(e) Set 5, K=24.



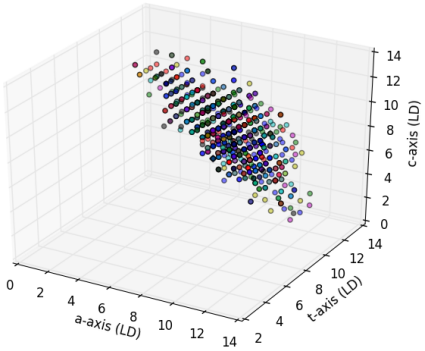
(f) Set 6, K=26.



(g) Set 7, K=24.

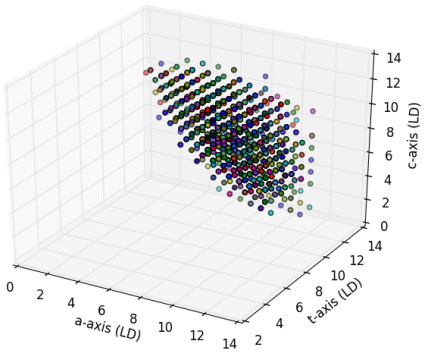


(h) Set 8, K=21.

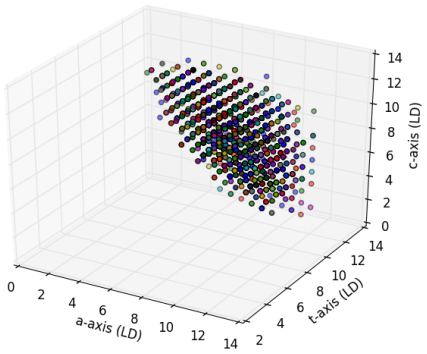


(I) Set, 9 K=25.

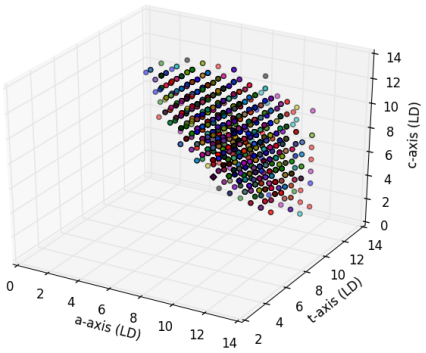
Figure 29. Cluster Plots of $\mathcal{A}$ with Set Centroids.



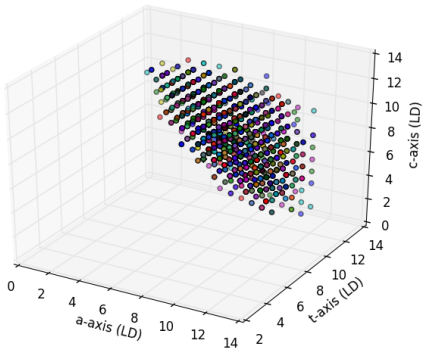
(a) Set 1, K=25.



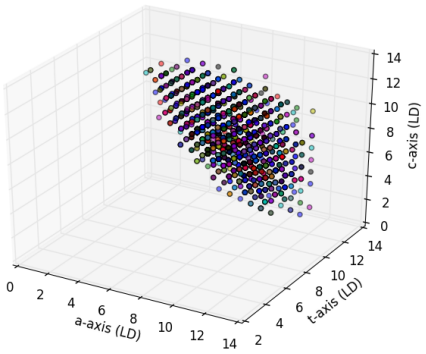
(b) Set 2, K=25.



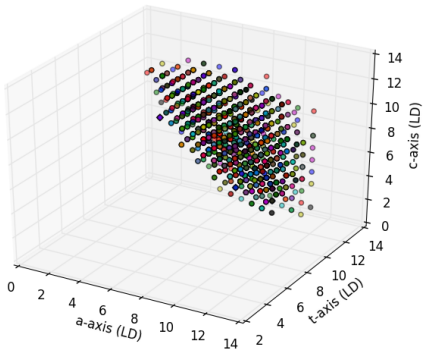
(c) Set 3, K=24.



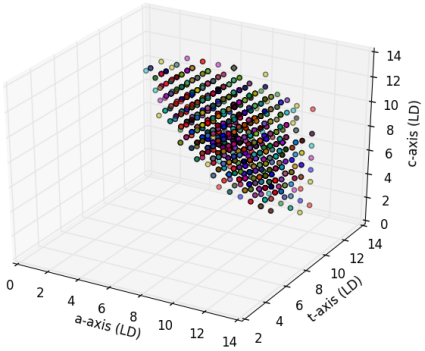
(d) Set 4, K=22.



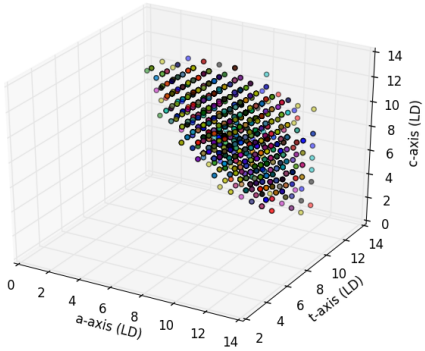
(e) Set 5, K=24.



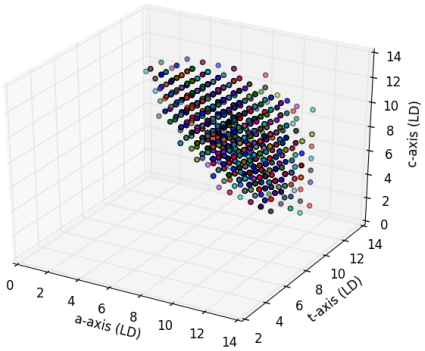
(f) Set 6, K=26.



(g) Set 7, K=24.



(h) Set 8, K=21.



(I) Set, 9 K=25.

Figure 30. Cluster Plots of $\mathcal{B}$ with Set Centroids.

Bibliography

1. "What is genotype? what is phenotype?." World Wide Web Page. Available at *https://pged.org/what-is-genotype-what-is-phenotype/*.

2. A. Kusiak and S. Shah, "Cancer gene search with data-mining and genetic algorithms," *Computers in biology and medicine*, vol. 37, no. 2, pp. 251–261, 2007.

3. J. Venter, "The sequence of human genome," *Science Magazine*, vol. 291, pp. 1304–1351, 2001.

4. Q. Shen, W. Shi, and W. Kong, "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," *Computational Biology and Chemistry*, vol. 32, pp. 53–60, 2007.

5. C. Bertrand, E. Fokoue, and H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*. New York, NY: Springer, 2009.

6. T. Brown, *Genomes Sequencing*. Manchester, UK: Oxford, 2002.

7. Y. Fei, *DNA Sequencing, Sanger and Next-Generation Sequencing*. College Publishing, 2008.

8. "An introduction to next-generation sequencing technology," tech. rep., Illumina, 2016.

9. S. Behjati and P. Tarpey, "What is next generation sequencing," *Arch Dis Child Educ Pract Ed*, vol. 98, pp. 236–238, 2013.

10. "Quality scores for next-generation sequencing," tech. rep., Illumina, 2011.

11. Q. Song, S. Merajver, and J. Li, "Cancer classification in the genomic era: Five contemporary problems," *Human Genomics*, vol. 9, no. 27, 2015.

12. M. Cusnir and L. Cavalcante, "Inter-tumor heterogeneity.," *Human Vaccines & Immunotherapeutics*, vol. 8, no. 8, pp. 1143–1145, 2012.

13. A. L. Bazinet and M. p. Cummings, "A comparative evaluation of sequence classification programs," *BMC Bioinformatics*, vol. 13, no. 92, 2012.

14. "Estimating sequence coverage.," tech. rep., Illumina, 2014.

15. "A highly sensitive method for measuring gene expression from single cells." World Wide Web Page. Available at *https://www.illumina.com/techniques/sequencing/rna-sequencing/ultra-low-input-single-cell-rna-seq.html* .

16. "Transcriptome." World Wide Web Page. Available at *https://www.genome.gov/13014330/transcriptome-fact-sheet/ /*.

17. I. Macaulay, C. Ponting, and T. Voet, "Single-cell multiomics: Multiple measurements from single cells," *Trends in Genetics*, pp. 155–168 volume = 33, number = 2, year = 2012.

18. K. Korthauer, L.-F. Chu, *et al.*, "A statistical approach for identifying differential distributions in single-cell rna-seq experiments.," *Gene Biology*, vol. 17, no. 222, pp. 1143–1145, 2016.

19. S. J. Altschuler and L. F. Wu, "Cellular heterogeneity: Do differences make a difference?," *Cell*, vol. 141, pp. 559–563, 2010.

20. X. Dai, T. Li, *et al.*, "Breast cancer intrinsic subtype classification, clinical use and future trends," *American Journal of Cancer Research*, vol. 5, no. 10, pp. 2929–2943, 2015.

21. T. Sørlie, R. Tibshirani, *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets.," *Proceedings of the National Academy of Science*, vol. 100, pp. 8418–8426, 2003.

22. K. Hoffman and R. Kunze, *Linear Algebra*. Englewood Cliffs, NJ: Prentice Hall, 1971.

23. E. Candés and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 21, pp. 22–30, 2008.

24. M. Davenport, M. Duarte, *et al.*, "Introduction to compressed sensing." World Wide Web Page, 2009.

25. H. Rauhut, "Compressive sensing and structured random matrices." World Wide Web Page. Available at *http://www.mathc.rwth-aachen.de/ rauhut/files/LinzRauhut.pdf*.

26. S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY: Springer, 2010.

27. A. S. Bandeira, E. Dobriban, *et al.*, "Certifying the restricted isometry property is hard," *arXiv:1204.1580*, vol. v2, 2012.

28. A. S. Bandeira, M. Fickus, *et al.*, "The road to deterministic matrices with the restricted isometry property," *arXiv:1202.1234*, vol. v1, 2012.

29. L. Granai and P. Vandergheynst, "parse decomposition over multi-component redundant dictionaries." Multimedia Signal Processing, 2004 IEEE 6th Workshop, 2004.

30. D. Mixon, W. U. Bajwa, and R. Calderbank, "Frame coherence and sparse signal processing," *arxiv1105.4279*, vol. v1, 2011.

31. W. U. Bajwa, R. Calderbank, and D. Mixon, *Two are better than one: Fundamental parameters of frame coherence.*

32. S. e. a. Kaur, "Mining text using levenshtein distance in hierarchical clustering," *International Journal of Computer Techniques*, vol. v2, no. I, pp. 92–97, 2015.

33. D. Katz, J. Baptista, *et al.*, "Obtaining confidence intervals for the risk ratio in cohort studies," *Biometrics*, vol. 34, no. 3, pp. 469–474, September 1978.

34. T. T. Cai and A. Zhang, "Sharp rip bound for sparse signal and low-matrix recovery," *arXiv:1302.1236*, vol. v1, 2013.

35. World Wide Web Page. Available at *https://www.ncbi.nlm.nih.gov/bioproject?LinkName=biosample_bioproject&from_uid=120456.*

36. H. Huang and C. Yu, "Clustering dna sequences using the out-of-place measure with reduced n-grams," *Journal of Theoretical Biology*, vol. 406, pp. 61–72, 2016.

37. T. E. of Encyclopedia Britannica, "Ribosomal rna." World Wide Web Page. Available at *https://www.britannica.com/science/ribosomal-RNA.*

38. World Wide Web Page. Available at *https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4794.*

39. T. Sato, A. Kaneda, *et al.*, "Prc2 overexpression and prc2-target gene repression relating to poorer prognosis in small cell lung cancer," *Scientific Reports*, vol. 3, no. 1911, 2013.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
27–08–2018	PhD Thesis	Oct 2015 — Sept 2018

4. TITLE AND SUBTITLE

COMPRESSIVE SAMPLING FOR PHENOTYPE CLASSIFICATION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Eric L. Brooks

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/ENC/DS/18-S001

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of Engineering Physics
2950 Hobson Way
WPAFB OH 45433-7765
DSN 271-0690, COMM 937-255-3636
Email: eric.broks@afit.edu

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Phenotype classification has become an increasingly important genomic research method for disease identification and treatment. Phenotype classification is the investigation into the genetic information concerned with locating biomarkers (features) in order to identify an observed effect. The primary challenge associated with phenotype classification is with analyzing the data due to the inherent high-dimensionality of DNA data. As a result, phenotype classification faces challenges with feature selection, and consequently, classification accuracy. This research developed a methodology to alleviate these challenges while improving classification performance. The methodology leverages concepts of compressive sampling, to arrive at a process that identifies features most relevant to the phenotype. Additionally, this research presents a probabilistic acceptance of the RIP and uses it to qualify dataframes constructed by the proposed methodology. Overall, I found this methodology as a viable approach to dimension reduction and feature selection, which improved phenotype classification accuracy.

15. SUBJECT TERMS

Compressive Sensing, Compressive Sampling, Genomics, Phenotype classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Alan Lair, AFIT/ENC
U	U	U	U	85	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4538; eric.brooks@afit.edu