METHODS FOR STATISTICAL ASSOCIATION MINING BY VARIABLE-TO-SET AFFINITY TESTING

Kelly Nicole Bodwin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2017

Approved by:

Andrew B. Nobel

Kai Zhang

J.S. Marron

Shankar Bhamidi

Yin Xia

ABSTRACT

KELLY BODWIN: Methods of Association Mining by Variable-to-Set Affinity Testing
(Under the direction of Andrew B. Nobel and Kai Zhang)

Statistical data mining refers to methods for identifying and validating interesting patterns from an overabundance of data. Data mining tasks in which the objective involves pairwise relationships between variables are known as *association mining*. In general, features sought by association mining methods are sets of variables, often small subsets of a larger collection, that are more associated internally than externally. Methods vary in both the measure of association that is studied and the algorithm by which associated sets are identified. This dissertation discusses provide a generalized framework for association mining called Variable-to-Set Affinity Testing (VSAT). Unlike conventional techniques for clustering or community detection, which usually maximize a score from a dissimilarity or adjacency matrix, the VSAT approach is an adaptive procedure grounded in statistical hypothesis testing principles. The framework is adaptable to a broad class of measurements for variable relationships, and is equipped with theoretical guarantees of error control.

This dissertation also presents in detail two new association mining methods built in the VSAT framework. The first, Differential Correlation Mining (DCM), identifies variable sets that have higher average pairwise correlation in one sample condition than in another. Such artifacts are of scientific interest in many fields, including statistical genetics and neuroscience. Differential Correlation Mining is applied to high-dimensional data sets in these two fields. The second method, Coherent Set Mining (CSM), is a novel approach to association mining in binary data. Dichotomous observations are assumed to derive from a latent variable of interest via thresholding. The Coherent Set Mining method identifies variable sets that are strongly associated in the latent measure, despite distortions in the association structure of the observed data due to the thresholding process. Coherent Set Mining is applied to problems in text mining, statistical genetics, and product recommendation.

"The unpredictable and the predetermined unfold together to make everything the way it is. It's how nature creates itself, on every scale, the snowflake and the snowstorm. It makes me so happy. To be at the beginning again, knowing almost nothing. [...] When you push the numbers through the computer you can see it on the screen. The future is disorder. A door like this has cracked open five or six times since we got up on our hind legs. It's the best possible time to be alive, when almost everything you thought you knew is wrong."

- Valentine Coverly, *Arcadia* (Tom Stoppard, 1993)

ACKNOWLEDGEMENTS

I completed this work all by myself without anyone's help or support. Wait - hold on - that's not right. The truth is the complete opposite. This dissertation and I owe an infinite debt of gratitude to a great many people.

First and foremost, to my advisors **Andrew Nobel** and **Kai Zhang**: Thank you so much for your wisdom and guidance in the past five years. I wasn't always the easiest student, and I truly appreciate the patience, flexibility, and genuine care you have shown me. Andrew, you showed confidence in me before anyone else did. Your attention to detail (including your *occasionally* superior grammar), your clarity of thought, and your statistical intuition have made me a much better writer and researcher. Kai, it has meant so much to me that you have always found time for me - whether for research help or for words of advice - during what were probably some of the busiest years of your life.

I am fortunate to have benefitted from fantastic UNC faculty and staff beyond my official advisors, particularly **Shankar Bhamidi**, **Robin Cunningham**, **Steve Marron**, and **Yin Xia**. Shankar, I can't thank you enough for your constant support and advocacy. The department is lucky to have you. Robin, your commitment to great teaching inspires me. Thank you so encouraging me and helping me improve. Steve, thank you for always showing interest my work - it was a pleasure to get to discuss ideas and applications with you every time our research intersected. Yin, I am grateful for your presence on my committee, especially since you generously signed on in your first year here.

The journey to this dissertation began long before I came to UNC. I am forever grateful to my early mentors: **Joe Blitzstein**, **Colin Anderson**, **Jeff Rosenthal**, and **Albert Yoon**. Blitzy, you taught me to love statistics, to love research, and to love teaching. I was, and still am, so lucky to have your guidance and your friendship. Ando, I have to admit that I hated Statistics AP. Sorry. But you're the reason I love math, and your jokes are sort of funny sometimes, I guess. Jeff and

Albert, thank you for taking a chance on a new and inexperienced research assistant. I learned so much from working with you.

Without my fellow UNC students, I would never have been able to navigate this experience. I am especially indebted to **John Palowitch** (with whom Chapter 2 is joint work), **Jimmy Jin**, **Suman Chakraborty** (who contributed substantially to the work in Chapter 4), and **James Wilson**. John, you are my favorite person on the planet collaborate with. You are dedicated and thorough in ways I aspire to be, and our discussions (about statistics and about life) have expanded my horizons. Jimmy, our graduation doesn't mean I will never again walk into your office uninvited and demand that you join me for coffee and complaining. Suman, thanks for sharing an office, a project, and pots of tea. James, your mentorship and friendship were invaluable to my growth in graduate school. Also, to Anna, Zane, Alex, Sepehr, and Rosie: I am so glad you are all a part of my life, and I hope our paths will cross again soon (probably at Wine Bar).

Finally, as in all things I am unbelievably grateful to my family, and to the friends who have become family. I love you all so much. **Mom and Dad**, I apologize for spending a decade on the opposite side of the country. I hope that my attempts to emulate you in all things (namely Google Calendar and terrible jokes) show you how much I look up to you, and I hope that I have made you proud both personally and professionally. **Greg**, you're my favorite human, but also: Ha-ha, I got my PhD first. To **Becka, Eric, Carol, and Aurora**: Suffice to say, I would never have finished this dissertation without you, and I am unbelievably lucky to have been welcomed into your family. Lastly, **Kirkland House**, past and present, you will always be home.

FUNDING AND DATA. I would like to thank the NSF for providing me with a Graduate Research Fellowship, which funded the majority of my graduate school years (grant DGE-1144081). Without this award I would likely not have been able to pursue an academic career. Work in this dissertation was also partially funded through NSF grant DMS-1310002 to Dr. Andrew Nobel et. al. and NSF grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute.

I am also very grateful to members of the Perou Lab at UNC's Lineberger Cancer Center for supplying datasets and valuable discussion for all genetic applications. I owe particular thanks to Katie Hoadley, whose expertise was invaluable to the work in Section 3.7, and Sara Selitsky, with whom work on DCM applications is ongoing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

VSAT	Variable-to-Set Testing
DCM	Differential Correlation Mining
CSM	Coherent Set Mining
FPR	False Positive Rate
TDR	True Discovery Rate
TCGA	The Cancer Genome Atlas
d	The dimension of data, i.e., the number of variables of interest
n	The sample size, i.e., the number of d-dimensional observations
$[d], [n]$	The set of indices $\{1, \ldots, d\}$ or $\{1, \ldots, n\}$
A	A variable set $A \subset [d]$
m	The size of a variable set, $\lvert A \rvert = m$
$\zeta(j, A)$	A population association measure between variable j and set A
$S(j, A)$	A statistic for testing $\zeta(j, A) > 0$
$\mathrm{p}(j, A)$	The p-value pertaining to a test of association between j and A
$\mathcal{A}(\mathbf{X}, \alpha)$	The set of VSAT fixed points at level α for data $\mathbf{X}$
$\mathbb{X}$	An observed data matrix of dimension $n \times d$
$\tilde{\mathbb{X}}$	An appropriately standardized version of $\mathbb{X}$
$\mathbf{R}_1, \hat{\mathbf{R}}_1$	The $d \times d$ population and sample correlation matrices for data $\mathbf{X}_1$
ρ_{jk}, r_{jk}	The population and sample correlations between variables j and k
$\Delta(j, A)$	The average correlation difference between j and A
Ω	An $n \times d$ matrix of realizations of random parameters
$\mathbb{P}_0(\cdot)$	The probability under a suitable null measure
$\phi(\cdot)$	The standard Normal cdf
$\varphi(\cdot)$	The Fisher transformation
$\mathcal{L}(\cdot \mid \cdot)$	The likelihood of a parameter or set of parameters given data
$\mathbb{I}\{\cdot\}$	The indicator of an event

CHAPTER 1
Introduction

The field of statistics first developed as a way of drawing conclusions from limited observations. Today, statisticians face a nearly opposite challenge: how to glean meaningful information from an ever-growing supply of data. The term "data mining", once a controversial epithet for questionable practices, now refers optimistically to the practice of extracting valuable material that has been buried in debris. In general, data mining methods seek notable patterns among noisy observations. Many such methods are exploratory, in that they focus on discovery rather than verification. *Statistical* data mining, more specifically, makes use of modeling and testing principles both to identify patterns and to make probabilistic claims about their validity. As available data becomes higher dimensional and more complex, approaches to problems of statistical data mining must continue to adapt in response.

This dissertation introduces novel statistical methods for the branch of data mining known as *association mining*. In general, association mining is concerned with detecting relational (or *second-order*) structure between variables. For example, a company might study association between its employees, with the goal of identifying distinct social subgroups or of understanding how low-level employees interact with management. Association structures of interest vary by data type and by research question. Most commonly, association mining targets take the form of subsets of variables that are either strongly internally associated or all associated with a common external feature. The work in this dissertation is specifically concerned with the former.

The discussion thus far has used the term "association" in the broadest possible sense to mean any quantification of a relationship between two variables. However, the choice of a specific association measure is a crucial element of any association mining method. Broadly speaking, there are two ways to measure association:

1. Distance or dissimilarity.

   Often, information about variables can be summarized in a single dissimilarity matrix, where entries represent pairwise relationships. In some cases, these matrices are calculated from data. For instance, if variables represent points in a vector space, the relationship between two points can be represented by a distance metric. In other cases, dissimilarity matrices are observed directly rather than computed. Such is the case in the analysis of networks, where available data takes the form of a set of variables (or *nodes*) and a set of links (*edges*) between the nodes. Edges may be *weighted*, taking on continuous values, or *unweighted*, consisting of 0/1 indicators of presence or absence. The measure of association between two variables is therefore the value (or presence) of the edge between those variables.

   In general, association mining based on dissimilarity measures or on network data is not statistical, in that the measures of association are treated as non-random. However, randomness can be introduced via addition of noise to dissimilarity measures or assumption of generating models. A prominent example of this is the Erdos-Renyi random graph model (Robins et al., 2007), which assumes an observed unweighted network was created by randomized edge assignments.

2. Estimators for statistical dependence.

   When data can reasonably be considered random samples from a population, it makes sense to infer relationships via estimates of parameters. Perhaps the most commonly studied parameter is the standard product-moment correlation,

$$\operatorname{cor}(X, Y) := \frac{\mathbb{E}\, XY \; - \mathbb{E}\, X\, \mathbb{E}\, Y}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

   The advantage of mining for association in the form of correlation is that two variables with nonzero correlation necessarily have nonzero dependence, and thus a "true" relationship may be said to exist. There are many alternatives to product-moment correlation, including partial correlation, rank-based correlation and covariance. It also should be noted that estimators $\hat{\beta}$ for the coefficients in a regression model $Y = \beta X + \epsilon$ may also be regarded as association measures estimating dependence.

One limitation of the correlation and related dependence measures is that they capture only linear relationships between variables. More complex dependence can be represented by summary statistics of graphical models, see e.g. Anderson (1959) Chapter 9 for further detail. Recent work has also produced useful estimators of nonlinear dependence, notably: Székely et al. (2007) defines the *distance correlation*, which is equal 0 for a variable pair if and only if the variables are independent; and Zhang (2017) proposes a distribution-free procedure for detecting dependence. Tan et al. (2002) provides a thorough overview of the many other available measures of dependence.

The methods in this dissertation are strictly statistical, and so content will primarily focus on association mining in the context of estimators for statistical dependence. Section 1.4.2 provides a more in-depth discussion of the consequences of different measures of association, as a motivation for the new measure (and corresponding mining methodology) introduced in Chapter 4.

1.1 Contributions and Outline

This dissertation consists of an in-depth treatment of two new methods for association mining: **Differential Correlation Mining (DCM)** and **Coherent Set Mining (CSM)**. Differential Correlation Mining is an algorithm for discovering variable sets that exhibit different correlation structure across two predefined sample conditions. Coherent Set Mining offers a method for mining for latent association structure from binary thresholded observations. Both methods are built on a novel algorithmic framework for mining strongly associated variable sets (or "communities"), known as **Variable to Set Affinity Testing (VSAT)**.

The remainder of this document begins with a brief overview of the major association mining methods that pertain to the topics in this dissertation and a discussion of related work to motivate the Differential Correlation Mining and Coherent Set Mining methods. Chapter 2 details in general terms the VSAT approach to association mining, and provides an important general theoretical result. Chapters 3 and 4 contain the full details, theoretical results, and real data applications for Differential Correlation Mining and Coherent Set Mining respectively. Finally, Chapter 5 closes with a final discussion and suggestions for future directions of inquiry.

1.2 Background: methods of association mining

Existing work in data mining can be characterized as either *unsupervised* or *supervised*. Unsupervised learning consists of searching for patterns in data without regard to a predictive goal. In supervised studies, datasets consist of a limited number observations for which a ground truth is known, from which one usually makes inferences about future behavior. For example, simple linear regression infers the nature of a linear relationship between a measurement x and random variable Y from observed pairs $(x_i, y_i)_{i=1}^n$. Data mining tasks can also be *semi-supervised*, in that a ground truth is known about only some of the available observations. This "training data" is then used to infer information about the remaining (or future) observations.

Perhaps the most well known class of unsupervised association mining methods is clustering algorithms. Clustering is the practice of dividing variables into groups, or *clusters*, that are highly internally associated. Typically, clustering methods seek a partition of the variables, such that each variable is assigned to exactly one cluster. In some settings, finding a partition of data is not appropriate to the research question at hand. For example, consider the problem of identifying social group memberships based on an observed Facebook friendship network. A strict partition of individuals does not capture the desired structure, since people may belong to many social groups or even none at all. Clustering tasks on networks are often referred to as community detection (cf. Fortunato (2010); Newman (2006) among others).

Although the methods of this dissertation are unsupervised, association mining also plays a role in some supervised and semi-supervised approaches. Most notable are classification and matrix completion problems. *Classification* refers to analyses of data for which there exist pre-specified categories. Typically, category labels are known for a given subset of observations, and new observations are assigned to categories based on their association with known members. *Matrix completion*, by contrast, usually assumes one has incomplete observations for a number of variables. One then "completes" the matrix by filling in missing values with estimates derived from similar variables. Chapter 5.3 discusses the use of association mining in matrix completion problems and suggests a future direction for improvement on these algorithms.

A thorough overview of the broad area of unsupervised association mining is provided by Everitt et al. (2011). This section provides a basic introduction to two of the most common methods of

clustering, which will provide a standard basis of comparison for the new methods discussed in this thesis, as well as for an area of supervised association mining that has a close relationship to the methods in this dissertation.

- k-means clustering.

  The k-means clustering algorithm consists of a simple iterative update to minimize an objective. The algorithm begins with a randomly chosen k data observations, designated as "centroids". The remaining observations are then assigned membership in one of the k clusters characterized by the centroids, in such a way as to minimize the total sum of squared error for the partition. The centroids are then updated to be the geometric centers of each of these clusters, and the process is repeated until convergence. The k-means method is similar to the common K-nearest neighbors approach, in that both cluster objects around a centroid. However, K-nearest neighbors analyses do not produce a partition, but rather study individual target objects via the K closest objects.

  There are two main reasons for the ubiquity of the k-means method. First, it is extremely computationally efficient, and can be run quickly and without large memory demands even for very high dimensional data. Second, it has a close tie to *dimension reduction* techniques. When a dataset consists of observations in d dimensions, it is common to use Principal Component Analysis (PCA) to reduce the dimension before applying k-means. The practice of dimension reduction before clustering includes a class of methods known as *spectral clustering*. PCA and k-means share a unique relationship even among spectral clustering methods due to a theoretical link between clusters centers and dimensions (Ding and He, 2004).

- Hierarchical clustering.

  Broadly, hierarchical clustering refers to the practice of progressively joining or separating variables according to some criteria. This process can be *agglomerative* (or "bottom up"), where variables begin as singletons and are merged sequentially, or *divisive* ("top-down") where the set of variables is split many times. Hierarchical clustering is extremely flexible, as any choice of metric (or *linkage* criteria) can be used to determine thresholds for joining

or splitting variable sets. For example, Section 4.5.1 compares the results of hierarchical clustering for a variety of choices of linkage metrics.

The output objects of a hierarchical clustering algorithm are *dendrograms* indicating at what height, or value of the linkage criteria, variable sets were divided/merged. To select a particular partition of the data, one typically determines a cutoff height of the dendrogram. When hierarchical clustering appears in this thesis, as a basis for comparison in simulation studies, we circumvent this problem by comparing our methods only to the best possible choice of cutoff as determined by "oracle" information about the true nature of simulated data. However, in practice, the decision about where to cut a dendrogram has enormous influence on the results. Contributions to the field of hierarchical clustering generally involve suggested algorithms for cutting a dendrogram for a particular data setting and linkage criteria.

- Variable selection by penalized regression.

  As a rule, data mining is descriptive rather than predictive; that is, its primary purpose is to use observations to identify structure in variables, rather than to forecast future observations. Regression analysis, on the other hand, is commonly used for prediction. Recent work has adapted regression principles to methods of variable selection. Perhaps the most notable examples are the penalized regression techniques of Tibshirani (1996) and Zou and Hastie (2005), which use an L_1 and L_2 penalty (respectively) to forcibly reduce the number of explanatory variables incorporated in the model. In many applications, researchers are more interested in which covariates are selected for inclusion rather than the predictive power of the model. For instance, in statistical genetics, one may use penalized regression to determine which genes among thousands are most correlated with a particular phenotypic response.

  Although regression-based variable selection is indeed an example of mining for association structure, it differs from the focus of this dissertation in its directionality. To apply penalized regression, one must first designate a response variable and many possible explanatory variables. Selected covariates represent those variables that are *associated with the response*. By contrast, our focus in this work is the identification of variable sets that are internally associated *with each other*; that is, we are interested in association mining from a single collection of variables without regard to a response.

1.3 Differential Correlation Mining

In many statistical problems, one has two datasets that measure the same variables under different conditions. It is common in the analysis of such data to assume that the samples in each dataset are generated from two underlying distributions. Even when the data is high dimensional, differences between the distributions may be present for only a small number of variables, and it is often of interest to identify these key variables. Most often, differential behavior between sample groups is measured by *first-order* statistics, which are functions of a single variable. Familiar first-order statistics include the sample mean and the sample variance. A well-studied example of first-order differential analysis is the study of differential gene expression in microarrays (see Cui and Churchill (2003) for a canonical example, or Soneson and Delorenzi (2013) and the references therein for an overview of several methods). Other applications of first-order differential analysis include text analysis for authorship identification (Stamatatos, 2009), studies of brain functionality based on regional activation (Phan et al., 2002), and investigation of cultural bias in standardized testing (Wainer and Braun, 2013).

The use of first-order statistics allows for analysis of only a single variable at a time. To study relationships between pairs of variables, one requires a measure of association such as correlation. Kriegel et al. (2009) provides a survey of clustering methods for high-dimensional data based on correlation distance. Datta and Datta (2002) and Jiang et al. (2004) and the references therein give an overview of methods developed specifically for clustering of gene expression. In general, typical clustering or community detection methods must be adapted for application to correlation distances to correct for bias (see e.g. MacMahon and Garlaschelli (2015) for an illustrative example). In applications of non-differential correlation mining, variable groups may represent, e.g., social groups communication networks (Lewis et al., 2008), genes in common protein pathways (Jiang et al., 2004), or functionally similar brain regions (Greicius et al., 2002).

While there is a large literature on clustering and community detection via correlation, there is relatively little work comparing association across two sample conditions. The many insights obtained from ordinary correlation studies lead us to believe that a second-order differential approach, or *differential association mining*, will be of scientific interest. As in all association mining, features of interest derive from pairwise behavior of variables; however, in the differential setting,

one studies two different sets of variable relationships. In some cases, simply taking the difference of dissimilarity matrices and applying ordinary clustering methods would suffice. However, most second-order statistics - including the linear correlation coefficient - require a more careful treatment. For instance, two sample correlation matrices will exhibit vastly different random behavior based on the sample sizes of the corresponding datasets, and will have a complex dependency structure when the corresponding population correlation matrices are not the identity.

Chapter 3 introduces **Differential Correlation Mining (DCM)**, a new method of second order comparative analysis that identifies sets of variables such that the average pairwise correlation between variables in the set is higher in one sample condition than in another. The method does not make use of auxiliary information, apart from the separation of samples into pre-determined groups (e.g. treatment vs control). Differential Correlation Mining is theoretically applicable to both low and high dimensional settings and is computationally feasible for high dimentional data (10^5 variables).

1.3.1 Example: TCGA

The following real-world example provides a brief illustration of and motivation for Differential Correlation Mining . Figure 1.1 shows a differentially correlated variable set identified by the DCM procedure in real data from The Cancer Genome Atlas (TCGA) Research Network (`http://cancergenome.nih.gov/`). The two sample conditions under consideration are Her-2 type breast cancer tumors and Luminal B type tumors, as classified by (Perou et al., 2000). Further results for the TCGA dataset are provided in Section 3.7.

Figure 1.1 shows the sample correlation matrices within each tumor type, restricted to a set of 202 variables consisting of a set of size 102 selected by DCM (A), and 100 randomly chosen variables (B). The variables B are included for contrast, and to show that the differential correlation observed in A is not present in the entire dataset. The figure illustrates the second-order behavior and the differential nature of the variable set A. The block pattern in the upper left corner of the Her-2 matrix shows that every entry in the correlation matrix of A is large, suggesting that all the variables of A are strongly pairwise correlated. The Luminal B sample correlation shows a similar pattern, but it is much less pronounced. No such pattern is seen among the variables in B.

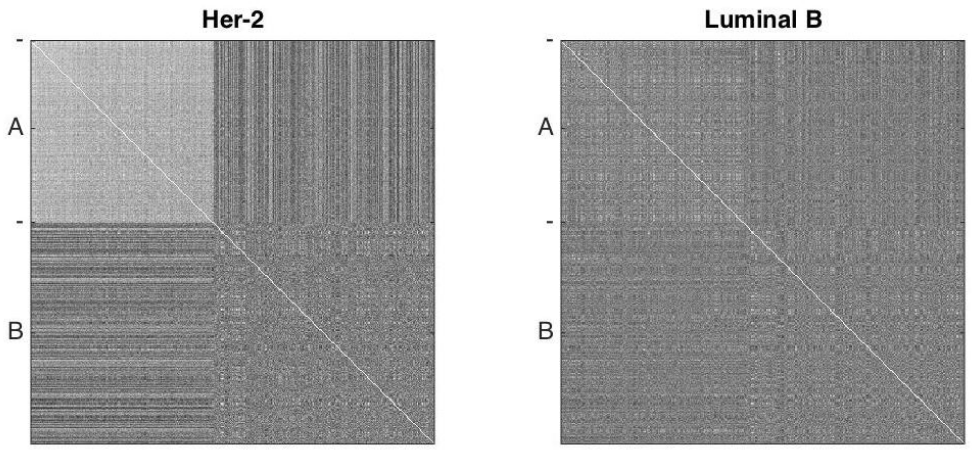


Figure 1.1: Sample correlation matrices for each of two breast cancer tumor subtypes, showing observed DC clique (A) and random genes (B) .

In general, the results of Differential Correlation Mining are distinct from those found by first-order analysis (e.g. differential expression). For example, Figure 1.2 shows the relative differential expression, overall expression level, and differential variation for the above estimated DC clique A. In this plot, all genes in the study (p =15,785) are ranked by (a) t statistic of differential mean expression between Her-2 and Luminal B samples, (b) overall expression in Her-2 samples, and (c) ratio of sample variations (F statistic) for Her-2 versus Luminal B samples. The histograms in Figure 1.2 show the ranking of the genes in A. The overall uniformity of the histograms indicates that the variables in the observed DC clique A do *not* exhibit standard first-order differential behavior.

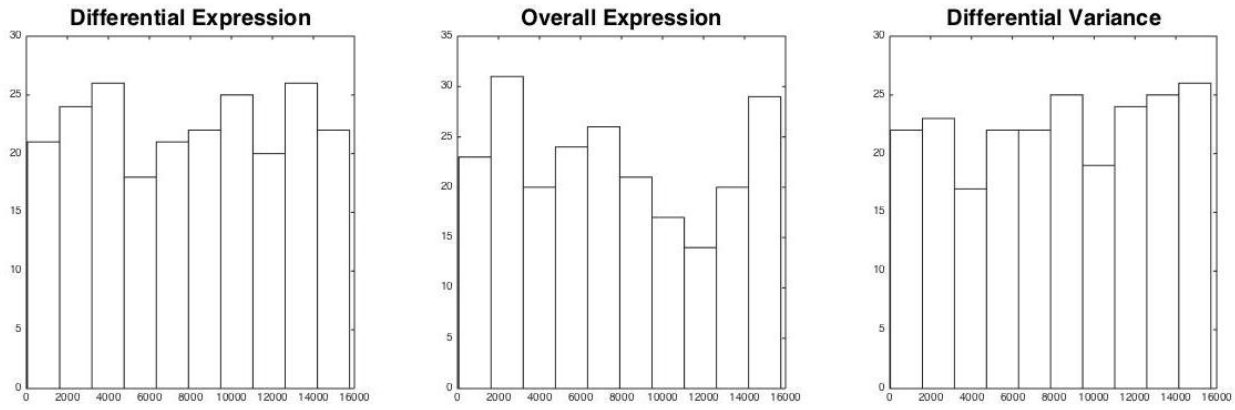


Figure 1.2: Ranks of genes in observed DC clique (A) out of 15,785 total genes.
(Ranked by: Differential expression, as measured by p-values of 2-sample t-tests; mean overall expression among Her-2 samples; and ratio of sample variances between Her-2 and Luminal B.)

By targeting differentially correlated variable sets, the Differential Correlation Mining method identifies variables whose *joint* behavior is different across sample conditions. The results are readily interpretable as sets of variables that interact strongly under one sample condition but only weakly (or not at all) under another.

1.3.2 Related work

Much existing work is either directly related to differential association or may be reasonably adapted to such a paradigm. In what follows, let $\mathbf{R}_1, \mathbf{R}_2$ denote the population correlation matrices of two data distributions, and let $\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_2$ denote the corresponding sample correlation matrices.

1. Mining from single correlation matrices.

   Non-differential correlation mining has been well-studied, typically in the context of clustering. These methods may be applied in the differential case by separately clustering the correlation matrices $\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_2$ and comparing results.

2. Detection of isolated changes in correlation structure.

   Existing approaches to differential correlation mining are largely based on examining individual variables for changes in second-order structure across two sample conditions. For example, one may treat $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ as the adjacency matrices of two fully connected, weighted networks, and then look for variables whose connectivity pattern is very different across the two networks (Xia et al., 2014; Gill et al., 2010). Most methods approach differential correlation mining by developing a statistic to measure the change in pairwise correlations of an individual variable: Hu et al. (2010) uses the covariance distance (total difference of covariances), Choi and Kendziorski (2009) uses a direct difference of sample correlations, Fukushima (2013) uses the difference of Fisher transformed sample correlations, and Liu et al. (2010) use a filtration (or thresholding) step before summing square correlation differences. These methods then permute samples across the two classes to measure the significance of the original differential correlation. Significant variables may then be selected by an appropriate multiple testing procedure.

3. Estimation and hypothesis testing.

There has been a great deal of theoretical work devoted to testing equality of high-dimensional covariance and correlation matrices. When the sample size n is substantially larger than the dimension p, classical results are applicable, e.g., likelihood ratio tests as discussed in Anderson (1958) and Muirhead (1982), or results like those of Steiger (1980) for testing individual sample correlation. In the high-dimensional ($p > n$) setting, Cai et al. (2010), Cai and Jiang (2011), and Cai et al. (2014) have developed minimax rate optimal tests for the equality of covariance matrices under sparsity assumptions. Results for correlation (rather than covariance) are less prevalent; recent work includes tests for sets of sample correlation coefficients (Donner and Zou, 2014), tests for rank-based correlation matrices (Zhou et al., 2015), and tests for detecting overall dependence (Bassi and Hero, 2012).

In some cases, optimal testing procedures can inform methods for estimation of high-dimensional covariance and correlation matrices. Particularly relevant is the work of Cai and Zhang (2014), which yields an estimator for the difference matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. This estimator is implemented and discussed further in Section 3.6. Other approaches to high-dimensional estimation include: Bickel and Levina (2008), who discuss a thresholding estimator for covariance matrices; Peng et al. (2008), who estimate partial correlations in sparse regression models; and Rajaratnam et al. (2008), who make use of graphical model techniques for covariance matrix estimation.

4. Direct mining of differential correlation

Finally, the work of Sheng et al. (2016) proposes an approach to correlation mining by testing subsections of the difference of correlation matrices $\mathbf{R}_1 - \mathbf{R}_2$. Like Differential Correlation Mining, the proposed method seeks to identify groups of differentially correlated variables by appealing to classical asymptotic results. However, the method relies on a sequential testing and screening procedure that is infeasible for high dimensional settings ($\sim 10^2$ or more). As such, despite the close relationship between this method and Differential Correlation Mining, we were not able to include it in the simulation study in Chapter 3.

1.4 Association Mining in Binary Data

The majority of well-known association mining methods are implicitly designed for continuous data. However, in some common settings, data may take the form of binary (0/1) observations. For example, purchasing information - known as *market basket data* - often consists of observations about d items available for purchase by n buyers. The resulting data matrix $\mathbb{X} \in \{0,1\}^{n \times d}$, where X_{ij} indicated whether buyer i bought item j, therefore may be interpreted as n samples of a d-dimensional binary random variable. It may be of interest to identify association structure in these d variables from the n samples.

In its basic form, this problem is not distinct from ordinary clustering and community detection methods. Algorithms like hierarchical clustering may be applied to any dissimilarity matrix, so as long as an appropriate measure of association is chosen, these methods still apply. However, binary data presents a unique challenge when it comes to standardization. Consider standardizing a vector in $\{0,1\}^n$ such that the sample mean is 0 and the sample variance is 1. The values $\{0,1\}$ are then each transformed to a different pair of values. No real transformation has been applied to the data; it is still dichotomous. Measurements such as product-moment correlation, which rely on a standardization step, are therefore not as appropriate as metrics for associations studies.

A further challenge arises when samples are not treated identically. Measures of association that involve an unweighted average over sample quantities, such as L1 and L2 distances, are unequipped to account for different behavior between buyers. In continuous data, differences between samples are often swept under the rug via pre-processing of data, usually by sample-standardizing before variable-standardizing. This option is less appealing in the binary case.

The method introduced in Chapter 4, Coherent Set Mining (CSM), was developed to be flexible to sample heterogeneity without disregarding the inherent dichotomous nature of binary observations. The following simple application motivates the need for such an approach, and give an overview of existing work in association mining that is specific to binary data.

1.4.1 Example: Grocery Store Data

The package `arules` (Hahsler et al., 2012) in **R** supplies software for several common frequent itemset mining and association mining methods. Also included in this package is a dataset from

grocery store transactions, `Groceries`, intended as ideal data for exploring and testing association mining methods. This dataset consists of observed 9835 transactions for 169 items. Tables 1.1 and 1.2 show the results of applying the well-known `eclat` algorithm and our new method, Coherent Set Mining, to the grocery store data. Since `eclat` screens for itemsets with support above a certain threshold, we applied the method with many thresholds. Table 1.1 shows the results for a threshold that yielded a moderate number of reasonably-sized itemsets. The Coherent Set Mining method, by contrast, is fully automatic and so the contents Table 1.2 are simply the direct output of the method.

The results in Table 1.1 lack an obvious interpertation. All three frequent sets contain whole milk, the most common item in the `Groceries` dataset. Intuitively, this makes sense, because the `eclat` algorithm seeks itemsets that appear in a large percentage of transactions; thus, items which are puchased more often overall are more likely to appear in frequent sets. The itemsets in Table 1.2, on the other hand, are readily interpretable in terms of real world grocery needs. For instance, Set 1 in Table 1.2 is easily recognizeable as a ham and cheese sandwich, Set 5 contains drinks one might buy for a party, and Set 7 evidently corresponds to baking staples.

Table 1.1: Results from **eclat** with support threshold = 0.05

1. `whole milk, other vegetables`
2. `whole milk, rolls/buns`
3. `whole milk, yogurt`

This example is provided to briefly justify the need for a new approach to association mining in binary data. Chapter 4 offers an in-depth discussion of settings where existing methods are susceptible may be measuring association that is not of scientific interest. The Coherent Set Mining approach is designed with such settings in mind, to work around challenges like the overall frequency of whole milk and produce more meaningful results such as those in Table 1.2.

1.4.2 Related Work

- Clustering with binary association measures.

Table 1.2: Results from CSM

1. white bread, processed cheese, ham
2. canned beer, soda, shopping bags
3. pip fruit, tropical fruit
4. root vegetables, herbs, beef, other vegetables, pork, chicken
5. soda, bottled water, bottled beer, red/blush wine, canned beer
6. berries, whipped/sour cream
7. sugar, flour, baking powder
8. Instant food products, hamburger meat
9. waffles, chocolate, long life bakery product, specialty bar, candy, specialty chocolate, salty snack, chocolate marshmallow

In principle, existing methods for clustering or community detection can easily be applied to binary data; one need only specifiy a measure of dissimilarity. However, there are many options for how best to infer relationships between variables from binary observations. (Choi et al., 2010) provide an overview of 76 different suggested dissimilarity measures (some of which are mathematically equivalent). Notable among these are the Phi Coefficient, which is equal to product-moment correlation, and the Jaccard distance, which considers a ratio of co-occurance to individual occurance. Some prior work also addresses the case of binary observations directly. Li and Li (2005) provide a general framework and methodology for clustering binary data, and Neuhaus et al. (1991) summarizes classic methods for analyzing correlated binary data.

- Frequent itemset mining and association rules.

Association mining in binary data is sometimes known as *itemset mining*, due to the prevalance of *market basket data*, in which variables take the form of items available for purchase, and observations (or transactions) represent an individual buyer's choice to purchase or not purchase each item. Associated itemsets, then, are items which tend to be bought together, which can be valuable information to researchers for purposes of advertising, inven-

tory control, and so forth. Methods for mining in market basket data fall under the heading of *frequent itemset mining* or *association rules*. In general, approaches to frequent itemset mining are non-stochastic; instead of modeling the data, they proceed by screening datasets for sets of items whose *support* - or percentage of buyers who purchased the entire itemset - is above a certain threshold. For example, a frequent itemset discovered from grocery store purchases might take the form `{milk, eggs, bread}`.

The study of frequent itemsets and association rules arguably began with the work of Agrawal et al. (1996), which introduced the `apriori` algorithm. This method is built on the *apriori principle*: that for an itemset to be frequent, all of its subsets must also be frequent. The apriori approach vastly reduces the number of itemsets that must be screened to search a dataset exhaustively. Subsequent methods improved on **apriori** by both algorithmic solutions and computational improvements. Some notable examples include `eclat` (Zaki et al., 1997a), MAFIA (Burdick et al., 2001), COBBLER (Pan et al., 2004), `fp-close` Grahne and Zhu (2003), and CHARM (Zaki and Hsiao, 2002). Zaki et al. (1997b), Prabha et al. (2013), (Zaki et al., 1999) and the references therein provide an excellent summary of early and recent work in frequent itemset mining. There are also some exceptions to the non-stochastic nature of itemset mining. Zhang et al. (2008) estimates the probability of itemsets exceeding a specified frequency, rather than simply screening for itemsets exceeding a threshold. Aggarwal et al. (2009) and Tong et al. (2012) take more complex model-based approaches to data uncertainty. Instead of screening for high support, they screen for high *expected* support under a probability model.

In general, frequent itemset mining methods are built to handle data that has a potentially very large number of samples (transactions). However, the number of items is taken to be moderate (commonly on the order 10^2 or less), since algorithms typically rely on screening *all* possible item subsets of many sizes. More recent work in itemset mining addresses the challenge of high dimensional data, in which the number of items studied may be very large (usually 10^4 or more). Such methods are known as *colossal itemset mining*. (Here "colossal" refers to the number of total items being searched for frequent sets, rather than the size of the discovered itemsets or the number of samples.) As with itemset mining in small data, existing

methods are primarily non-stochastic, and the research focus is algorithmic and computational alternatives to an exhaustive search over all possible itemsets. For important examples, see Liu et al. (2006), Sohrabi and Barforoush (2012), and Zhu et al. (2007). Unfortunately, public software is not readily available for large data, and foundational small-data methods like `apriori` and `eclat` are still the norm in analyses of market basket data.

CHAPTER 2

Variable-to-Set Affinity Testing

2.1 Introduction

In general terms, the goal of the association mining algorithms in this dissertation is to identify subsets of variables that are more associated internally than externally. Classical approaches to problems of this type typically rely on an optimization algorithm. That is, every variable set or every partition is given a score intended to measure the strength of in-group versus out-group association. For example, in k-means clustering (MacQueen, 1967), a particular partition is scored by the within-cluster sum of squared distances to means. These methods then apply a maximization (or minimization) procedure with the goal of identifying clusters or communities with high (or low) scores.

There three main challenges inherent to such approaches. First, if one has a moderate to large number of variables, it is in most cases computationally intractable to find global optima for a score. Instead, methods most commonly apply algorithms guaranteed to reach local optima, then run these localized procedures many times and select the "best" output. For instance, the k-means algorithm consists of a greedy iterative procedure to refine k cluster centers until the within sum-of-squares distance to center is locally minimized. Since the locally minimal choice of cluster centers may be different depending on the starting point, it is common to apply k-means multiple times to a particular dataset and report only the most optimal partition.

Secondly, even if the global optimum for an association score is accessible, most methods do not come equipped with any kind of assessment for statistical significance of the results. *Any* dataset that is fed to a score maximization type clustering method will yield a result, even if no true association structure exists in the data. There have been some attempts to quantify the statistical significance by assuming an underlying generative model; for example, Liu et al. (2008) assigns significance to clusters under the assumption of an underlying multivariate Gaussian distribution,

and Lancichinetti et al. (2011) imposes a model on networks and then measures significance of communities based on asymptotic extreme value results. It is also sometimes possible to derive measure of significance via a bootstrapping or permutation-based approach, see e.g. Jakobsson and Rosenberg (2007). However, these solutions all represent *ex post facto* assessments of clusters or communities; statistical principles are not embedded in the search algorithm itself.

Finally, existing methods of clustering and community detection commonly require crucial user input. In k-means and many other clustering methods, one must pre-select k, the number of clusters. In hierarchical clustering, the final selected partition depends on the choice of where to cut the dendrogram. For community extraction methods, it is usually necessary to specify a score threshold above which a community is considered "interesting". Reliance on user-specified information weakens the conclusions of association mining as compared to a fully data-driven method.

As a response to some of the limitations of existing association mining techniques, the methods in this dissertation make use of the **Variable-to-Set Testing (VSAT)** algorithmic framework first introduced by Wilson et al. (2014). VSAT is a general approach to statistical association mining based in hypothesis testing principles. Methods built from the VSAT algorithm enjoy many advantages over classical clustering and community detection, such as flexibility to unusual data types and/or particular measures of association. Associated variable sets selected by VSAT algorithms have natural statistical interpretations and error control guarantees. Because VSAT type methods choose variable sets adaptively from significance testing results, one need not pre-specify a number or size of clusters or a score cutoff. Finally, implementations of VSAT algorithms tend in general to be computationally efficient. Importantly, the VSAT approach is not an optimization procedure. No score is involved; rather, sets are chosen organically via testing-based iterative update.

At present, two VSAT type methods are available for association mining in networks: The ESSC method of Wilson et al. (2014), for community extraction on unweighted random networks, and the CCME method of Palowitch et al. (2016), which generalizes ESSC to weighted networks. Chapters 3 and 4 detail two more VSAT type methods for association mining. This chapter generalizes the VSAT framework for non-network settings, particularly in the context of the methods in this thesis, and provides a simple theoretical guarantee.

2.2 The variable-to-set testing algorithm

The VSAT algorithm relies on the determination of a population quantity of interest, a test statistic, and a null model. The choice of measure of association and assumptions about data for a particular analysis dictate the appropriate test statistic and null. This chapter provides a discussion of the VSAT framework in terms of a general choice of measure of association.

Formally, define $\zeta(j, A)$ to be a measure of *affinity* between variable j and variable set $A \subset [d]$. In general, $\zeta(j, A)$ is a function of the set of pairwise associations $a(j, k)$ between j and $\{k : k \in A\}$ for some choice of association measure $a(\cdot, \cdot)$. Most VSAT methods will define $\zeta(j, A)$ to be a simple average over $k \in A$, that is, $\zeta(j, A) := |A|^{-1} \sum_{k \in A} a(j, k)$. For example, to mine for highly correlated variable sets, one might set $a(j, k)$ to be the population correlation ρ_{jk} between variables j and k, then let $\zeta(j, A)$ be the average of these correlations. In general, however, $\zeta(j, A)$ may be chosen to reflect the association structure of interest in a particular research problem. Given a choice of $\zeta(j, A)$, the VSAT algorithm is designed to use statistical principles to search for **ζ-connected sets**, defined as follows.

Definition 1. (ζ-connected set) *An index set $A \subset [d]$ with at least two elements is* ζ-connected *with regard to an affinity measure $\zeta(\cdot, \cdot)$ if*

(i) for all $j \in A$, $\zeta(j, A) > 0$, and

(ii) for all $j \notin A$, $\zeta(j, A) \leq 0$.

A ζ-connected set may be thought of as "closed", in the sense that only elements in the set have positive affinity (as measured by ζ) with the rest of the set. The VSAT search procedure for ζ-connected sets is summarized as follows.

1. Initialization: Set $A_0 \subset [d]$.

2. Testing: Given A_t, simultaneously test hypotheses for the affinity of $j \in [d]$ to A_t,

$$H_0(j) : \zeta(j, A_t) = 0 \quad \text{vs} \quad H_1(j) : \zeta(j, A_t) > 0 \tag{2.1}$$

by an appropriate multiple testing procedure.

3. UPDATE: Set $A_{t+1} = \{ j : H_0(j) \text{ was rejected } \}$.

4. ITERATION: Repeat steps 3 and 4 until $A_t = A_{t'} := A^*$.

5. OUTPUT: If A^* is not empty, select it as an estimated ζ-connected variable set.

6. REPETITION: Repeat steps 2-5 as many times as desired, or until no further sets are found.

Steps 2-5 may be considered a refinement process, during which a proposed ζ-connected set A_t is updated in accordance with the results of simultaneous hypothesis tests. Regardless of the size of initial set A_0, the size of output sets A^* is chosen adaptively by the application of multiple testing. Furthermore, because updates require statistical significance, not every initial set A_0 is guaranteed to produce convergence to a non-empty set A^*.

If $t' = t - 1$, the sets A^* are *fixed points* of convergence of the VSAT algorithm in that further updates will not change the elements of A^*. When $t' \neq t - 1$, the algorithm has reached a cycle of three or more sets $A_t, \ldots, A_{t'}$ that will continue ad infimum as the algorithm continues. Although these sets are not as ideal as fixed points, which are discussed below, they are often highly overlapping and may be of interest. Particular implementations of VSAT methods take different approaches to cycles. For the remainder of this chapter, we restrict our discussion only to fixed points, or *stable sets*, which have several desirable properties.

Definition 2. (Stable Set) *Let $U_\alpha(A, \mathbb{X})$ denote the index set of the rejected hypothesis tests for $\zeta(j, A) = 0$ from observed data $\mathbb{X}$. An index set $A^* \subset [d]$ is a* stable set *in $\mathbb{X}$ if $U_\alpha(A^*, \mathbb{X}) = A^*$.*

Note, trivially, that $A^* = \emptyset$ is always stable set, albeit not one of scientific interest. There is a close relationship between nonempty stable sets and the ζ-connected variable sets they approximate. A stable set A^* has the property that for hypothesis tests $H_0(j) : \zeta(j, A) = 0$ performed on a particular observed dataset,

(i) for all $j \in A$, $H_0(j)$ was rejected, and

(*ii*) for all $j \notin A$, $H_0(j)$ was accepted.

It is clear that A^* exhibits the properties of Definition 1 up to a level of statistical significance. As such, the VSAT algorithm is a natural approach to estimating ζ-connected sets from data $\mathbb{X}$.

2.3 Deriving hypothesis tests

The crucial element of the VSAT algorithmic framework is the ability to test the hypotheses in (2.1) for a desired affinity measure ζ. In order to develop a VSAT type method for a particular association mining setting, one requires

1. A random vector or matrix $\mathbf{X}$, containing information about variables $1, \ldots, d$
2. A test statistic $S(j, A|\mathbf{X})$ for $\zeta(j, A)$; and
3. A null model $\mathbb{P}_0$ specifying the distribution of $S(j, A|\mathbf{X})$ when $\zeta(j, A) = 0$.

The test statistic $S(j, A \,|\, \mathbf{X})$ will in most cases be a direct estimator for $\zeta(j, A)$ from $\mathbf{X}$, such that large positive values provide evidence for $\zeta(j, A) > 0$. Then, p-values pertaining to the hypotheses in (2.1) can be computed from an observed dataset $\mathbb{X}$ by

$$\mathrm{p}(j, A \,|\, \mathbb{X}) := \mathbb{P}_0 \left(S(j, A \,|\, \mathbf{X}) > S(j, A \,|\, \mathbb{X}) \right) . \tag{2.2}$$

In other words, the p-values measure the extremity of the observed data $\mathbb{X}$ under the a null distribution on $\mathbf{X}$.

The determination of the null distribution $\mathbb{P}_0$ is an important aspect of developing a VSAT type method. The null should reflect, in some sense, an association structure that is *not* of scientific interest. In some settings, datasets $\mathbb{X}$ will consist of n i.i.d observations of a d-dimensional random vector $\mathbf{X}$. Then, it is often possible to derive an asymptotic approximation for $\mathbb{P}_0$ via a central limit theorem, under the null assumption $\zeta(j, A) = 0$ and mild regularity conditions on $\mathbb{P}_0$. For example, a simple test statistic for the average correlation between j and A, is the average of *sample* correlations, computed in the usual way from samples of $\mathbf{X}$. Steiger and Hakstian (1982) provides a central limit theorem for a vector of sample correlations under mild conditions. Chapter 3 provides a more extensive derivation of this type of approximation.

In other settings, $\mathbb{X}$ may simply represent a single instance of a dissimilarity matrix. Such is the case in the existing VSAT methods for networks (Wilson et al. (2014), Palowitch et al. (2016)). Observed data in these cases takes the form of a network or a series of networks, represented by a node set $[d]$ and a (possibly weighted) edge set E capturing relationships between nodes. To perform inference on these artifacts, a specific null generative model is assumed appropriate to the data context. Then, the distribution $\mathbb{P}_0$ on $S(j, A \,|\, \mathbf{X})$ can be derived from this null model.

Remark: Ancillary Statistics. Commonly, the null measure $\mathbb{P}_0$, or asymptotic approximation thereof, will depend on an unknown parameter η that has no bearing on the measure ζ of interest, but that nonetheless must be estimated from data. For instance, in the example of basic correlation mining, the asymptotic distribution supplied by Steiger and Hakstian (1982) depends in part on the covariance between sample correlations, for which an explicit form is derived in terms of the population moments. In practice, one must estimate the covariances from sample estimates of moments in order to compute p-values. Ideally asymptotic results regarding $\mathbb{P}_0$ continue to hold when η is replaced by a data-driven estimate $\hat{\eta}$.

2.4 Flexibility in objective

An association mining method by the VSAT approach is fully specified by a choice of $S(j, A \,|\, \mathbf{X}$ estimating $\zeta(j, A)$ and a null model $\mathbb{P}_0$. Therefore, in principle, one can use this framework to perform association mining for *any* data feature of scientific interest that can be reasonably represented as a ζ-connected set for some ζ. Notions of pairwise variable relationships that do not lend themselves well to the creation of a single summarizing dissimilarity matrix, necessary for most association mining methods, are still possible to study under via VSAT algorithm. The two methods derived in this thesis, Differential Correlation Mining and Coherent Set Mining, take full advantage of the flexibility inherent to the VSAT algorithm.

2.4.1 VSAT and Differential Correlation Mining

As introduced in 1.3.2, the Differential Correlation Mining (DCM) method was created to identify variable sets that are more highly correlated in one sample condition than in another. In

terms of the VSAT framework,

$$\zeta(j, A) = \frac{1}{|A|} \sum_{k \in A} (\mathbf{R}_1 - \mathbf{R}_2)_{jk} \quad \text{and} \quad S(j, A \,|\, \mathbf{X}) = \frac{1}{|A|} \sum_{k \in A} \left(\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2\right)_{jk}, \tag{2.3}$$

where $\mathbf{R}_1, \mathbf{R}_2$ are population correlation matrices under two sample conditions and $\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_2$ are the usual sample correlation matrices computed from observed datasets of samples from the two conditions, $\mathbb{X}_1$ and $\mathbb{X}_2$. In the Differential Correlation Mining method, $\mathbb{P}_0$ is approximated by a Gaussian measure based on a central limit theorem for S.

The VSAT flexibility comes into play with regard to the covariance matrix of the test statistic, $\{\text{cov}\,(S(j, A), S(k, A))\}_{j,k \in A}$, which is needed to approximate $\mathbb{P}_0$. If one were to apply ordinary clustering or community detection to the adjacency matrix $\left(\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2\right)$, one would not be accounting for variability in average correlations of a set A. The Differential Correlation Mining method allows us to estimate the necessary covariance from data, and therefore to test $\zeta(j, A)$ directly. Estimated ζ-connected sets are then interpretable as variable sets that have higher average pairwise correlation in the first sample condition, up to a level of statistical significance.

2.4.2 VSAT and Coherent Set Mining

Section 1.4.2 introduced the challenges of association mining from binary observations, and gave an example of data for which existing methods are not appropriate. Our approach to this problem, described in detail in Chapter 4, is to model binary data as a thresholded version of unobserved latent data. The measure of association of interest is the correlation in the *latent* data, i.e., $a(j, k) = \text{cor}(Z_j, Z_k)$ for some latent variable $\mathbf{Z} \in \mathbb{R}^d$ and $\zeta(j, A)$ is the average of correlations between Z_j and $\{Z_k\}_{k \in A}$. However, only a thresholded version of $\mathbf{Z}$ given by $\mathbf{X} \in \{0, 1\}^d$ is observed. In this setting, one does not have enough information to directly estimate the correlation structure of $\mathbf{Z}$. That is, it is not possible to craft a test statistic that is a reasonable estimator for $\zeta(j, A)$.

Fortunately, the VSAT approach does not require an estimator for $\zeta(j, A)$, only a procedure for testing departures from $\zeta(j, A) = 0$. The Coherent Set Mining method relies on a carefully define a statistic $S(j, A \,|\, \mathbf{X})$, referred to as "coherence", such that large values of $S(j, A \,|\, \mathbf{X})$ are evidence

for large values of $\zeta(j, A)$. Our null model $\mathbb{P}_0$ is then derived in part from a central limit theorem, and in part from imposed null assumptions about the thresholding of $\mathbf{Z}$ to $\mathbf{X}$.

Coherent Set Mining, in other words, is an association mining method capable of estimating ζ-connected sets, *even though ζ itself cannot be estimated.* The power of the VSAT framework lies in its flexibility to uncommon choices of ζ driven by unique data, and to choices of $\mathbb{P}_0$ driven by specific research questions

2.5 Control of global familywise error under the null

Since VSAT algorithms incorporate a multiple testing step at each iteration of the set update process, it is reasonable to expect that error control properties hold for the entire procedure. Indeed, it can be shown that in an idealized setting for small α, the probability of false identification is controlled.

This result, while simple, is important: it guarantees that in data where no signal is present (as defined by $\mathbb{P}_0$), the probability of *any* stable set being present is controlled at level α. Interestingly, (2.4) is a familywise error control property, even though the multiple testing procedure of (Benjamini and Hochberg, 1995) only controls False Discovery Rate. Theorem 1 allows us to have confidence that stable sets discovered by a VSAT type algorithm are likely to reflect true population structure.

Theorem 1. (VSAT global error control)

Fix $\alpha \in (0, 0.15]$. Let $\mathcal{A}(\mathbf{X}, \alpha)$ be the class of all stable sets of a VSAT algorithm using the Benjamini-Hochberg multiple testing procedure (Benjamini and Hochberg, 1995) at level α. Assume that for any $A \subseteq [d]$, the p-values $\{\mathrm{p}(j, A \,|\, \mathbf{X}) : j \in [d]\}$ are independent and uniformly distributed. Then,

$$\mathbb{P}_0\big(\, |\mathcal{A}(\mathbf{X}, \alpha)| > 0 \,\big) \;<\; \alpha\,. \tag{2.4}$$

where $\mathbb{P}_0$ denotes the probability under the null model for $\mathbf{X}$.

2.5.1 Example

The result of Theorem 1 is powerful, but it relies on strong assumptions about the p-values in a VSAT update step; namely, that they are uniform and independent. Typically, in VSAT methods

asymptotic uniformity of p-values can be guaranteed by deriving a limiting distribution on the test statistic. Independence, however, is not always a reasonable assumption. The following example provides a setting where both conditions of Theorem 1 are met.

Example 2.1. Let $P = \{p_{jk}\} \in [0,1]^{(d\times d)}$ be a matrix of fixed probabilities, with p_{jk} not necessarily equal to p_{kj}. Define the measure of affinity for an index j and a set $A \subset [d]$ to be

$$\zeta(j, A) \;=\; \frac{1}{|A|}\sum_{k\in A} p_{jk} - \frac{1}{d-|A|}\sum_{k\in A^C} p_{jk}\,. \tag{2.5}$$

Let $\mathbf{X} \in \{0,1\}^{(d\times d)}$ be a random matrix with $X_{jk} \sim \text{Bernoulli}(p_{jk})$, all independently. Suppose i.i.d. copies $\mathbf{X}(1), \ldots, \mathbf{X}(n)$ are observed. Define the test statistic for an index j and a set $A \subset [d]$ to be

$$S(j, A \,|\, \mathbf{X}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{|A|}\sum_{k\in A} X_{jk}(i) - \frac{1}{d-|A|}\sum_{k\in A^C} X_{jk}(i)\right]. \tag{2.6}$$

Finally, let the null model $\mathbb{P}_0$ be that $p_{j1} = \ldots = p_{jd} = p_j$ for every j. Then, $\zeta(j, A) = 0$ for all j and for any $A \subset [d]$. $\Diamond$

The data in Example 2.1 can be interpreted as a set of i.i.d. samples of a directed unweighted network. For example, the data might represent observations of behavior over time for a group of individuals, with $X_{jk}(i)$ representing whether individual j visited the Facebook page of individual k on day i. Then, a ζ-connected set under the definition of $\zeta(j, A)$ would be a group of individuals who visit each others pages on average more than they visit other people's. The null model may be interpreted as an assumption that each individual visits all her friend's pages equally often.

Since observations $X_{jk}(i)$ are binary, $S(j, A \,|\, \mathbf{X})$ is bounded in $[1, -1]$. Therefore, since $S(j, A \,|\, \mathbf{X})$ is a sum of bounded i.i.d. variables, an ordinary central limit theorem applies. Note that the mean of $S(j, A \,|\, \mathbf{X})$ is 0 under the null model, and its variance is given by

$$\begin{aligned}\text{var}(S(j, A \,|\, \mathbf{X})) \;&=\; \frac{1}{n}\left(\frac{1}{|A|^2}\sum_{k\in A}\text{var}(X_{jk}) + \frac{1}{(d-|A|)^2}\sum_{k\in A^C}\text{var}(X_{jk})\right)\\ &=\; \frac{1}{n}\left(\frac{1}{|A|} + \frac{1}{(d-|A|)}\right) p_j(1-p_j)\,.\end{aligned} \tag{2.7}$$

It is straightforward to show that

$$\hat{p}_j \;:=\; \frac{1}{nd}\sum_{i=1}^{n}\sum_{k\in[d]} X_{jk} \tag{2.8}$$

is a consistent estimator for p_j under the null. Then, $\hat{\sigma}(j,A) := d\,n^{-1}(d-|A|)^{-1}(\hat{p}_j(1-\hat{p}_j)$ is consistent for the variance of $S(j,A\,|\,\mathbf{X})$. Therefore, under $\mathbb{P}_0$, for fixed A, p-values given by

$$\mathrm{p}(j,A\,|\,\mathbf{X}) \;=\; 1-\Phi\left(\frac{S(j,A\,|\,\mathbf{X})}{\hat{\sigma}(j,A)}\right), \tag{2.9}$$

where $\Phi(\cdot)$ is the standard normal cdf, are asymptotically uniformly distributed. Finally, due to the fact that $p_{jk}\neq p_{kj}$ and that the variables X_{jk} are independent, it follows that $S(j,A\,|\,\mathbf{X})$ and $\hat{p}_j$ are independent of $S(k,A\,|\,\mathbf{X})$ and $\hat{p_k}$ for any $j\neq k$. Then, $\mathrm{p}(j,A\,|\,\mathbf{X})$ and $\mathrm{p}(k,A\,|\,\mathbf{X})$ are independent for $j\neq k$. We conclude that the setting in Example 2.1 asymptotically satisfies the conditions of Theorem 1.

2.5.2 Proof

Define $\mathcal{C}_m := \{A : |A| = m\}$. By construction of the Benjamini-Hochberg procedure, the event that $A\in\mathcal{C}_m$ is a fixed point only if

$$\bigcap_{j\in A}\left\{p(j,A;\mathbf{X})\leq\frac{m\alpha}{d}\right\}\bigcap\bigcap_{j\in[d]\setminus A}\left\{\mathrm{p}(j,A\,|\,\mathbf{X})>\frac{m\alpha}{d}\right\} \tag{2.10}$$

Since the p-values are independent and uniformly distributed, this implies that for any $A\in\mathcal{C}_k$,

$$\mathbb{P}_0\big(u_\alpha(A)=A\big)=\Big(\frac{m\alpha}{d}\Big)^m\Big(1-\frac{m\alpha}{d}\Big)^{d-m} \tag{2.11}$$

Define $\mathcal{A}_m(\mathbf{X},\alpha)$ to be the class of all stable sets of size m. Then, using equation 2.11 and a union bound,

$$\mathbb{P}_0\big(\,|\mathcal{A}_m(\mathbf{X},\alpha)|>0\,\big)\;\leq\;\binom{d}{m}\Big(\frac{m\alpha}{d}\Big)^m\Big(1-\frac{m\alpha}{d}\Big)^{d-m} \tag{2.12}$$

Applying the inequality $\binom{d}{m}\leq\frac{1}{\sqrt{2\pi}}(\frac{ed}{m})^m$ gives

$$\sqrt{2\pi}\;\mathbb{P}_0\big(\,|\mathcal{A}_m(\mathbf{X},\alpha)|>0\,\big)\leq(e\alpha)^m\Big(1-\frac{m\alpha}{d}\Big)^{d-m}\leq(e\alpha)^m$$

Since $\mathcal{A}(\mathbf{X}, \alpha) = \cup \mathcal{A}_m$, a union bound gives

$$\sqrt{2\pi}\, \mathbb{P}_0\big(\, |\mathcal{A}(\mathbf{X}, \alpha)| > 0 \,\big) \;\leq\; \sum_{m=2}^{d} (e\alpha)^m = \sum_{m=1}^{d} (e\alpha)^m - (e\alpha)$$

As $\alpha \leq 0.15 < 1/e$, the sum on the right-hand side is a geometric series. Thus,

$$\sqrt{2\pi}\, \mathbb{P}_0\big(\, |\mathcal{A}(\mathbf{X}, \alpha)| > 0 \,\big) \leq \frac{e\alpha[1 - (e\alpha)^d]}{1 - e\alpha} - e\alpha \leq \frac{(e\alpha)^2}{1 - e\alpha} \tag{2.13}$$

We want to show that $\mathbb{P}_0\big(\, |\mathcal{A}(\mathbf{X}, \alpha)| > 0 \,\big) \leq \alpha$, i.e., that

$$\frac{(e\alpha)^2}{1 - e\alpha} \;\leq\; \sqrt{2\pi}\alpha\,. \tag{2.14}$$

Re-arranging, we find that $\alpha \leq .15 < \sqrt{2\pi}(e^2 + \sqrt{2\pi}e)^{-1}$ satisfies (2.14). □

CHAPTER 3
Differential Correlation Mining

3.1 Introduction

Given data obtained under two sampling conditions, it is often of interest to identify variables that behave differently in one condition than in the other. In this chapter, we present a method for differential association mining called **Differential Correlation Mining (DCM)**. The Differential Correlation Mining method identifies differentially correlated sets of variables, with the property that the average pairwise correlation between variables in a set is higher under one sample condition than the other. Differential Correlation Mining is a VSAT-style algorithm, so updates are performed via hypothesis testing of individual variables, based on the asymptotic distribution of their average differential correlation.

We refer to the target variable sets of Differential Correlation Mining as differentially correlated (DC) cliques. In a graph, a clique is a set of nodes that is fully connected, in the sense that there is an edge between every pair of nodes in the set. Informally, a DC clique is a set of variables such that each variable in the set has a positive (usually large) average differential correlation with the other variables in the set. More formally, let $\mathbf{R}_1, \mathbf{R}_2$ be the $d \times d$ population correlation matrices of the distributions underlying sampling conditions 1 and 2, respectively. Let $A \subset [d]$, where $[d]$ is the index set $\{1, ..., d\}$, and define

$$\Delta(j, A) \;=\; \frac{1}{|A|} \sum_{k \in A} (\mathbf{R}_1 - \mathbf{R}_2)_{jk} \tag{3.1}$$

to be the average difference of correlations between variable j and variables in index set A. Here the subscript jk denotes the element in the j-th row and k-th column of the corresponding matrix, and $|A|$ is the cardinality of the set A. We formally define DC cliques as follows.

Definition 3. *Let* $\mathbf{R}_1, \mathbf{R}_2$ *be given and let* $\Delta(\cdot, \cdot)$ *be defined as in (3.1). An index set* $A \subseteq [d]$ *with at least two elements is a* DC clique *for* $\mathbf{R}_1 - \mathbf{R}_2$ *if*

1. $\Delta(j, A) > 0$ *if and only if* $j \in A$,
2. *The set* A *cannot be written as a disjoint union of nonempty index sets* $A_1, A_2 \subset [d]$ *such that* A_1 *and* A_2 *satisfy condition 1 above.*

Condition 1 ensures that no relevant variables are omitted from a DC clique (every variable that is positively differentially correlated relative to the set A is included in A) and that a DC clique does not contain any extraneous elements. Condition 1 implies that a DC clique has larger average pairwise correlation under the first distribution than under the second. Condition 2 ensures that a DC clique cannot be subdivided into two smaller DC cliques. Importantly, the definition places *no conditions* on the correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. In particular, $\mathbf{R}_1$ and $\mathbf{R}_2$ need not be sparse, and need not satisfy any structural constraints such as bandedness. For a given pair $\mathbf{R}_1, \mathbf{R}_2$, it may happen that no DC cliques exist, or that the entire variable set forms a DC clique.

Note that the definition of DC cliques is not symmetric: in general, the DC cliques for $\mathbf{R}_1 - \mathbf{R}_2$ will be different from those for $\mathbf{R}_2 - \mathbf{R}_1$. The difference lies not in the relational structure itself, but rather in how we order the sample conditions (1 or 2). For example, in biological data, one sample group may involve a treatment condition, while the other is a reference or control group. A DC clique for $R_1 - R_2$ would contain genes that are more highly correlated in Condition 1 than Condition 2, for example, a protein pathway that is more active in Condition 1. This structure is illustrated in Figure 1.1.

The asymmetry in DC cliques could be eliminated by replacing the relevant section of (3.1) by a symmetric notion of difference such as $|\mathbf{R}_1 - \mathbf{R}_2|$. However, a variable set based on absolute difference (or similar) could contain a mixture of elements with positive correlation to A *and* elements with negative correlation to A. Such mixed groups would not exhibit the unified block structure of the type seen in Figure 1.1. Further, large variable sets with strong average negative correlation cannot occur. Simple algebra shows that since $\mathbf{R}_1$ is positive definite, the average pairwise correlation in Condition 1 of any set A with m elements must be at least $-\frac{1}{(m-1)}$.

As defined above, DC cliques are features of the underlying population distributions of the data. In practice, we will replace $\mathbf{R}_1, \mathbf{R}_2$ with estimates from observations, accounting for the uncertainty

in these estimators, to select empirical DC cliques. The broad objective of Differential Correlation Mining is to use observed data to identify DC cliques, or approximations of these, without prior knowledge of the identity, number, or size of the DC cliques present in the population. It is worth noting that the Differential Correlation Mining algorithm and supporting analysis described here are easily adapted to a non-differential correlation mining algorithm. An implementation of a correlation mining procedure is included along with the public DCM software.

Notation. In what follows, we assume that the data under condition 1 consists of n_1 independent samples drawn from a distribution F_1 with correlation matrix $\mathbf{R}_1$, and that the data under condition 2 consists of n_2 independent samples drawn from a distribution F_2 with correlation matrix $\mathbf{R}_2$. Let $\mathbb{X}_1 = (\mathbf{U}_1, ..., \mathbf{U}_d) \in \mathbb{R}^{n_1 \times d}$ and $\mathbb{X}_2 = (\mathbf{V}_1, ..., \mathbf{V}_d) \in \mathbb{R}^{n_2 \times d}$ denote the resulting data matrices in standard sample-by-variable form. Thus $\mathbf{U}_j \in \mathbb{R}^{n_1}$ denotes the measurements of variable j under condition 1, while $\mathbf{V}_j \in \mathbb{R}^{n_2}$ denotes the measurements of variable j under condition 2. Let $\mathbb{X}_{1,A} = (\mathbf{U}_j)_{j \in A}$ and $\mathbb{X}_{2,A} = (\mathbf{V}_j)_{j \in A}$ denote the restriction of $\mathbb{X}_1$ and $\mathbb{X}_2$, respectively, to a variable set $A \subset [d]$. Similarly, let $\mathbf{R}_{1,A}$ and $\mathbf{R}_{2,A}$ denote the correlation matrices under the distributions of F_1 and F_2 restricted to the variables in A.

Let $\tilde{\mathbf{U}}_j$ and $\tilde{\mathbf{V}}_j$ be the standardized versions of $\mathbf{U}_j$ and $\mathbf{V}_j$ respectively, such that $\|\tilde{\mathbf{U}}_j\| = \|\tilde{\mathbf{V}}_j\| = 1$, and define $\tilde{\mathbb{X}}_1 = (\tilde{\mathbf{U}}_1, ..., \tilde{\mathbf{U}}_d)$ and $\tilde{\mathbb{X}}_2 = (\tilde{\mathbf{V}}_1, ..., \tilde{\mathbf{V}}_d)$. Finally, let $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ denote the usual sample correlation matrices of $\mathbb{X}_1$ and $\mathbb{X}_2$, respectively (and $\widehat{\mathbf{R}}_{1,A}$ and $\widehat{\mathbf{R}}_{2,A}$ those of the appropriate restricted datasets). Thus $\left(\widehat{\mathbf{R}}_1\right)_{jk} = \widehat{\mathrm{cor}}\left(\mathbf{U}_j, \mathbf{U}_k\right) = \left(\tilde{\mathbb{X}}_1^t \tilde{\mathbb{X}}_1\right)_{jk}$ and a similar relation holds for $\widehat{\mathbf{R}}_2$.

3.2 The Differential Correlation Mining Method

The Differential Correlation Mining procedure has two main components: initialization and set update. These are discussed in detail in Sections 3.3 and 3.4. In brief, the Differential Correlation Mining procedure first employs a simple greedy algorithm to select an initial variable set A. Once the initial set is determined, it is passed to an update algorithm that iteratively refines the set, making use of a hypothesis testing framework to test variables for differential correlation. When an estimated DC clique is found, a residualization process prepares the data for further search by removing the differential correlation of the discovered set.

An important advantage of this type of approach is that the number and size of output sets are chosen adaptively based on testing principles. The Differential Correlation Mining method does not require pre-specification of number of clusters (as in kmeans), nor does it require an additional decision about cluster size (as in hierarchical clustering). Rather, the multiple testing procedure in the iterative step of Differential Correlation Mining naturally determines the number of variables in an output set. Differential Correlation Mining also differs from typical clustering procedures in that it does not require the calculation of a full $d \times d$ dissimilarity matrix, which can be a computational advantage in high dimensional data.

The Differential Correlation Mining procedure is summarized below. Detailed pseudocode is in Appendix A.

The Differential Correlation Mining Method

1. *Initialization:* Identify a good initial variable set A using a greedy algorithm that identifies a local maximum of a simple score function.

2. *Iteration:* Refine the initial set A. At each iterative step, repeat the following until termination.

   ▷ *Test* the differential correlation of each variable j with respect to A. Let A' be the set of variables with significant differential correlation, as determined by an FDR controlling multiple testing procedure.

   ▷ *Terminate* if $A' = A$ or a cycle is observed.

   ▷ *Update:* Set A to be A'.

3. *Return:* Output variable set A.

4. *Residualization:* Remove the effect of the DC clique A.

5. *Repeat* search with new initial set as many times as desired.

Iterative updating using multiple testing was first applied by Wilson et al. (2014) in the context of community detection for binary networks. Differential Correlation Mining makes use of the same search paradigm; however, a fundamentally different treatment is required to address differential correlation. In particular, the work of Wilson et al. (2014) performs hypothesis tests based on a fully constructed null model, whereas Differential Correlation Mining requires no structural assumptions on the null distribution of the data beyond equal correlation ($\mathbf{R}_1 = \mathbf{R}_2$) and some mild moment conditions (see Theorem 2).

3.2.1 Minor Algorithmic Details

Residualization In general, we expect multiple DC cliques in a dataset. The residualization step allows the Differential Correlation Mining procedure to search the same dataset many times, avoiding repeated results. Suppose an empirical DC clique A has been selected. Our approach is to estimate a rank one approximation of correlation matrices $\widehat{\mathbf{R}}_{1,A}$ and $\widehat{\mathbf{R}}_{2,A}$ via factor analysis (Harman, 1960). We then substitute the relevant submatrices, $\mathbb{X}_{1,A}$ and $\mathbb{X}_{2,A}$, with residualized data for which the estimated rank one correlation has been removed. Methods of estimation and removal of low-rank correlation have been well established in the literature. In the DCM software, we use the implementation of Friguet et al. (2012) for the R Statistical Software version and the method of Bishop (2006) for the Matlab version.

By opting for rank-one approximation, we are taking a conservative approach to residualization. It is conceivable that the correlation structure of A is of higher rank. If so, A may be selected more than once; however, since each time the data is being further residualized, we are guaranteed to eventually remove all structure of A. In practice, we have yet to encounter a duplicate result from real data.

Minimality. A nonempty fixed point A of the set update procedure has the property that, analogously to Definition 3, $H_0(j, A)$ is rejected if and only if $j \in A$. The second condition of Definition 3, however, is not guaranteed in general. It is possible that Differential Correlation Mining may select a large set that in truth consists of two (or more) disjoint population DC cliques. These cases are well addressed by the residualization step. When a conglomerate estimated DC clique is residualized, the *joint* structure is removed, leaving behind the individual structure of the

disjoint DC cliques. Further runs of the Differential Correlation Mining algorithm are then able to identify the separate DC cliques.

In extreme cases, the sampled data may be such that the disjoint DC cliques are, by chance, correlated enough to have negligible remaining individual structure after residualization. This correlation may render the multiple DC cliques indistinguishable in the data from a combined DC clique.

Cycles. Under certain conditions, the main search procedure terminates in a cycle of two or more sets. When the set update procedure oscillates between two sets A_1 and A_2, we restart the search on the intersection $A = A_1 \cap A_2$. In this case, the algorithm usually converges to fixed point in the vicinity of the intersection. If the oscillation persists, we output the intersection $A = A_1 \cap A_2$. This overlap set has the property that $H_0(j, A)$ will be rejected for all $j \in A_1, A_2$, so it is worth attention as an empirical DC clique.

Cycles of length greater than two are rarely observed in real or simulated data. However, to protect against longer cycles leading to infinite loops, the algorithm terminates at a maximum iteration limit.

Completion. In principle, the Differential Correlation Mining procedure can be run from many initial sets. In practice, we consider the procedure to have been "run to completion" if every variable has been included in at least one initial set and/or output set. Our implementation of the method is thus designed to randomly choose seed sets at each run from among the remaining unused variables. Note that this approach does **not** prevent variables from appearing in multiple output sets.

Data cleaning. Certain data artifacts that contradict our base assumptions can skew the DCM results. The software implementation is built to detect and remove (a) missing data, (b) rows or columns with more than 10% zeros, and (c) rows with approximately zero variance. Further, the software checks for extreme deviations from normality, which might indicate improper tail behavior, as well as large overall correlation difference between conditions. These cases are flagged for the user, but not forcibly prevented.

3.3 Initialization

The set update procedure in the second step of Differential Correlation Mining readily identifies variables that are significantly differentially correlated relative to a given variable set A, and is most effective when the initial set of variables exhibits at least low levels of differential correlation. (When applied to a randomly chosen set of variables, the set update procedure typically returns an empty set.) The core search procedure could be run exhaustively, beginning with every variable set $A \subset [d]$, but this is not computationally feasible for data sets of high or moderate dimension. As an alternative, we identify initial variable sets exhibiting a moderate degree of differential expression using a greedy search procedure. We then pass this initial skeleton clique to the set update process to be fleshed out into a final estimated DC clique.

The initialization procedure seeks a local maximum of the score function

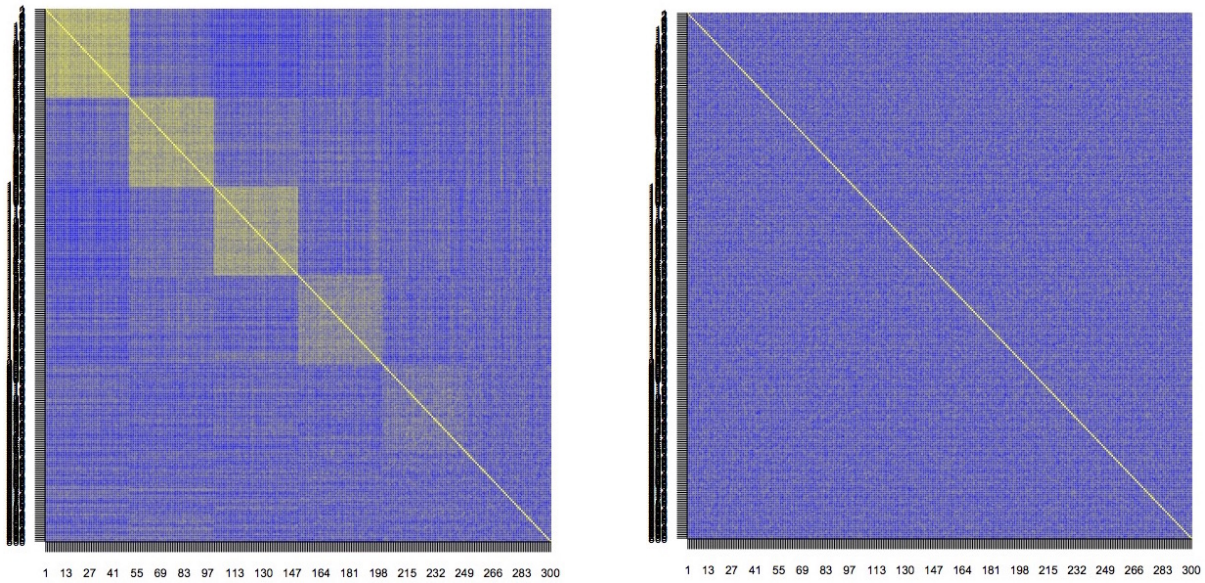
$$S(A) = \sum_{j,k \in A} \left\{ (n_1 - 3)^{1/2} \, \varphi\left(\widehat{\mathbf{R}}_1\right) - (n_2 - 3)^{1/2} \, \varphi\left(\widehat{\mathbf{R}}_2\right) \right\}_{jk} \tag{3.2}$$

where φ is the element-wise Fisher transformation of sample correlations, namely

$$\varphi(r) = \frac{1}{2} \log\left(\frac{1-r}{1+r}\right). \tag{3.3}$$

To find a local maximizer of $S(\cdot)$, we begin with a random seed A. We consider only pairwise swaps in which we replace an element of A with one from A^c. The set A is then updated by making the swap that produced the largest increase in the score. Since exactly one element is added and removed at each stage, the size of the variable set remains constant. Because of the random seeding, the algorithm is not purely deterministic. However, in practice the same local maximum is reached from most seeds.

We make use of the variance-stabilizing Fisher transformation in the initialization procedure as a way of roughly capturing *significance* of differential correlation instead of simply maximizing over absolute differences $\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2$. The transformation, and subsequent weighting by degrees of freedom, ensures that the first and second terms in the sum are approximately standardized. As such, sets maximizing $S(\cdot)$ are good ballpark guesses for true DC cliques. In the core set update procedure (Section 3.4), we employ a precise testing approach to measure significance of average



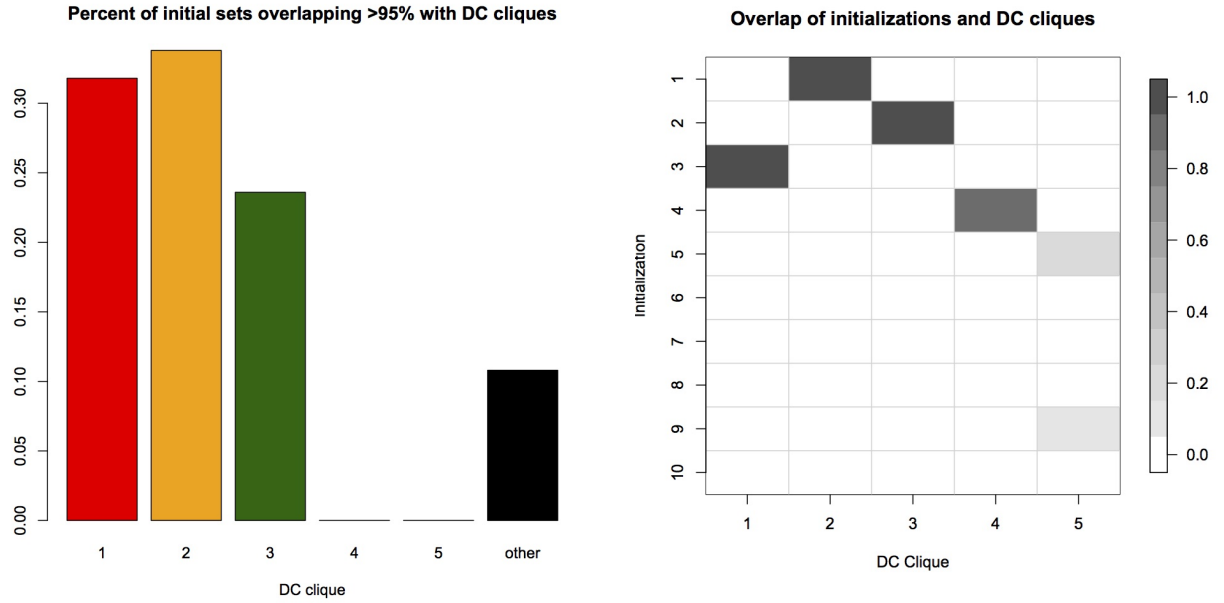
(a) Sample correlation, Condition 1 **(b)** Sample correlation, Condition 2

Figure 3.1: Sample correlation of simulated data.

differential correlation, so the initial sets need not be perfect. It is simply computationally more efficient to "warm-start" the algorithm with a reasonable set than to apply the core refinement procedure from random starting points.

Importantly, the cardinality of A is user-specified (with a default of 50). Due to the subsequent set update procedure, which adaptively chooses the size of a final output set A^*, we need not be completely confident in our choice of initial choice of cardinality. We also can generally expect results of the initialization procedure to be similar for similar cardinalities $|A| = m$. As an illustration of this phenomenon, we demonstrate the behavior of the initializing algorithm on artificial data. We generate 101 samples of a Gaussian random of 2,000 variables for each of two conditions. In Condition 2, the data is fully uncorrelated. In Condition 1, we include five correlated blocks with different correlation strength. Figure 3.1 shows the sample correlations for this simulated dataset.

It is clear that five distinct DC cliques are present, with decreasing signal size. A good initializing search procedure would have two properties: First, that when true DC cliques, selected sets of the correct size usually approximate these well; and second, that if the chosen cardinality m of the search procedure is too small or too large, selected sets will be sub- or super-sets of the true DC cliques. We find that our initializing method indeed exhibits these properties, as illustrated by Figures 3.2 and 3.3 for the artificial dataset.



(a) 500 initial sets, without removal

(b) First 10 initial sets, with removal

Figure 3.2: Overlap between initialized sets and DC cliques.

Figure 3.2(a) shows the percent of times, out of 500 separate runs with different random seeds, the initializing algorithm with $m = 50$ selected each of the DC cliques at less than 5% error. The algorithm selects one of the first three DC cliques nearly perfectly a high percentage of the time. Figure 3.2(a) shows 10 runs of the initializing algorithm, this time with the selected set removed from consideration in future seeds after each run. This figure shows that all five DC cliques are discovered to some degree in the first five runs of the initializing procedure. Although DC cliques 4 and 5 were never found in the 500 runs of 3.2(a), Figure 3.2(b) makes it clear that these lesser cliques are discoverable once the overshadowing signal of the stronger cliques is ignored.

In Figure 3.3, 5 distinct variable sets were selected for each value of m, and these are plotted according to their difference of average sample correlation. Colored points indicate that the set had at least 90% overlap with one of the true DC cliques in Figure 3.1. It is clear that even for misspecified m, the initializing procedure mostly selects sets that either contain or are contained by true DC cliques.

Pseudocode for the implementation of the initializing algorithm is provided as supplemental material. A closely related method is implemented in Section 3.6 for comparison with Differential Correlation Mining.

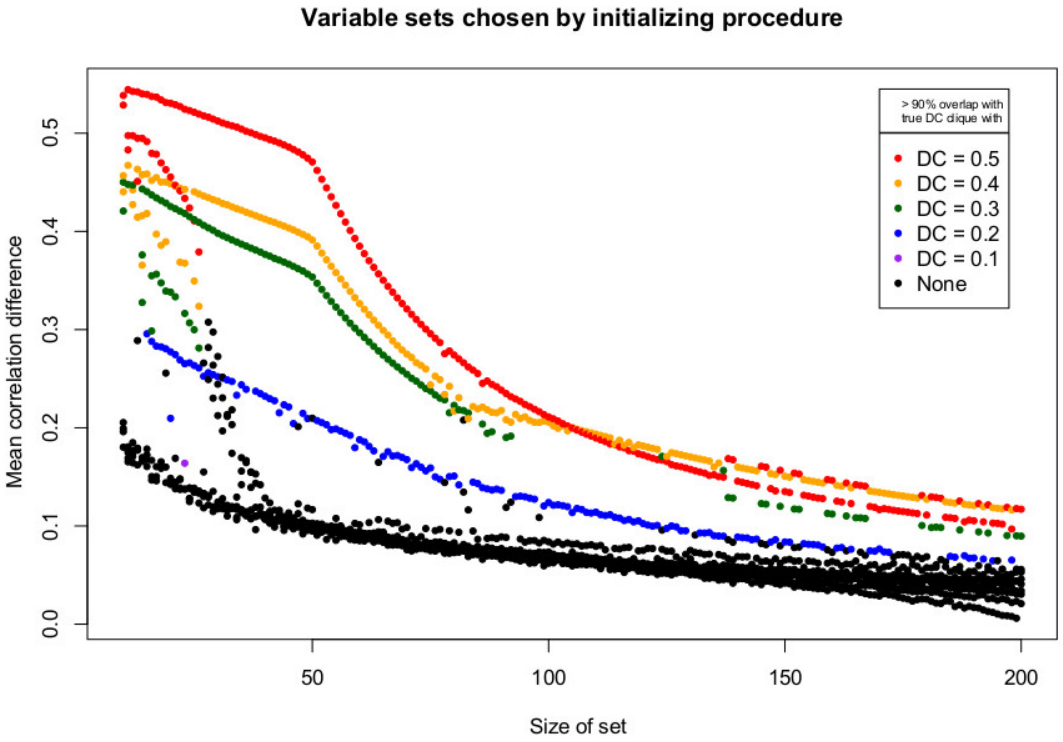


Figure 3.3: Initial sets at various sizes, colored by overlap with true DC cliques

3.4 Core set update procedure

The heart of the Differential Correlation Mining procedure is the set update algorithm, which makes use of multiple testing principles to iteratively refine a variable set A. Recall that the goal of Differential Correlation Mining is to estimate DC cliques from the data. To this end, the set update procedure is designed to identify variable sets exhibiting the properties of a true DC clique up to a level of statistical significance.

Consider a single iterative step, at which we update a given variable set A. We wish to determine whether each variable j (including those in A itself) ought to be included in the updated set A'. Since our eventual goal is to discover a DC clique, we perform hypothesis tests based upon the principles of Definition 3. For a given variable set A, the tests for variable j may be written as

$$H_0(j; A) : \Delta(j, A) = 0 \qquad \text{vs.} \qquad H_1(j, A) : \Delta(j; A) > 0\,. \tag{3.4}$$

Recall that $\Delta(j, A)$, as defined in (3.1), is a difference of average pairwise correlations between j and elements of A, so (3.4) is a test of differential correlation *relative* to the fixed set A. We then update the set A to $A' = \{j : H_0(j, A) \text{ was rejected}\}$ by simultaneous multiple hypothesis testing. This process continues until a fixed point $A = A'$ is reached.

To test the hypotheses in (3.4), we require a test statistic. A natural choice is the corresponding sample quantity,

$$\hat{\Delta}(j, A) = \frac{1}{|A|} \sum_{k \in A} (\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2)_{jk}. \tag{3.5}$$

In addition to being a straightforward choice, this test statistic exhibits several desirable properties discussed in Section 3.5.

Let $\delta(j, A)$ denote the realized value of the test statistic $\hat{\Delta}(j, A)$ for a particular dataset. It is clear that large positive values of $\delta(j, A)$ provide support for the alternate hypothesis in (3.4), while values that are negative or close to zero provide evidence in favor of the null. Thus, to test the hypotheses, for each $j = 1, \ldots, d$ we calculate a p-value of the form

$$\mathrm{p}(j, A) = \mathbb{P}_0 \left(\hat{\Delta}(j, A) > \delta(j, A) \right), \tag{3.6}$$

where the probability $\mathbb{P}_0$ is the (unknown) distribution of $\hat{\Delta}(j, A)$ under the null hypothesis $\Delta(j, A) = 0$. Since we make no assumptions about the distributions of data under Conditions 1 and 2 , we make use of asymptotic results to approximate the above probability. We show in Section 3.5.2 that, under appropriate regularity assumptions, and for large enough sample sizes n_1 and n_2,

$$\mathrm{p}(j, A) \approx 1 - \Phi \left(\frac{\delta(j, A)}{\hat{\sigma}_0(j, A)} \right), \tag{3.7}$$

where $\hat{\sigma}_0^2(j, A)$ is an estimate of the variance of $\hat{\Delta}(j, A)$ that can be computed from the available data. (The exact form of $\hat{\sigma}_0^2$ is given in Section 3.10.2.)

The collection of p-values $\{\mathrm{p}(j, A)\}_{j=1}^d$ measure the significance of the differential correlation of each variable relative to A. To select a set of significant variables A', we apply the modified FDR procedure of Benjamini and Yekutieli to the p-values. Specifically, we carry out the following steps

1. Order the p-values $\{\mathrm{p}(j, A)\}_{j=1}^d$ as $\{\mathrm{p}_{(1)}, ..., \mathrm{p}_{(d)}\}$.

2. Define the cutoff index k^* by

$$k^* = \max\left\{k \,:\, \mathrm{p}_{(k)} < \left(\sum_{j=1}^{d} 1/j\right)^{-1}\left(\frac{k\alpha}{d}\right)\right\}. \tag{3.8}$$

3. Let $A' = \{j \,:\, \mathrm{p}(j; A) \leq \mathrm{p}_{(k^*)}\}$.

Recall that we impose no assumptions on the structure of correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. In particular, it is possible that p-values $\mathrm{p}(j, A)$ and $\mathrm{p}(k, A)$ may be negatively correlated. For example, it is common in genetics for individual pairs of genes to exhibit negative correlation; in this case, a low p-value for one gene will imply a high p-value for the other. Most common multiple testing methods assume independence or positive dependency between p-values. The possibility of negative dependency of p-values necessitates a more conservative multiple testing method such as that of of Benjamini and Yekutieli (2001), which controls the expected False Discovery Rate at level α under negative dependence.

The main search procedure terminates when it degenerates ($A = \emptyset$) or converges ($A = A' \neq \emptyset$). For the degenerate case, the interpretation is simple: the initial set (chosen in the first step of the Differential Correlation Mining procedure) was not significantly differentially correlated. In the second case, we have identified an empirical DC clique, in the sense that by design, a nonempty fixed point A meets the first requirement of a DC clique in Definition 3 up to a level of statistical significance. The only other possible outcome of the iterative search procedure is a multi-set cycle, which is discussed in Section 3.2.1.

3.5 Properties of the Test Statistic

We now discuss some properties of the test statistic $\hat{\Delta}(j, A)$ used in the calculation of p-values for the set update procedure.

3.5.1 Geometric Interpretation

The equation for $\hat{\Delta}(j, A)$ given in (3.5) expresses the test statistic directly in terms of average differential correlation. However, we may also write $\hat{\Delta}(j, A)$ in an alternate form that yields an informative geometric interpretation. Let $\tilde{\mathbf{U}}_j \in \mathbb{R}^{n_1}$ and $\tilde{\mathbf{V}}_j \in \mathbb{R}^{n_2}$ be the standardized measurements

of variable j under sample conditions 1 and 2, respectively; and let

$$\mathbf{W}_1 := \frac{1}{|A|}\sum_{k\in A}\tilde{\mathbf{U}}_k \text{ and } \mathbf{W}_2 := \frac{1}{|A|}\sum_{k\in A}\tilde{\mathbf{V}}_k \tag{3.9}$$

be the vector means of the standardized measurements of the variables in A under each condition. It is easily shown that

$$\frac{1}{|A|}\sum_{k\in A}\widehat{\text{cor}}\left(\mathbf{U}_j,\mathbf{U}_k\right) = \mathbf{W}_1^t\tilde{\mathbf{U}}_j = \|\mathbf{W}_1\|\,\widehat{\text{cor}}\left(\tilde{\mathbf{U}}_j,\mathbf{W}_1\right)$$

and therefore

$$\hat{\Delta}(j,A) = \|\mathbf{W}_1\|\,\widehat{\text{cor}}\left(\mathbf{W}_1,\tilde{\mathbf{U}}_j\right) - \|\mathbf{W}_2\|\,\widehat{\text{cor}}\left(\mathbf{W}_2,\tilde{\mathbf{V}}_j\right).$$

Note that the vector $\tilde{\mathbf{U}}_j$ and the vectors $\{\tilde{\mathbf{U}}_k : k \in A\}$ lie on the surface of an (n_1-2)-dimensional sphere embedded in $\mathbb{R}^{n_1}$, and that $\mathbf{W}_1$ is the geometric center (centroid) of the latter collection. The norm $||\mathbf{W}_1||$ is between 0 and 1; large values of $||\mathbf{W}_1||$ correspond to the centroid being closer to the surface of the sphere, indicating that the vectors $\{\tilde{\mathbf{U}}_k : k \in A\}$ are tightly clustered, or equivalently, highly intercorrelated. Thus the quantity $\|\mathbf{W}_1\|\,\widehat{\text{cor}}\left(\mathbf{W}_1,\tilde{\mathbf{U}}_j\right)$ weights the similarity of $\mathbf{U}_j$ and the centroid $\mathbf{W}_1$ according to the overall similarity of the vectors $\{\tilde{\mathbf{U}}_k : k \in A\}$. Similar remarks apply to $\{\tilde{\mathbf{V}}_k : k \in A\}$ and $\mathbf{W}_2$. The statistic $\hat{\Delta}(j,A)$ is the difference of the summary measures in conditions 1 and 2.

Figure 3.4 gives a simple two-dimensional representation of the geometric picture discussed above. In Condition 1, $\mathbf{U}_j$ is not strongly correlated with $\mathbf{W}_1$, but $\|\mathbf{W}_1\|$ is large because the vectors indexed by A are tightly clustered. In Condition 2, $\mathbf{V}_j$ is strongly correlated with $\mathbf{W}_2$, but $\|\mathbf{W}_2\|$ is small because the vectors indexed by A are not tightly clustered. In this example, $\hat{\Delta}(j,A)$ is close to zero, and we would likely conclude no differential correlation is present.

3.5.2 Asymptotic distribution of the test statistic

We now discuss the asymptotic distribution of $\hat{\Delta}(j,A)$, from which the p-values used in Section 3.4 are derived. First, we make note of a classical result concerning sample correlations.

Theorem 2. (Steiger and Hakstian, 1982) *Let* $\mathbf{R}$ *be a* $d \times d$ *correlation matrix, and* $\widehat{\mathbf{R}}$ *the corresponding sample correlation matrix based on* n *i.i.d. samples of* d*-variate data with finite 4th*

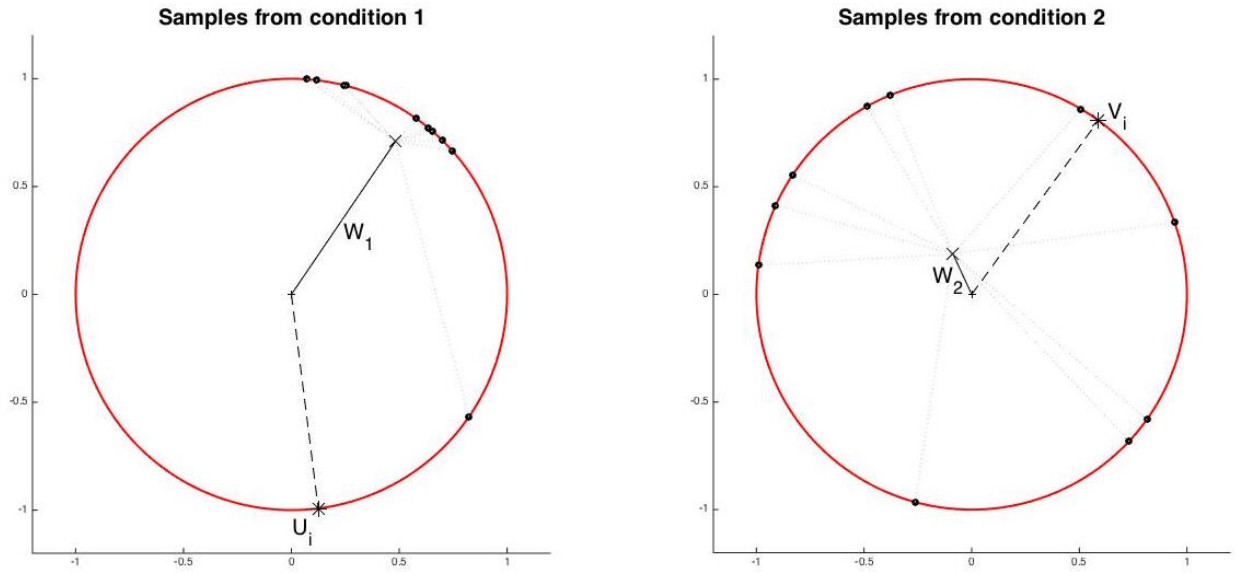


Figure 3.4: Geometric representation of data in two dimensions.

moment. Let $\mathbf{P}$ *and* $\widehat{\mathbf{P}}$ *be the vectorized versions of the matrices, of dimension* $d^2 \times 1$*. Then, as* n *tends to infinity*

$$\sqrt{n}\left(\widehat{\mathbf{P}} - \mathbf{P}\right) \Rightarrow \mathcal{N}_{d^2}\left(\mathbf{0}, \mathbf{\Sigma}\right),$$

where $\mathbf{\Sigma}$ *is a* $d^2 \times d^2$ *covariance matrix for which a general form is given equations (3.1-3.5) in Browne and Shapiro (1986).*

Using Theorem 2 one may evaluate the asymptotic distribution of $\hat{\Delta}(j, A)$, which is a function of $\mathbf{P}$ and $\widehat{\mathbf{P}}$.

Corollary 1. *Let* A *be a fixed index set and let* $\hat{\Delta}(j, A)$ *be defined as in* (3.5)*, with sample correlation matrices* $\widehat{\mathbf{R}}_1$ *and* $\widehat{\mathbf{R}}_2$ *based on* n_1 *and* n_2 *independent samples from distributions* F_1 *and* F_2 *respectively. Let* $\sigma_0^2(j, A) := \mathrm{var}\Big(\hat{\Delta}(j, A) \,|\, H_0\Big)$*, where* H_0 *is the null hypothesis in (3.4). Then, under* H_0*,*

$$\frac{\hat{\Delta}(j, A)}{\sigma_0(j, A)} \Rightarrow \mathcal{N}(0, 1) \tag{3.10}$$

as $\min(n_1, n_2) \to \infty$*.*

A proof of Corollary 1 is supplied in Section 3.10.1.

In practice, the variance $\sigma_0^2(j, A)$ is not known. We can use the results in Steiger and Hakstian (1982) for the asymptotic variance of $\hat{\Delta}(j, A)$, which leads to a consistent estimator $\hat{\sigma}_0(j, A)$, the

derivation of which is detailed in Section 3.10.2. We note that regardless of the size of A, the calculation of $\hat{\sigma}_0(j, A)$ requires basic operations on only three n_1 vectors and three n_2 vectors. Such algebraic simplification is important, since in practice the variance estimate must be calculated separately for *every* variable $j \in [d]$ and for multiple iterative steps of the Differential Correlation Mining algorithm.

Remark. The results of Corollary 1 apply to variable sets of fixed cardinality ($|A| = m$) as n_1 and n_2 tend to infinity. In practice, one may encounter variables sets for which $m > n_1, n_2$. Simulations suggest that the Differential Correlation Mining algorithm still identifies DC cliques with high success and controls false discovery in such settings even when the cardinality of $|A|$ greatly exceeds the sample size.

3.6 Simulation Study

To test the Differential Correlation Mining method against possible alternatives, we implemented a simple study of performance on simulated data. We created artificial datasets containing a single DC clique and compared the results of several methods to the known truth. Although the simulated setting is not a perfect representation of real data situations, it readily illustrates the advantages of Differential Correlation Mining as opposed to existing methods.

3.6.1 Simulated Data

We generated data with a single embedded DC clique, consistent with Definition 3. Our study varied the following parameters: size of the DC clique (m), total number of variables (d), strength of the true correlations in each sample condition (ρ_1 and ρ_2), and samples sizes of the two conditions (n_1 and n_2). In both sample conditions, the DC clique signal was layered on top of either (a) uncorrelated Gaussian noise or (b) a randomly real data sample from The Cancer Genome Atlas gene expression data. For an illustration of the form of the simulated data, refer to Figure 3.1 and the discussion thereof, where the data has five DC cliques rather than only one, but is otherwise generated in an identical fashion.

3.6.2 Methods implemented

To compare Differential Correlation Mining to alternate approaches, we implemented or adapted representative methods from those discussed in Section 1.3.2 to search for DC cliques.

Detection of isolated changes **(DCP)**. Although the goal of Differential Correlation Mining is to identify *sets* of variables, certain existing methods are designed to find *individual* (or isolated) variables whose correlations structure changes across conditions. The Differential Correlation Profile (DCP) method of Liu et al. (2010) is one such approach, using permutation of samples to determine the significance of correlation difference for each individual variable. Importantly, this approach identifies a list of individual differentially correlated variables, rather than a united set. For the purposes of this study, we treated the collection of selected variables as an estimated DC clique.

Mining a single correlation matrix **(WGCNA, NetTop)**. One approach to mining differential correlation is to analyze each sample condition separately, then compare results. The Network Topology (NetTop) method of Bockmayr et al. (2013) creates network representations for each of the two sample conditions by thresholding the corresponding Fisher-transformed sample correlation matrices. Connected components that appear in one network and not the other are considered to be differentially correlated variable sets.

The Weighted Gene Co-Expression Network Analysis (WGCNA) method of Langfelder and Horvath (2008) is a hybrid approach which mines for clusters (or "modules") in a single correlation matrix, then tests each module for differential *expression* across conditions. Thus, although the WGCNA method involves both differential and second-order elements, it is not designed to search for DC cliques or similar structures. For the purposes of this simulation study, we applied WGCNA to samples from condition 1 only. We then tested the output module for differential correlation via a standard t-test over sample correlations in conditions 1 and 2. In this way, we attempted to only select variable sets exhibiting differential correlation, even though WGCNA does not naturally identify modules with this property.

Mining dissimilarity matrices **(hclust, D-Est, DiffCoEx)**. Another possible approach is to summarize differential correlation in a single dissimilarity matrix, then select variable sets via ordinary clustering methods. We implemented a straightforward version of this approach, applying

hierarchical clustering to the difference of sample correlation matrices, $\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2$. To circumvent the challenge of selecting a cutoff in the dendrogram, we instead chose the first cluster of size less than or equal to the true DC clique. (In practice, the true size would not be known, so we would be less sure of the "best" cutoff point.) We also applied this idealized hierarchical clustering to $\widehat{\mathbf{D}}$, the estimator suggested in Cai and Zhang (2014) for directly estimating $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. Finally, the DiffCoEx method of Tesson et al. (2010) is a modification of WGCNA; a dissimilarity matrix is created based on adjusted sample correlations, then the clustering approach of WGCNA is applied.

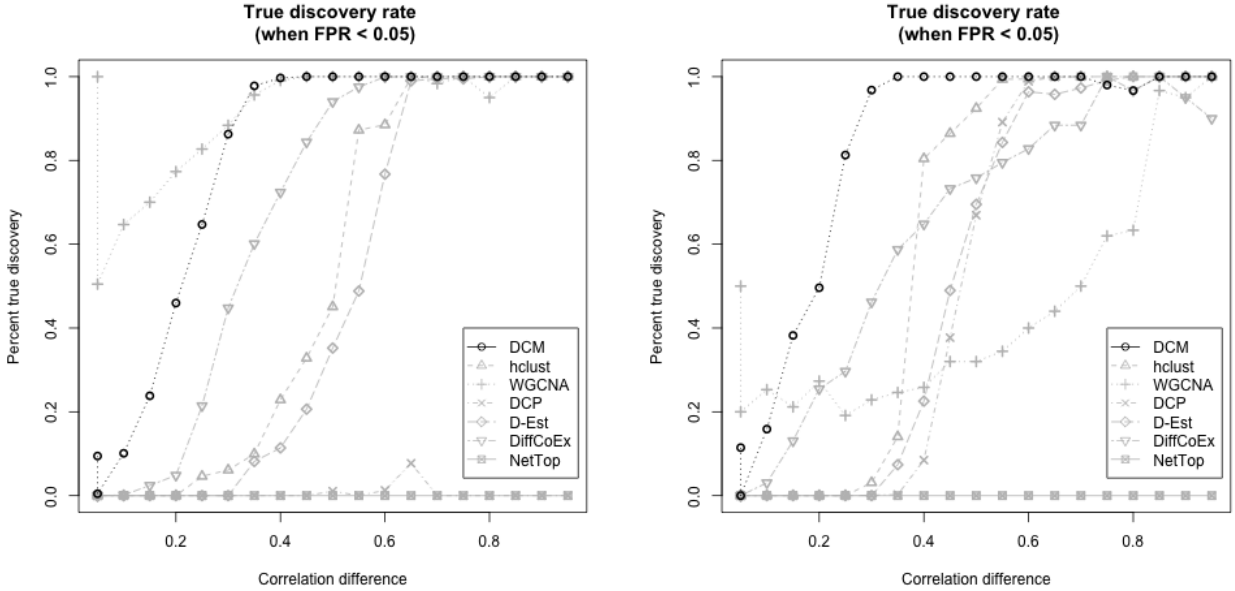
3.6.3 Results

We applied the seven proposed methods (DCM, DCP, NetTop, WGCNA, hclust, D-EST, and DiffCoEx) to several simulated datasets at each of many parameter combinations. We found that all methods behaved similarly with regard to changes in sample sizes n_1, n_2 and clique size m (relative to d). Here, we present only the results regarding the correlation signal size (ρ_1 vs. ρ_2) and the different background types, to illustrate key differences in performance between methods. By default, the other parameters were set to be $n_1 = n_2 = 100$, $m = 100$, and $d = 1000$.

The success of the methods was measured by the false positive rate (FPR), the percentage of variables in a selected set that were not in the seeded DC clique, and the true discovery rate (TDR), the percentage of detected variables from the true DC clique. That is, if B was the output variable set of a procedure and $A = (1, \ldots, m)$ was the embedded DC clique, then

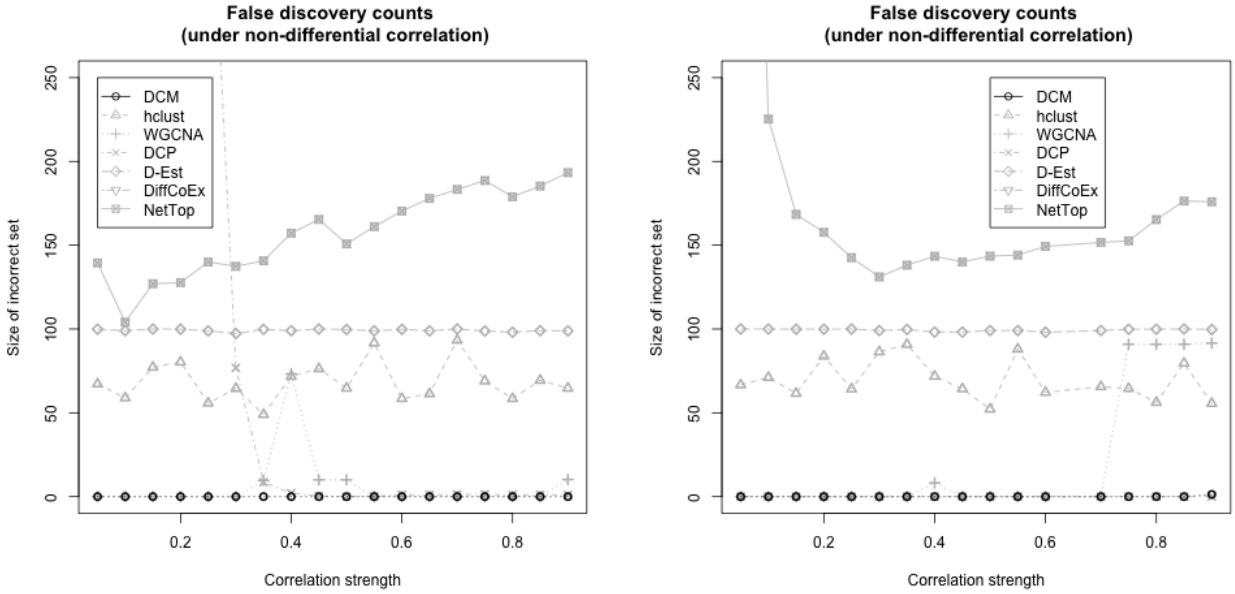
$$\text{False Positive Rate} = \frac{|B \setminus A|}{|B|} \quad \text{and} \quad \text{True Discovery Rate} = \frac{|B \cap A|}{|A|}.$$

To control false discovery, we disregarded output variable sets with more than 5% FPR. Figure 3.5 shows the percent of variables in the seeded DC clique that were successfully identified by each method (the TDR) after false discovery screening, for various strengths of true differential correlation ($\rho_1 - \rho_2$ grows). Figure 3.6 examines the scenario where $\rho_1 = \rho_2 \neq 0$; that is, when correlation was present in both sample conditions but not differential. Figure 3.6 shows the size of selected variable sets - ideally, DC mining methods would return no results in these cases. All result reflect an average of 10 simulations at each data point, with all other parameters set to default values.



(a) Gaussian noise background **(b)** Real data background

Figure 3.5: True discovery rates when false positive controlled at 0.05 level.



(a) Gaussian noise background **(b)** Real data background

Figure 3.6: Sizes of incorrect variables sets when no differential correlation is present.

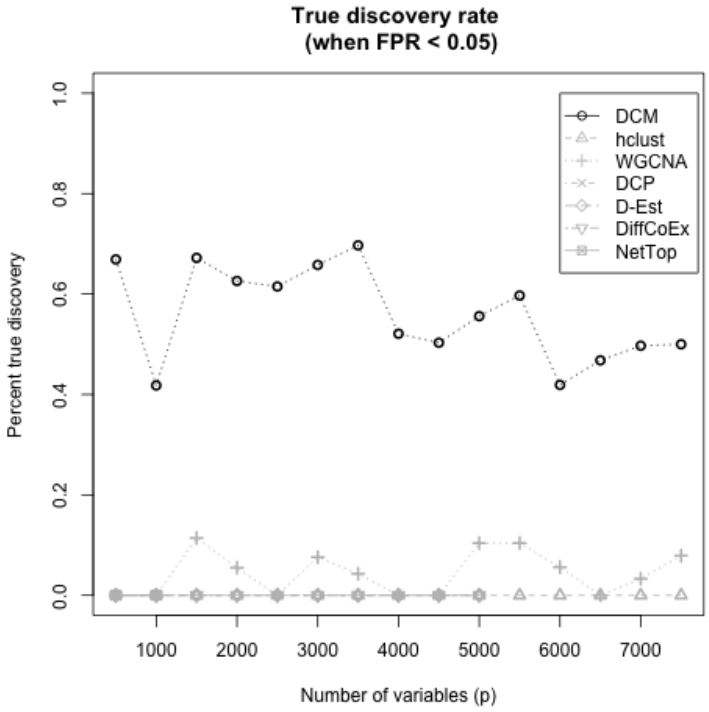


Figure 3.7: Detection rate for various dimensions.
($m = 100, \rho_1 = 0.3, \rho_2 = 0.1$)

DCM was able to control false positives in all cases except for some error when there was very low signal in the real data background, which may be due to actual signal being present in the randomized real data. Differential Correlation Mining also began to reliably detect DC cliques at a lower signal (around a correlation difference of 0.2 at the default parameters) than every method except WGCNA with Gaussian background. We find that this discovery rate is not noticeably affected by the total number of variables d; Figure 3.7 provides evidence of stable discovery rate for Differential Correlation Mining over different values of d.

In randomized real data (Figure 3.5b), **WGCNA** did not control the false positive rate. This is because WGCNA is a method for non-differential analysis, so when applied to Condition 1 data, it (correctly) identifies many correlated variables - even though these are often equally correlated in Condition 2. Although we have adapted the method to test selected modules for differential correlation, true DC cliques are obscured by existing non-differential structure.

The **hclust** and **D-EST** approaches behave as expected: because we chose a cutoff of the hierarchical clustering dendrogram by size, our approach necessarily returns a nonempty variable

set. This caused the false positive rate to be high for small or no signal. Similarly, **NetTop** relies on a thresholding procedure to maximize differences between conditions, so it is likely to find signal even when none is present. However, even if the false positives were perfectly controlled in some way, these methods show a lower detection point than Differential Correlation Mining.

DiffCoEx performed the strongest in our simulations, as it was able to control false discovery in most cases while still detecting DC cliques at a reasonable rate. Differential Correlation Mining, however, proved more sensitive without sacrificing error control.

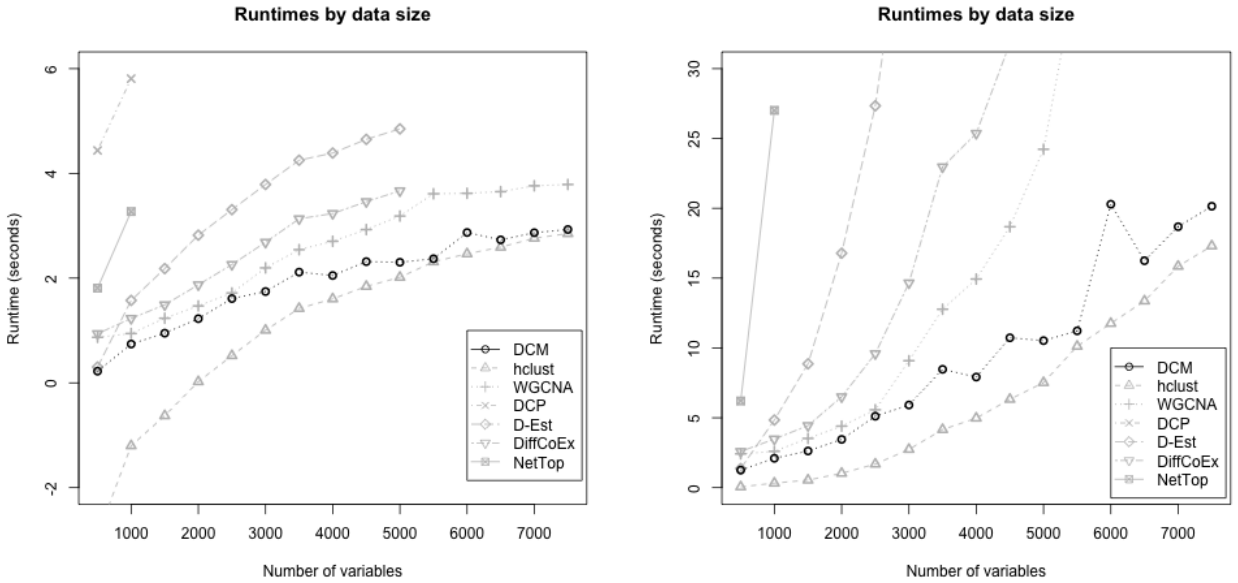
Finally, **DCP**, and any approach that seeks isolated structure rather than unified sets, is likely to greatly overselect variables in the uncorrelated background case because the mutual behavior of the variables in a DC clique will induce some correlation structure in the extraneous variables. Figure 1.1 illustrates this phenomenon, as there is some pattern in the cross correlation between variables in B and A. This result emphasizes the danger of the common approach of looking for isolated changes in correlation structure of individual variables, rather than searching for DC cliques: vestigial correlation patterns may be misleading.

Remark. We also implemented versions of the iterative testing update procedure using different hypothesis testing approaches, including a Normal approximation to Fisher-transformed data and a classic likelihood ratio test as derived in Muirhead (1982). We found that neither approach yielded a higher discovery rate (with controlled FDR) than Differential Correlation Mining.

3.6.4 Computation

Figure 3.8 shows the computation times for all tested methods on a log scale and an absolute scale. Since modern datasets tend to have dimension in the tens or hundreds of thousands of variables, the exponential differences between method runtimes are crucial to the practicality of analysis. All methods except the basic hclust required exponentially more runtime than Differential Correlation Mining.

One important limitation of common approaches to correlation mining (including DCP, D-Est, hclust, and NetTop) is that memory demands scale on the order of at least d^2, as they necessitate estimation of full d by d dissimilarity matrices. Permutation- or repetition-based methods such as DCP and NetTop are even more infeasible in high dimensions, since they require the computation of a d by d correlation matrix for each of many permutations (this is why the simulations were



(a) Log scale **(b)** Exact scale

Figure 3.8: Computation time to find a single variable set.

truncated in Figure 3.8). An advantage of Differential Correlation Mining is that only a the $|A| \times d$ portion of sample correlation matrices corresponding to proposed set A must be computed at any given time.

3.7 Data Analysis: TCGA

As introduced in Figure 1.1, we applied the Differential Correlation Mining procedure to data from The Cancer Genome Atlas, with samples from two pre-determined breast cancer subtypes: Her-2 and Luminal B. The dataset consisted of 51 tissue samples from the Her-2 subtype and 152 samples from the Luminal B subtype. A total of 14 empirical DC cliques (more correlated in Her-2 than in Luminal B) were discovered, ranging in size from 8 to 102 genes. These sets are summarized in Table 3.1, which is ordered by a rough measure of "signal" calculated from the square root of the set size multiplied by the average differential correlation of the set. The gene memberships of the sets are available in Table B.1 in Appendix B.

To illustrate how this information may be useful to genomic research, we briefly discuss one of the discovered gene sets. The set of interest contained 48 genes, listed alphabetically in Table 3.2. These genes are found to be highly associated with immune response, particularly the HLA

Table 3.1: Summary of DC cliques found in TCGA data

Label	Size	Mean Corr, Her-2	Mean Corr, Lum-B
1	31	0.85	0.05
2	102	0.74	0.40
3	48	0.62	0.04
4	22	0.89	0.07
5	73	0.48	0.07
6	59	0.48	0.03
7	123	0.35	0.05
8	63	0.53	0.18
9	30	0.45	0.08
10	32	0.52	0.16
11	25	0.50	0.16
12	15	0.48	0.08
13	13	0.49	0.07
14	8	0.42	0.09

Table 3.2: Genes selected in empirical DC Clique for Her2 vs. Luminal B samples.

AGER	amt	APOL1	ARPC4	B2M	BATF2	BTN3A2
BTN3A3	C19orf38	calml4	CCDC146	CHKB-CPT1B	echdc1	ETV7
EXOSC10	FBXO6	GBP1	GBP4	GJD3	gnb3	**HLA-A**
HLA-B	**HLA-C**	**HLA-E**	**HLA-F**	**HLA-H**	HSH2D	IDO1
IL15	Irf1	LOC115110	LOC400759	LOC91316	micB	Myo15b
OASL	PILRB	Rec8	Rufy4	SAMD9L	SEC31B	STAT1
tap1	Tapbp	TTLL3	TXNDC6	Ube2l6	Zbp1	

(Human Leukocyte Antigen) gene class, represented by six of the genes in the set (emphasized in bold). Researchers are interested in understanding how and why some cancer subtypes trigger immune response while others do not. For example, Iglesia et al. (2014) showed that prognosis was improved for patients with Her-2 and Basal-like subtypes showing higher immunoreactive response. Further exploration of DC cliques such as the one in Table 3.2 may further understanding of the gene interactions that drive immune response.

Although no methods besides Differential Correlation Mining are feasible for data of this size, we compared the performance of Differential Correlation Mining and related methods in a limited set of the TCGA data. We included the first large DC clique selected by Differential Correlation Mining (size 102) and 500 randomly selected genes. Table 3.3 shows the output of competing methods applied to this data. For the primary selected set for each method, we measure the number of genes that overlapped with the DC clique and the number that did not. Using the

Table 3.3: Results from competing methods, compared to Differential Correlation Mining result

Method	Size Found	Num in DC Clique	FDR	TDR
hclust	270	101	0.63	0.99
WGCNA	87	74	0.15	0.73
DCP	56	50	0.11	0.49
D-Est	100	71	0.29	0.70
DiffCoEx	6	6	0.00	0.06
NetTop	332	99	0.70	0.97

selected DC clique from Differential Correlation Mining as a reference, we compute the False and True Discovery Rates for each method as in Section 4.

3.8 Data Analysis: The Human Connectome Project

The Human Connectome Project is a multi-institutional venture aimed at mapping functional connections between parts of the human brain. The project has collected vast amounts of brain scan data, all of which is publicly available to researchers online at `www.humanconnectome.org`.[1] In this analysis, we made use of a dataset from the "500 Subjects MR" data release, which consists of functional magnetic resonance imaging (fMRI) brain scans for 542 healthy adult subjects. Participants performed a variety of tasks during the MR scan, designed to isolate certain types of brain functionality. Measurements of brain activation were taken at frequent time steps over the course of the tasks (316 steps for language tasks; 284 for motor tasks) at locations corresponding to ∼30,000 voxels (the brain's white matter interior) and ∼60,000 greyordinates (the grey matter brain surface). We applied Differential Correlation Mining to data from a single subject.[2] Our analysis compared two task categories:

Language-based tasks: During the scan, subjects were told brief stories and asked to answer questions after each one about what they were told.

Motor-based tasks: Subjects were attached to motion sensors at the hands, feet, and tongue. They were then asked to move one appendage at a time, in blocks of repetitions.

[1]Data was available in pre-processed form; see `http://www.humanconnectome.org/about/project/MR-preprocessing.html` for further detail.

[2]Subject #101006, a 35-year-old female.

Differential Correlation Mining was applied the data for 91,282 brain locations to find DC cliques of voxels and greyordinates that exhibit more correlation over time during language tasks than during motor tasks, as measured by sample correlation across measurements at time steps. On a home computer, this process took under a minute to find the first DC clique, running in Matlab. Continuing to completion took approximately an hour. No additional methods were applied, as the dataset was too large to be computationally feasible for any of the approaches suggested in Section 3.6. The DCM algorithm discovered 10 total empirical DC cliques, summarized in Table 3.4.

Table 3.4: Summary of DC cliques found in Human Connectome Data

Label	Size	Mean Corr, Lang Tasks	Mean Corr, Motor Tasks
1	1688	0.2000	0.1000
2	137	0.2044	0.0506
3	407	0.1856	0.0143
4	111	0.2497	0.0359
5	377	0.1658	0.0097
6	82	0.3253	0.0639
7	266	0.1649	0.0121
8	259	0.1482	0.0098
9	198	0.1732	0.0116
10	20	0.2981	0.1019

The first empirical DC clique selected by Differential Correlation Mining is very large, containing 1688 nodes located on the cortical surface. These nodes, or "greyordinates", are visualized as points on the smoothed exterior of the brain in Figure 3.9. The clear locational pattern in the nodes - despite the fact that the analysis did not take location into account - is striking. Additionally, the empirical DC clique in Figure 3.9 includes a concentrated group in the rear of the left cortex. This general brain region is known to be specifically associated with language processing and auditory input (Wernicke's Area, see Wang et al. (2015)).

We also studied two other artifacts of the data for comparison, displayed in Figure 3.10. First, we identified the 1000 nodes exhibiting the strongest differential first-order behavior. These show higher mean activation during the language tasks than during the motor tasks, as measured by standard two-sample t-tests. We saw a clear grouping of nodes in the right frontal lobe. This pattern is unsurprising and appears in many studies of brain functionality that examine differential activation for language processing (Voets et al., 2006). This basic first-order analysis suggests that differential correlation is not redundant. None of the empirical DC cliques selected by Differential

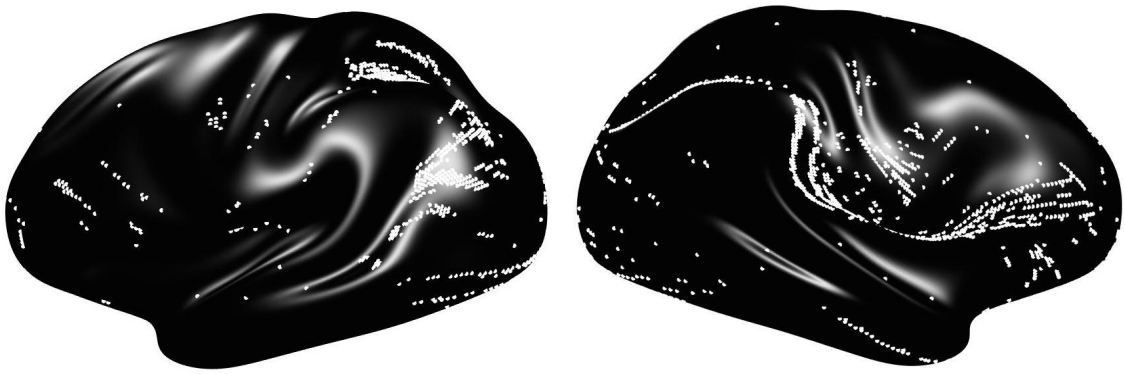
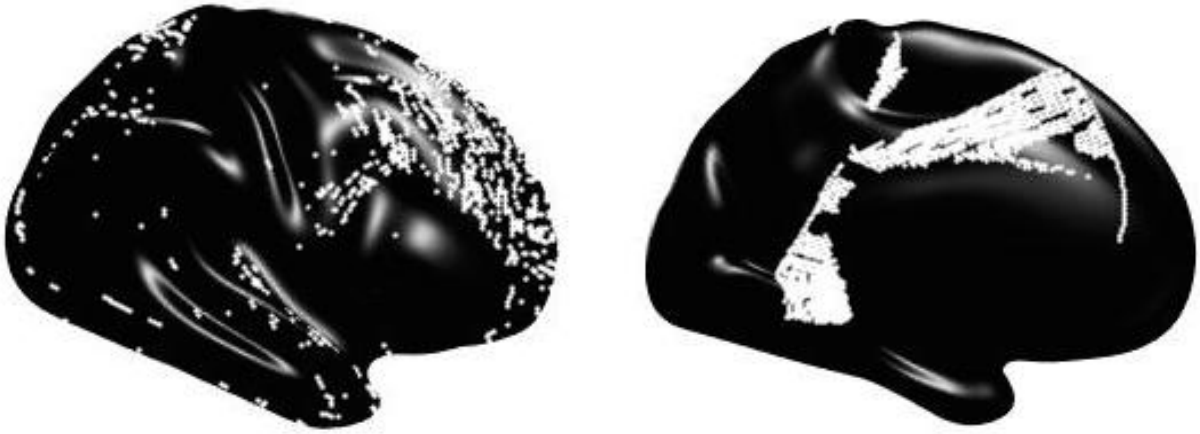


Figure 3.9: Brain locations of DC clique for languages tasks versus motor tasks.



(a) High differential mean activation. *(Right cortex, exterior view.)*

(b) High correlation during language tasks. *(Left cortex, interior view.)*

Figure 3.10: Brain locations showing high first-order differences and high non-differential correlation.

Correlation Mining show high frontal lobe concentration; instead, they exhibit "trail-like" patterns such as the ones shown in Figure 3.9.

Second, we identified 1000 nodes found to be highly correlated over time for the language task data, irrespective of their behavior in the motor task data. These nodes were observed to be very tightly grouped in the interior left hemisphere. This is likely due to the nature of data measurement: fMRI brain scans measure oxygen flow in the brain, so measurements for adjacent regions tend to "blur" and show high artificial correlation (Derado et al., 2010). In this case, the same node set is also highly correlated during motor tasks, suggesting that it is likely a byproduct of data collection. Even if this node set does represent a meaningful result - regions, perhaps, that are universally correlated regardless of task - it is not differential.

This example illustrates the advantage of taking a differential approach like Differential Correlation Mining. Effects due to fMRI-driven spatial correlation or strong universal correlation can drown out signal that is truly specific to a particular sample condition. By comparing language tasks to the similar but distinct condition of motor tasks, we are able to isolate signals that are unique to language processing. The fact that the identified DC cliques show emergent locational patterns suggests that Differential Correlation Mining is capturing a true facet of the data rather than arbitrary correlation. Since this output is unique in form, while maintaining some consistency with known brain functionality, we believe it merits further scientific investigation.

3.9 Discussion

There is ample motivation in data for methods of differential second-order analysis, especially in the area of statistical genetics, where analyses of differential correlation are beginning to emerge. We argue that the Differential Correlation Mining method represents an important new tool in differential association mining. There are three main advantages of Differential Correlation Mining over existing methods:

1. Differential Correlation Mining is designed to search specifically for DC cliques, a precisely defined population quantity. Simulation suggests that the Differential Correlation Mining method will detect cliques within reasonable error at a much lower signal threshold than existing methods. We believe the DC clique structures has scientific merit in many settings, including those demonstrated in the data analyses in this chapter.

2. Since Differential Correlation Mining is an VSAT-type algorithm, with foundations in classical and asymptotic theory, the analysis accounts for random behavior and results are interpretable in a hypothesis testing framework. In particular, control of false positives is guaranteed theoretically in ideal settings and holds in complex simulations.

3. The initialization and core update procedures of Differential Correlation Mining do not require the computation of a $d \times d$ dissimilarity matrix, nor do they rely on permutation scores. As such, DCM has low memory demands and computation time even for very large data, without sacrificing accuracy. The efficiency of DCM allowed us to study differential correlation in two

very high dimensional settings: gene expression data ($\sim 10^4$ variables) and fMRI brain scan data ($\sim 10^5$ variables). Both these datasets are beyond the computation limits of the alternate methods discussed in Section 3.6 without access to extraordinary computing resources.

Software packages in R and Matlab for the Differential Correlation Mining procedure are publicly available at `http://github.com/kbodwin/Differential-Correlation-Mining`.

3.10 Proofs and Derivations

3.10.1 CLT for difference of sample correlations (Corollary 1)

Let A be a fixed index set and let $\hat{\Delta}(j, A)$ be defined as in (3.5), with sample correlation matrices $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ based on n_1 and n_2 independent samples from distributions F_1 and F_2 respectively. Let $\sigma_0^2(j, A) := \mathrm{var}\Big(\hat{\Delta}(j, A) \,|\, H_0\Big)$, where H_0 is the null hypothesis in (3.4). Then, under H_0,

$$\frac{\hat{\Delta}(j, A)}{\sigma_0(j, A)} \Rightarrow \mathcal{N}(0, 1) \tag{3.11}$$

as $\min(n_1, n_2) \to \infty$.

Proof: For clarity, we first examine only one "half" of $\hat{\Delta}(j, A)$. Let

$$\bar{r}_1(j, A) \,=\, \frac{1}{|A|} \sum_{k \in A} \Big(\widehat{\mathbf{R}}_1\Big)_{jk} \quad \text{and} \quad \bar{\rho}_1(j, A) \,=\, \frac{1}{|A|} \sum_{k \in A} (\mathbf{R}_1)_{jk} \,. \tag{3.12}$$

Note that $\bar{r}_1(j, A)$ is a linear function of $\widehat{\mathbf{R}}_1$ and that $\bar{\rho}_1(j, A)$ is the same function applied to the population correlation matrix $\mathbf{R}_1$. It follows from Theorem 2 that

$$\sqrt{n_1} \left(\frac{\bar{r}_1(j, A) - \bar{\rho}_1(j, A)}{\tau_1^2(j, A)} \right) \Rightarrow \mathcal{N}(0, 1) \,, \tag{3.13}$$

with $\tau_1^2(j, A) := \mathrm{var}\big(\sqrt{n_1}\, \bar{r}_1(j, A)\big)$, which has a finite limiting value that can be expressed as the mean of appropriate elements of the covariance matrix $\mathbf{\Sigma}$ in the theorem. To apply this result for the full test statistic, we note that $\hat{\Delta}(j, A) = \bar{r}_1(j, A) - \bar{r}_2(j, A)$. Samples from F_1 are independent of those from F_2, so $\bar{r}_1(j, A)$ is independent of $\bar{r}_2(j, A)$, and thus $\hat{\Delta}(j, A)$ is asymptotically normal.

Under the null hypothesis in (3.4), $\bar{\rho}_1(j, A) = \bar{\rho}_2(j, A)$, and hence the mean of the limiting distribution of $\hat{\Delta}(j, A)$ is 0. The variance of $\hat{\Delta}(j, A)$ can be expressed as the weighted sum

$$\sigma_0^2(j, A) \;=\; \frac{\tau_1^2(j, A)}{n_1} + \frac{\tau_2^2(j, A)}{n_2}\,. \tag{3.14}$$

□

3.10.2 Variance Estimator

Let r_{jk} be the sample correlation of $\mathbf{U}_j$ and $\mathbf{U}_k$, and let $r_A := \frac{1}{|A|}\sum_{k\in A} r_{jk}$. Let $\mathbf{Y}$ and $\mathbf{W}$ be vectors of length n_1 such that for $i = 1, 2, ..., n_1$

$$W_i \;:=\; \frac{1}{|A|}\sum_{k\in A}\tilde{U}_{ik}\,, \quad \textit{and} \quad Y_i \;:=\; \frac{1}{|A|}\sum_{k\in A} r_{jk}\tilde{U}_{ik}^2\,. \tag{3.15}$$

Let $\hat{\tau}_1(j, A)$ be the consistent variance estimator given by equation (5.1) of Steiger and Hakstian (1982),

$$\hat{\tau}_1 \;=\; \frac{1}{|A|^2}\sum_{k,l\in A}\left[r_{jjkl} \;+\; \frac{1}{4}r_{jk}r_{jl}(r_{jjjj} + r_{jjkk} + r_{jjll} + r_{kkll})\right], \tag{3.16}$$

where

$$r_{jkls} \;:=\; \sum_{i=1}^{n_1}\tilde{U}_{ij}\tilde{U}_{ik}\tilde{U}_{il}\tilde{U}_{is}\,. \tag{3.17}$$

An equivalent form for $\hat{\tau}_1(j, A)$ is given by

$$\hat{\tau}_1(j, A) \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\frac{r_A^2}{4}\tilde{U}_{ij}^4 \;-\; r_A W_i\,\tilde{U}_{ij}^3 \;+\; \left(\frac{r_A Y_i}{2} \;+\; W_i^2\right)\tilde{U}_{ij}^2 \;-\; W_i Y_i\,\tilde{U}_{ij} \;+\; \frac{Y_i^2}{4}\right\}. \tag{3.18}$$

Proof. We begin by expanding (3.16),

$$\begin{aligned}\hat{\tau}_1 \;&=\; \frac{1}{|A|^2}\sum_{k,l\in A}\Big[r_{jjkl} \;+\; \frac{1}{4}r_{jk}r_{jl}(r_{jjjj}+r_{jjkk}+r_{jjll}+r_{kkll})\\ &\quad -\; \frac{1}{2}r_{jk}(r_{jjll}+r_{jkkl}) \;-\; \frac{1}{2}r_{jl}(r_{jjjk}+r_{jkll})\Big]\\ &=\; \frac{1}{|A|^2}\sum_{k,l\in A} r_{jjkl} \;+\; \frac{1}{4}r_A^2 r_{jjjj} \;+\; \frac{1}{2|A|}\sum_{j\in A} r_{jk}r_A r_{jjkk} \;+\; \frac{1}{4|A|^2}\sum_{k,l\in A} r_{jk}r_{jl}r_{kkll}\\ &\quad -\; \frac{1}{|A|}\sum_{j\in A} r_A r_{jjjk} \;-\; \frac{1}{|A|^2}\sum_{k,l\in A} r_{jk}r_{jkkl}\,.\end{aligned}$$

We then derive equivalent matrix forms for each summation. Here $\circ$, as in $\mathbf{W}^{\circ 2}$, denotes elementwise exponentiation of a vector.

$$\frac{1}{|A|^2}\sum_{k,l\in A} r_{jjkl} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{1}{|A|^2}\sum_{k,l\in A}\tilde{U}_{ik}\tilde{U}_{il}\right)\tilde{U}_{ij}^2 \;=\; \frac{1}{n_1}\left(\mathbf{W}^{\circ 2}\right)^t\,\tilde{\mathbf{U}}_i^{\circ 2}\,.$$

$$\frac{1}{|A|}\sum_{k\in A} r_{jk}r_{jjkk} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{1}{|A|}\sum_{k\in A} r_{jk}\tilde{U}_{ik}^2\right)\tilde{U}_{ij}^2 \;=\; \frac{1}{n_1}\mathbf{Y}^t\,\tilde{\mathbf{U}}_i^{\circ 2}\,.$$

$$\frac{1}{|A|^2}\sum_{k,l\in A} r_{jk}r_{jl}r_{kkll} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{1}{|A|}\sum_{k\in A} r_{jk}\tilde{U}_{ij}^{\circ 2}\right)^2 \;=\; \frac{1}{n_1}\mathbf{Y}^t\,\mathbf{Y}\,.$$

$$\frac{1}{|A|^2}\sum_{k,l\in A} r_{jk}r_{jkkl} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{1}{|A|}\sum_{k\in A} r_{jk}\tilde{U}_{ij}^2\right)\left(\frac{1}{|A|}\sum_{k\in A}\tilde{U}_{ik}\right)\tilde{U}_{ij} \;=\; \frac{1}{n_1}\left(\mathbf{W}\cdot\mathbf{Y}\right)^t\,\tilde{\mathbf{U}}_i\,.$$

Substituting the above into the expanded equation for $\hat{\tau}_1(j,A)$ gives

$$n_1\hat{\tau}_1 \;=\; \frac{1}{4}r_A^2\mathbf{1}^t\tilde{\mathbf{U}}_i^{\circ 4} \;+\; r_A\left[\frac{1}{2}\mathbf{Y}^t\,\tilde{\mathbf{U}}_i^{\circ 2} \;-\; \mathbf{W}^t\tilde{\mathbf{U}}_i^{\circ 3}\right] \;+\; \left(\mathbf{W}^{\circ 2}\right)^t\,\tilde{\mathbf{U}}_i^{\circ 2} \;-\; \left(\mathbf{W}\cdot\mathbf{Y}\right)^t\,\tilde{\mathbf{U}}_i \;+\; \frac{1}{4}\mathbf{1}^t\,\mathbf{Y}^{\circ 2}\,,$$

and we may rewrite the vector operations into a single summation over elements,

$$\hat{\tau}_1(j,A) \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\frac{r_A^2}{4}\,\tilde{U}_{ij}^4 \;-\; r_A W_i\,\tilde{U}_{ij}^3 \;+\; \left(\frac{r_A Y_i}{2} \;+\; W_i^2\right)\tilde{U}_{ij}^2 \;-\; W_i Y_i\,\tilde{U}_{ij} \;+\; \frac{Y_i^2}{4}\right\}\,. \qquad (3.19)$$

□

Remark . *We note that although the estimator* $\hat{\tau}_1(j, A)$ *is consistent for a very general set of sampling distributions, it may in some cases converge slowly. For very small sample sizes, we find the estimator to be negatively biased; that is, tests involving this estimator may be anticonservative. Although the full DCM procedure appears in simulations to control false positive rate even for small sample sizes, we caution against its use when* $\min(n_1, n_2) < 30$.

CHAPTER 4

Coherent Set Mining for Binary Data

4.1 Introduction

In this chapter, we introduce Coherent Set Mining (CSM), a new method of association mining in binary data. Coherent Set Mining makes use of a VSAT-type algorithm for extracting associated variable sets. Our approach relies a new measure of association, coherence, which is designed to be identified with latent-space relationships between variables when only thresholded binary observations are observed. We propose an estimator for coherence built upon a novel null model and corresponding consistent estimation of parameters. Relevant significance tests for coherence are derived from asymptotic results. We demonstrate the effectiveness of Coherent Set Mining via applications in text mining, music recommendation, and genetics.

4.1.1 The problem of non-identical samples

As discussed in Chapter 1.4.2, many existing association mining methods are applicable (or even tailored specifically) to binary data. These techniques cover a variety of approaches: some treat observed data as stochastic, some seek to maximize an association score, and some simply screen observations for pre-defined features. However, one similarity in common methods is that measures of association or dissimilarity treat observations as homogeneous. Frequent itemset mining deals with raw counts of equally weighted transactions, and in statistical association mining, models typically assume that samples are i.i.d. or approximately so. In reality, the assumption of identically distributed or indistinguishable samples may not be reasonable. For example, in market basket data, it may be unrealistic to assume that all buyers tend to buy the same overall number of items. Variation in available spending money, household size, etc. may effect the quantity of items that a particular buyer is inclined to purchase. To further illustrate the problems that arise from giving

all samples equal treatment, consider the following toy dataset, consisting of 12 samples (buyers) and 14 items.

	Buyers											
	1	2	3	4	5	6	7	8	9	10	11	12
Item 1	0	0	0	0	0	1	0	1	1	1	1	1
Item 2	0	0	0	0	0	0	1	1	1	1	1	1
Item 3	1	1	1	1	1	1	0	0	0	0	0	0
Item 4	1	1	1	1	1	0	1	0	0	0	0	0
Item 5	1	1	1	1	0	1	0	0	0	0	0	0
Item 6	1	1	1	0	1	0	1	0	0	0	0	0
Item 7	1	1	0	1	1	1	0	0	0	0	0	0
Item 8	1	0	1	1	1	0	1	0	0	0	0	0
Item 9	0	1	1	1	1	1	0	0	0	0	0	0
Item 10	1	1	1	1	1	0	1	0	0	0	0	1
Item 11	1	1	1	1	1	1	0	0	0	0	1	0
Item 12	1	1	1	1	1	0	1	0	0	1	0	0
Item 13	1	1	1	1	1	1	0	0	1	0	0	0
Item 14	1	1	1	1	1	0	1	1	0	0	0	0

Figure 4.1: Toy Dataset

Consider the item pairs `(Item 1, Item 2)` and `(Item 3, Item 4)`. These two sets show identical behavior, in that both are purchased by five buyers, neither is purchased by six buyers, and only one is purchased by one buyer. Thus, any measure of association for which the order of buyers does not matter will consider these two sets to be equally internally associated. Common measures of association in literature for binary data, such as frequent itemset mining and association mining, include the following.

- The *support* of an itemset. This is the percentage of buyers who bought the full itemset.

$$\text{support}(1,2) = \text{support}(3,4) = 5/11 = 0.455$$

Generally, methods of frequent itemset mining screen for support over a particular threshold.

- The *confidence* of an itemset with respect to a particular item is the support of the full set divided by the support of the item.

$$\text{confidence}(2 \to 1) = \text{confidence}(4 \to 3) = 5/5 = 1$$

Methods of association rule mining typically screen for confidence over a given threshold.

- The Manhattan distance. This is the L_1 norm between the observed item vectors, i.e., the number of buyers who buy one item but not the other.

$$d_1(1,2) = d_1(3,4) = 1 \text{ (out of 12)}$$

- The *Jaccard coefficient* (Jaccard, 1901). This is a measure of similarity between binary vectors, that divides the intersection of the items by their union.

$$d_J(1,2) = d_J(3,4) = 5/6 = 0.833$$

- The Pearson correlation. This is the inner product of standardized item vectors.

$$\rho_{12} = \rho_{34} = 0.667$$

As discussed in Chapter 1, these measures of association differ in their interpretation: dissimilarity, distance, or statistical dependence. However, for the toy dataset, they *all* indicate that `(Item 1, Item 2)` is a highly associated pair with association equal to that of `(Item 3, Item 4)`. In other words, no matter what algorithmic approach one takes to association mining or to testing for significant association, with these measures the conclusions will be the same for sets `(Item 1, Item 2)` and `(Item 3, Item 4)`. To illustrate this, consider taking an agglomerative hierarchical clustering approach, such as that described in Section 1.2, to perform association mining on the toy dataset. For any measure of association from those above, which treat buyers identically, the dendrogram for the hierarchical joining process will look identical to Figure 4.2 up to the scaling

of the height. It is clear that if `(Item 1, Item 2)` is considered an associated set, `(Item 3, Item 4)` must also be part of an associated set. (In fact, in this data, the sets `(Item 3, Item 5)` and `(Item 4, Item 6)` appear even more strongly associated than `(Item 1, Item 2)`.)

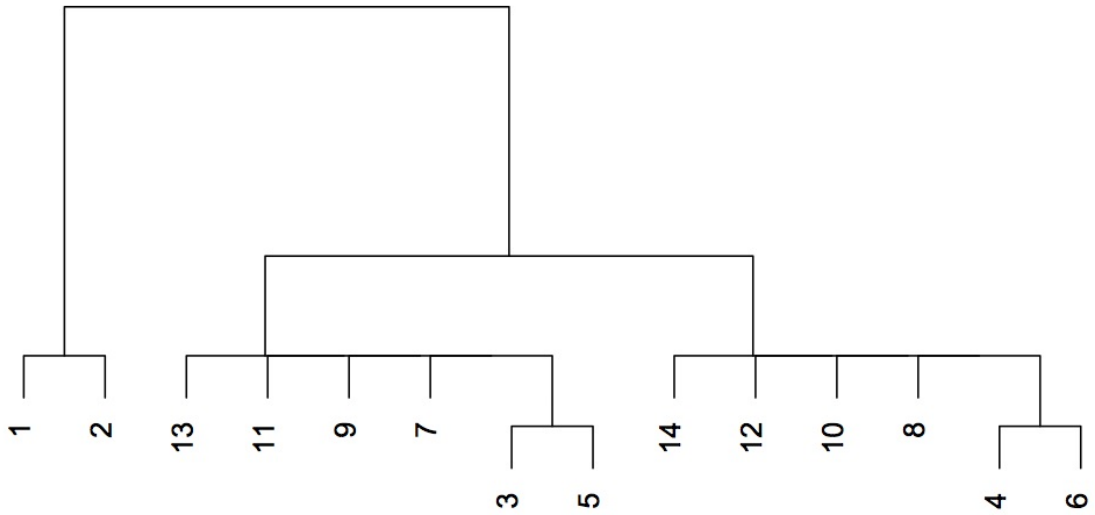
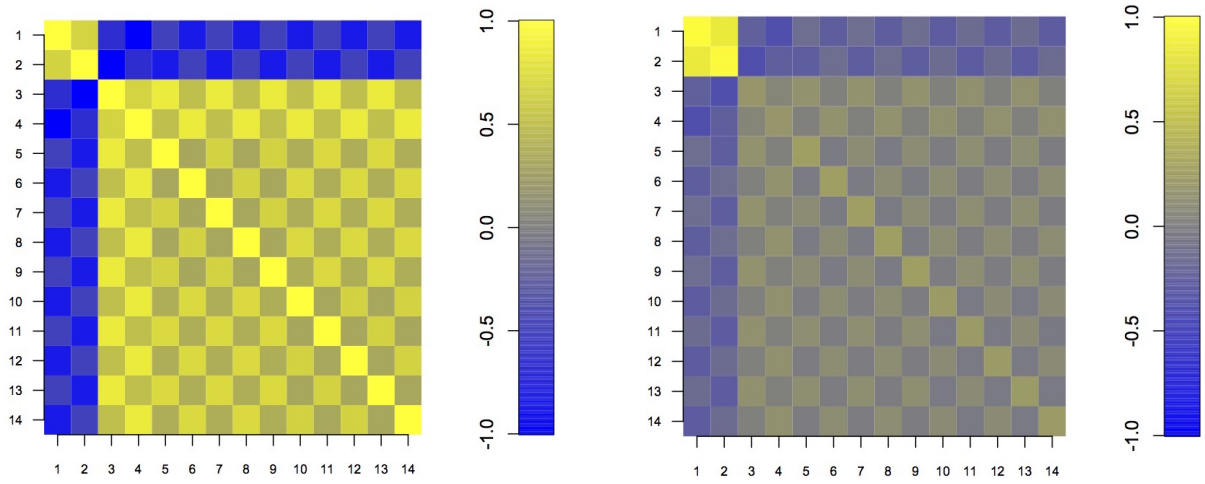


Figure 4.2: Hierarchical clustering dendrogram for toy dataset.

Despite the results of these common approaches, it is not clear that we should believe there is any association structure in items 3-14 aside from an overall pattern in buyer behavior. Buyers 1-5 bought most available items, while buyers 8-12 bought only three items each. Items 3 and 4 may not be meaningfully related beyond the fact that they both respond to differences between buyers. That is, it may not be true that an individual buyer's decision about Item 3 is in any way influenced by his decision about Item 4. Items 1 and 2, on the other hand, show similar buying patterns that can *not* be explained by buyer differences. Buyers who do not purchase many items overall still tend to purchase Items 1 and 2 together, which is a strong indicator of a true relationship between these items. For an association mining method to treat these the sets `(Item 1, Item 2)` and `(Item 3, Item 4)` differently, a new measure of association is required.

The goal of the method described in this chapter is to analyze association between binary variables beyond what be attributed to patterns sample (buyers) behavior. In continuous data, observations can be easily transformed to be roughly identically distributed, e.g., by standardizing the data matrix to achieve identical low-order moments for each sample. In general, such transformations are not appropriate for binary data, as they can only translate 0/1 dichotomous data to a



(a) Sample correlation **(b)** Sample coherence

Figure 4.3: Association matrices based on correlation and coherence for toy dataset.

different pair of numerical values. The remainder of this chapter is devoted to defining and testing a new measure of association, *coherence*, that incorporates the concept non-identical buyers into a formal model.

Figure 4.3 illustrates the difference between coherence and standard linear correlation. 4.3(a) is the sample correlation matrix for the toy dataset, for which we expect `(Item 1, Item 2)` and `(Item 3, Item 4)` to have the same values. 4.3(b) is the estimated coherence matrix, calculated by the methods outlined in this chapter, for the toy dataset. In 4.3(b), `(Item 1, Item 2)` remains associated, but all other associations are devalued since they are not distinguishable from the overall pattern in the data. When the full Coherent Set Mining procedure is applied to this data, only one set, `(Item 1, Item 2)` is identified, which is consistent with the pattern of coherence seen in Figure 4.3(b).

In the rest of this chapter, a formal model and definition is given for coherence, and the Coherent Set Mining procedure is described in detail and applied to both artificial and real data.

4.2 Coherence

Our approach is based on a latent-space model for binary data. In what follows, we will assume that we wish to infer associations between jointly distributed variables $\mathbf{Z} = (Z_1, \ldots, Z_d)^t \in \mathbb{R}^d$, as measured by linear correlation. We further assume that instead of observing $\mathbf{Z}$, we observe

$\mathbf{X} \in \{0,1\}^d$, a binary random vector derived by thresholding $\mathbf{Z}$ in accordance with some random parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. A feature of this model, and the key to the misleading association structure in Figure 4.1, is that association in $\mathbf{X}$ is a result of both association structure of $\mathbf{Z}$ *and* that of $\boldsymbol{\theta}$. To perform association mining on latent vector $\mathbf{Z}$, we require a measure of association calculated from $\mathbf{X}$ that bypasses $\boldsymbol{\theta}$. The following definition formalizes the latent model from which we will define an appropriate measure of association.

Definition 4. (Basic model) *Let $\mathbf{Z} \sim \varphi$ be a real-valued d-dimensional random vector, $\mathbf{Z} = (Z_1, \ldots, Z_d)^t$. For $j = 1, \ldots, d$ let F_j denote the marginal cdf of Z_j, where F_j is taken to be continuous with quantile function F_j^{-1}. Let $\boldsymbol{\theta} \sim \nu$ be a d-dimensional random vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^t \in (0,1)^d$, that is independent of $\mathbf{Z}$. Let $\mathbf{X} \in \{0,1\}^d$ be defined by $\mathbf{X} = \mathbb{I}\{\mathbf{Z} > F^{-1}(\boldsymbol{\theta})\}$, that is, $\mathbf{X}$ is defined elementwise by*

$$X_j = \mathbb{I}\{Z_j > F_j^{-1}(\theta_j)\} \tag{4.1}$$

for $j = 1, \ldots, d$.

We assume $\boldsymbol{\theta}$ takes values in $(0,1)^d$ since, trivially, if $\theta_j = 1$ or 0, X_j is nonrandom. Standard results and the continuity of F_j ensure that $F_j(Z_j)$ is uniformly distributed for any joint measure on $\mathbf{Z}$, so that $\mathbb{E}[X_j \,|\, \boldsymbol{\theta}] = \theta_j$ and the marginal distribution of X_j is fully specified by the marginal distribution of θ_j. The joint distribution of $\mathbf{X}$, however, derives from both $\boldsymbol{\theta}$ and $\mathbf{Z}$. Thus, the pair (X_j, X_k) may be associated even when (Z_j, Z_k) are not. Beyond continuity of F_j, Definition 4 imposes no assumptions on the form of φ or ν. The Coherent Set Mining approach to modeling ν, detailed in Section 4.4, assumes that randomness in $\boldsymbol{\theta}$, and therefore association between individual components (θ_j, θ_k), derives from a common univariate random variable τ. However, in principle the definition of coherence and corresponding asymptotic results are valid for any choice of φ and ν that satisfies certain mild conditions.

For an illustrative example, consider the case of market basket data. A possible interpretation of the model is to let $\mathbf{Z}$ represent the desirability of each available item at a grocery store to a random shopper. Although buyers will, at random, have different item preferences, in some settings it is reasonable to assume that the desirability of items for each buyer comes from a common underlying distribution captured by $\mathbf{Z}$. Individual variables Z_j may have a wide range of means and variance, as some items are naturally more universally popular or controversial than others (e.g., eggs versus

SPAM). Variables in $\mathbf{Z}$ may also be highly dependent, for example, a person who strongly desires peanut butter may be far more likely to also strongly desire jelly.

If one were to somehow gather direct i.i.d. samples of $\mathbf{Z}$ from many buyers, one might reasonably estimate item-item correlations. However, data measuring the abstract notion of "desirability" is difficult to obtain. Perhaps a survey questionnaire or carefully designed behavioral experiment could access $\mathbf{Z}$, but these techniques are expensive and require experts to design and execute. Instead, data may easily be collected in the form of purchasing behavior of buyers, which is a natural proxy for desirability. In other words, one can observe a binary vector $\mathbf{X} \in \{0,1\}^d$ for each customer representing whether or not each item was bought or not. Generally speaking, association in $\mathbf{Z}$ will translate to $\mathbf{X}$; that is, if two items are mutually desirable, it is uncommon for one to be purchased without the other. (Nobody buys peanut butter without jelly.) However, purchasing behavior is not a direct consequence of item desirability; a buyer's decisions are also influenced by factors like wealth. These external factors will determine the desirability cutoffs $\boldsymbol{\theta}_j$ above which item j is bought. More money to spend at the store means that the cutoffs will be lower. Even if a wealthy shopper and a non-wealthy shopper have the same attitude towards desirability of items, the wealthy shopper is still likely to buy many more total items. A pair of expensive, highly desirable items may nearly always either both be purchased by wealthy shoppers or neither purchased by shoppers who cannot afford the items. Thus, we may observe that two items are rarely purchased one without the other - even if they have no association in terms of desirability. Simply put, differences between buyers ($\boldsymbol{\theta}$) can produce association structure in transactions $\mathbf{X}$ even when desirability $\mathbf{Z}$ has none.

We now provide a simple example of a setting in which $\mathbf{X}$ has association induced by $\boldsymbol{\theta}$ despite independence in $\mathbf{Z}$.

Example 4.1. Let $\mathbf{Z} \sim N_d(\mathbf{u}, \mathbf{I}_d)$, for some fixed $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix. Let $\theta_1 = \ldots = \theta_d$, with

$$\theta_j = \begin{cases} \epsilon & \text{with probability } 1/2, \\ 1-\epsilon & \text{with probability } 1/2 \end{cases} \tag{4.2}$$

for some $0 < \epsilon < 1/2$, and let $\mathbf{X} = \mathbb{I}\left\{\mathbf{Z} > \Phi^{-1}(\boldsymbol{\theta}) + \mathbf{u})\right\}$, where $\Phi(\cdot)$ is the standard Normal cdf. Then, for any $j \neq k$, Z_j is independent of Z_k, but $\mathrm{cov}\,(X_j, X_k) > 0$, since

$$\mathbb{E}\, X_j X_k \;-\; \mathbb{E}\, X_j\, \mathbb{E}\, X_k \;=\; \left(\frac{\epsilon^2}{2} + \frac{(1-\epsilon)^2}{2}\right) - \left(\frac{1}{2}\right)^2 \;=\; \frac{1}{4} - \epsilon(1-\epsilon)\,. \tag{4.3}$$

$\Diamond$

As ϵ approaches 0, the covariance between X_j and X_k gets arbitrarily close to $1/4$, which is the maximum possible for binary variables. In other words, because individual variables Z_j are simultaneously thresholded at either very large or very small values, we are likely to observe $X_j = X_k = 0$ or $X_j = X_k = 1$ for any pair (j, k). In the market basket example, this corresponds to purchases from non-wealthy and wealthy buyers respectively. Then, the covariance structure of $\mathbf{Z}$ represents association (or lack thereof) unique to the items, without the effect of the buyer. The absence of dependence in $\mathbf{Z}$ does not prevent dependence in $\mathbf{X}$ produced by the dependence in $\boldsymbol{\theta}$. Common measures of association may indicate - correctly - that structure exists in $\mathbf{X}$, even when the vector of interest $\mathbf{Z}$ has none. To isolate the structure in $\mathbf{Z}$ in our analysis, we introduce the concept of *coherence*.

Definition 5. (Coherence) *As in Definition 4, let $\mathbf{Z} \sim \varphi$ and $\boldsymbol{\theta} \sim \nu$ be independent, and let $\mathbf{X}$ be such that $\mathbf{X} = \mathbb{I}\left\{\mathbf{Z} > F^{-1}(\boldsymbol{\theta})\right\}$. Then, the* coherence *between X_j and X_k with respect to θ_j and θ_k is*

$$\psi(j,k) \;=\; \mathbb{E}_{\varphi,\nu}\left[\frac{(X_j - \theta_j)\,(X_k - \theta_k)}{\sqrt{\theta_j(1-\theta_j)\,\theta_k(1-\theta_k)}}\right], \tag{4.4}$$

where the expectation is taken over the joint distribution of $(\mathbf{X}, \boldsymbol{\theta})$ inherited from (ν, φ).

If $\boldsymbol{\theta}$ is non-random, the coherence reduces to standard Pearson correlation between binary variables X_j and X_k with fixed means θ_j and θ_k. A simple conditioning argument shows that, like correlation, the coherence $\psi(j,k)$ is contained in $[-1, 1]$ for any $j, k \in [d]$, with values close to 1 or -1 indicating stronger dependence. However, while correlation directly measures dependence between X_j and X_k, coherence is designed to measure dependence between Z_j and Z_k only.

Under Definition 4, $\psi(j,k)$ depends only on the joint distributions of (θ_j, θ_k) and of (Z_j, Z_k). Coherence is not identifiable without knowledge of the distributions φ, ν on $\mathbf{Z}$ and $\boldsymbol{\theta}$. Model assumptions for the Coherent Set Mining method are discussed in Section 4.4. In general, multiple

measures on $\mathbf{Z}$ may produce the same coherence. For example, for *any* $\mathbf{Z}$ such that Z_j is independent of Z_k, $\psi(j,k) = 0$. This feature of the framework is neither a limitation nor a lack of specificity; rather, it is the key characteristic that separates variable dependence in $\mathbf{Z}$ from that in $\boldsymbol{\theta}$. Nonzero coherence indicates that $\mathbf{X}$ is conditionally dependent given $\boldsymbol{\theta}$, which implies dependence in $\mathbf{Z}$. For this reason, coherence serves as a reasonable proxy for studying latent association in $\mathbf{Z}$ from observations of $\mathbf{X}$.

In general, not all forms of dependence in $\mathbf{Z}$ result in nonzero coherence - as in ordinary product-moment correlation, only linear association is captured. Further, while nonzero coherence guarantees dependence in (Z_j, Z_k), it does not guarantee a specific type of association; specifically, it does not guarantee positive covariance between Z_j and Z_k. However, as the following proposition shows, in a basic Gaussian setting an equivalence does hold.

Proposition 1. Let $\boldsymbol{\theta} \sim \nu$ and $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{u}, \Sigma)$ for fixed $\mathbf{u} \in \mathbb{R}^d$ and $\Sigma_{jj} = \sigma^2$ for all j. Then, for any ν,

$$\psi(j,k) > 0 \quad \text{if and only if} \quad \Sigma_{jk} > 0\,. \tag{4.5}$$

Proposition 1 is proven in Section 4.8.5 . Note, importantly, that the latent Gaussian model for $\mathbf{Z}$ is not an assumption of the Coherent Set Mining method. It merely provides a familiar setting in which coherence is easily interpretable as a surrogate for underlying correlation.

4.3 Testing for Coherent Sets

Our goal in Coherent Set Mining is to discover *coherent sets* of variables, defined as follows.

Definition 6. (Coherent Set) *Let $\psi(\cdot,\cdot)$ be defined as in Definition 6, for a particular model $\mathbf{X} = \mathbb{I}\{\mathbf{Z} > F^{-1}(\boldsymbol{\theta})\}$. A subset $A \subset \{1,\ldots,d\}$ is a* coherent set *if*

(i) $\psi(j,A) > 0$ for $j \in A$, and

(ii) $\psi(j,A) \leq 0$ for $j \notin A$.

where $\psi(j,A)$ is the average coherence between j and A; that is,

$$\psi(j,A) := \frac{1}{|A|} \sum_{k \in A \setminus \{j\}} \psi(j,k)\,. \tag{4.6}$$

Coherent sets are self-contained variable sets such that each element has positive average coherence with the rest of the set, while no element outside the set does. Since average coherence is a population quantity, in practice it must be estimated from observations. We will assume observations are i.i.d. copies $\mathbf{X}_i = \mathbb{I}\left\{\mathbf{Z}_i > F^{-1}(\boldsymbol{\theta}_i)\right\}$ as in Definition 4. These observations are summarized in data matrix $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^t \in \{0,1\}^{n\times d}$, relative to $\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^t \in \mathbb{R}^{n\times d}$ and $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)^t \in \mathbb{R}^{n\times d}$. We lay the foundation for the full Coherent Set Mining method by first discussing properties when both $\mathbb{X}$ and Θ are observed.

Definition 7. (Idealized sample coherence) *Given observed data matrices* $\mathbb{X}$ *and* Θ*, the* idealized sample coherence *between variables* X_j *and* X_k *is*

$$\widehat{\psi}(j,k) \;=\; \frac{1}{n}\sum_{i=1}^{n} U_{ij}\, U_{ik} \quad \textit{where} \quad U_{ij} := \frac{X_{ij} - \theta_{ij}}{\sqrt{\theta_{ij}(1-\theta_{ij})}}\,. \tag{4.7}$$

The formula in (4.7) is a straightforward estimator for coherence if sample matrices $(\mathbb{X}, \Theta)$ are available: the expectation in (4.4) is replaced with an average over sample quantities. We refer to this estimator as a "idealized" quantity, because $\boldsymbol{\theta}$ is taken to be observed. We later discuss estimation of $\boldsymbol{\theta}$, which is not observed in practical settings. For the time being, however, we will proceed as though $(\mathbb{X}, \Theta)$ is available, in order to show useful properties of the idealized sample coherence.

Since idealized sample coherence is an average of i.i.d.
copies of U_jU_k, it is unbiased for $\mathbb{E}\left[U_jU_k\right] = \psi(j,k) \in [-1,1]$. However, for small samples, $\left|\widehat{\psi}(j,k)\right|$ maybe be larger than 1. Proposition 2 ensures that for large enough sample size it will fall in (or arbitrarily close to) the range $[-1,1]$. In large sample settings this allows us interpret the idealized sample coherence as indicating strong positive association when values are close to 1 and strong negative association near -1.

Proposition 2. If $\sup_{j\le d}\theta_j = o_p(1)$ and $\mathbb{E}\left[\theta_j^{-1}\right] = o(n)$ for all $j \in [d]$, then for any $\epsilon > 0$ and any j, k,

$$\mathbb{P}\left(\left|\widehat{\psi}(j,k)\right| > 1+\epsilon\right) \to 0 \tag{4.8}$$

as $n \to \infty$.

Proposition 2 is proven in Section 4.8.2. The assumptions for this result are closely related to conditions (ii) and (iii) in Theorem 3, which establishes a full central limit theorem for the idealized sample coherence.

The results of Proposition 2 allow us to interpret $\widehat{\psi}(j,k)$, but not to perform inference about it. To apply the Coherent Set Mining procedure (fully outlined in Section 4.5), we require a procedure for testing whether an item X_j , $j \in [d]$, is coherent with a set $A \subset [d]$. That is, we must test hypotheses of the form

$$H_0(j) : \psi(j,A) \,=\, 0 \quad \text{vs.} \quad H_1(j) : \psi(j,A) \,>\, 0\,, \tag{4.9}$$

with $\psi(j,A)$ as in (4.6). The obvious corresponding test statistic is the average sample coherence between X_j and $\{X_k\}_{k\in A}$ denoted by $\widehat{\psi}(j,A)$. In practice an exact p-value for $\widehat{\psi}(j,A)$ cannot be computed without knowledge of φ, ν, so we use an asymptotic approximation. Theorem 3 guarantees asymptotic normality of this test statistic, under appropriate conditions. We now lay out the notation and assumptions for this theorem.

For purposes of asymptotic approximation, we assume that both the sample size n and the number of variable d_n are increasing. Formally, for each n let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \overset{i.i.d.}{\sim} \varphi_n$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \overset{i.i.d.}{\sim} \nu_n$, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id_n})^t$ and $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id_n})^t$. Denote the marginal cdfs of Z_{ij} by F_{jn} and define $X_{ij} = \mathbb{I}\{Z_{ij} > F_{jn}(\theta_{ij})\}$ for $i \in [n], j \in [d_n]$.

We then consider the coherence between a particular variable and a sequence of variable sets. Fix j and for each n let $A_n \subset [d_n]\backslash\{j\}$ with $m_n := |A_n|$. Let $\widehat{\psi}_n$ denote the average idealized sample coherence $\widehat{\psi}(j, A_n)$, and let $\sigma_n^2 := \mathrm{var}(\sqrt{n}\psi_n)$. Define

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_n}\sum_{k\in A_n} U_{ij}U_{ik}\right)^2\,, \tag{4.10}$$

which will serve as an estimator for σ_n^2. Let $\bar{\Psi}_n(A_n)$ denote the average of the matrix of pairwise coherences for variables in A_n, i.e.,

$$\bar{\Psi}_n(A_n) \,:=\, \frac{1}{m_n^2}\sum_{j,k\in A_n} \psi_n(j,k)\,. \tag{4.11}$$

Finally, let s_{jn}^2 be the expected conditional variance of X_j under (φ_n, ν_n), that is,

$$s_{jn}^2 \;=\; \mathbb{E}_{\nu_n}\left[\frac{1}{\theta_j(1-\theta_j)}\right]. \tag{4.12}$$

Theorem 3. (Limiting Distribution) *Let* $\mathbf{Z} \sim \varphi_n$, $\boldsymbol{\theta} \sim \nu_n$, *and* $X_j = \mathbb{I}\{Z_j > \theta_j\}$. *Fix* j *and for each* n *let* $A_n \subset [d_n]\backslash\{j\}$ *be an index set with cardinality* $|A_n| = m_n$. *Let* $\bar{\Psi}_n(A_n)$ *be the average of the coherence matrix for* A_n, *as in* (4.11). *Assume that*

(i) For each n, Z_j *is independent of* $\{Z_k\}_{k\in A_n}$ *under* φ_n;

(ii) $\lim_{n\to\infty}\left(\sup_{k\in\{j\}\cup A_n}\theta_k\right) = o_p(1)$; *and*

(iii) $\left(\frac{1}{m_n}\sum_{k\in A_n} s_{jn}^2 s_{kn}^2\right)\bar{\Psi}_n(A_n)^{\text{-}2} \;=\; o(n)\,.$

Then,

$$\sqrt{n}\left(\frac{\widehat{\psi}_n(j,A_n)}{\hat{\sigma}_n(j,A_n)}\right) \xrightarrow{d} \mathcal{N}(0,1) \quad \textit{as} \quad n\to\infty. \tag{4.13}$$

Theorem 3 is proven in Section 4.8.5. The assumption of independence between Z_j and $\{Z_k\}_{k\in A_n}$ implies the null hypothesis in (4.9), $\psi(j, A_n) = 0$, since independence between Z_j and Z_k guarantees that $\psi(j,k) = 0$. Conditions (ii) and (iii) say, roughly, that the marginal probabilities θ_{ik} get small asymptotically, but not too quickly. This can be interpreted as a sparsity constraint: as the number of samples (n) and variables (d_n) grows, the expected number of 1's in the data matrix $\mathbb{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^t$ must not become too high or too low, or it is impossible to infer association.

Although a general version of Condition (iii) is provided, there are readily interpretable settings under which this condition holds. For example, a common assumption in showing a central limit theorem is that variance of a particular variable grows slower than the sample size. Analogously, suppose we assume that the relevant conditional variances of observations X_j are bounded in the sample size,

$$\left(\frac{1}{m_n}\sum_{k\in A_n} s_{jn}^2 s_{kn}^2\right) \;=\; o(n)\,. \tag{4.14}$$

Then, for (ii) to hold, we need $\bar{\Psi}_n(A_n)^{\text{-}1} = \mathrm{O}(1)$. There are two simple settings for which this is true.

1. *Coherence of A_n bounded away from zero.* In practice, we are interested in sets A_n with large average pairwise coherence. This implies a strong condition on the asymptotic strength of the coherence of A_n may be appropriate. If in addition to (4.14), one assumes $\bar{\Psi}_n(A_n) > \rho$ for all n and for some fixed $\rho > 0$, then (iii) holds regardless of the set sizes m_n.

2. *Non-negative coherence of A_n and bounded set size.* Recall that $\bar{\Psi}_n(A_n)$ is an average over an array of pairwise coherences $\psi(k,\ell) : k,\ell \in A_n$. We may therefore rewrite it in terms of diagonal and off-diagonal terms,

$$\bar{\Psi}_n(A_n) = \frac{1}{m_n^2} \sum_{k \in A_n} \psi_n(k,k) + \frac{1}{m_n^2} \sum_{k \neq \ell \in A_n} \psi_n(k,\ell) \tag{4.15}$$

$$= \frac{1}{m_n} + \left(\frac{1 - m_n^{-1}}{2}\right) \binom{m_n}{2}^{-1} \sum_{k \neq \ell \in A_n} \psi_n(k,\ell)\,, \tag{4.16}$$

   where the last line follows from the fact that $\psi_n(k,k) = 1$ for any n by definition. If one assumes that the average of off-diagonal elements of $\Psi_n(A_n)$ is non-negative, then $\bar{\Psi}_n(A_n) \geq m_n^{-1}$. Then, if $m_n < M$ for some fixed M, $\bar{\Psi}_n(A_n)^{-1}$ is bounded even if the off-diagonal coherences are shrinking in n.

These are, of course, not the only two possible settings for asymptotic normality of the idealized sample coherence. In general, Theorem 3 holds in *any* settings for which Condition (iii) is fulfilled. In particular, (4.14) may be weakened and corresponding assumptions may be made about the off-diagonal behavior of $\Psi_n(A_n)$ and/or the rate of m_n. Note, importantly, that the conditions of Theorem 3 only explicitly involve the sample size n and the size m_n of the sets of interest A_n, not the total size of the dataset d_n. However, Condition (ii), which requires that θ_j is shrinking, is best understood as a sparsity condition for over a growing number of variables d_n. (The results in Section 4.4 do require assumptions on the relative rates of d_n and n.)

In brief, Theorem 3 guarantees the approximate normality of the test statistic for average coherence, under an appropriate null hypothesis, even for growing sets A_n. Our approach to testing the hypotheses in (4.9) is therefore to calculate p-values from the Normal approximation,

$$\mathrm{pv}(j, A) = 1 - \Phi^{-1}\left(\frac{\widehat{\psi}(j,A)}{\hat{\sigma}(j,A)}\right), \tag{4.17}$$

where $\hat{\sigma}(j, A)$ is calculated as in (4.10).

4.4 Model assumptions and parameter estimation

The results of Section 4.3 suggests an hypothesis testing procedure based on the idealized sample coherence, which requires that $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)^t$ is observed alongside $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^t$. In practice, samples of $\boldsymbol{\theta}$ are not observed. Our approach is to derive consistent estimators for Θ from observations $\mathbb{X}$. We then treat these estimates as observed values and insert them directly into the idealized sample coherence equation (4.7). Although this plug-in approach does not account for the variance in the estimation of Θ, our positive results from simulation and applications, as well as the theoretical consistency of our estimators, leads us to believe this is a reasonable approximation.

In order to estimate Θ from $\mathbb{X}$, we assume that for each j,

$$\theta_j \;=\; 1 - \exp(-\tau\alpha_j)\,, \tag{4.18}$$

where α_j is a single fixed parameter and $\tau \sim \pi$ is a univariate random variable. In other words, we assume that the dependence structure and randomness of $\boldsymbol{\theta}$ derives from a single shared random parameter τ. Differences between the marginal distributions of $\theta_1, \ldots, \theta_d$ are then entirely captured by the fixed parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. In the buyer-item paradigm, we may interpret τ as the wealth of a particular buyer, who is randomly selected from the population, and α_j as a measure of the overall prevalence of a particular item. We are thus assuming that the probability of buyer i purchasing an item j is fully determined by wealth (τ_i) and an inherent quality of the item (α_j). A similar model known as *Poisson factorization* is used by Gopalan et al. (2014) and in subsequent work to model expected counts by an exponentiated product of random sample and variable parameters.

This model imposes a rank-one structure on the marginal sample expectations θ_{ij}. The number of quantities to estimate is reduced from $(n \times d)$ parameters $\{\theta_{ij}\}$ to $(n + d)$ parameters $(\alpha_1, \ldots, \alpha_d, \tau_1, \ldots, \tau_n)$. This reduction and model specification allows us to estimate Θ from $\mathbb{X}$ under certain mild conditions.

4.4.1 Parameter estimation

As in Section 4.3, we assume that both the sample size n and the number of variables d_n are increasing, and that $\mathbf{Z} \sim \varphi_n$, $\boldsymbol{\theta} \sim \nu_n$ and $X_j = \mathbb{I}\{Z_j > F_{jn}(\theta_j)\}$. Under the assumption of (4.18), ν_n is fully specified by a univariate measure π with $\tau \sim \pi$ and a set of parameters $\boldsymbol{\alpha}_n = (\alpha_{1n}, \ldots, \alpha_{d_n n})$. We first derive a method of moments estimator for $\boldsymbol{\alpha}_n$ that is consistent under certain conditions by integrating out τ. Let $\boldsymbol{\mu}_n = (\mu_{n1}, \ldots, \mu_{nd_n})^t$ denote the unconditional mean of $\mathbf{X}$, that is, $\mu_{jn} := \mathbb{E}_n[X_j]$. We can then write the μ_{jn} as a function of α_{jn},

$$\mu_{jn} \;=\; g(\alpha_{jn}) \;:=\; \int_{\mathcal{T}} \left(1 - e^{-t\,\alpha_{jn}}\right) \pi(t) \mathrm{d}t\,, \tag{4.19}$$

where $\mathcal{T}$ is the support of π. Note that $g(\cdot) = 1 - M_\tau(\cdot)$, where $M_\tau(\cdot)$ is the moment generating function of τ. Then, if π is such that $M_\tau(\cdot)$ is continuous and invertible, $g(\cdot)$ will also be a continuous invertible function. In these cases, there exists a straightforward estimator for α_{jn} via μ_{jn},

$$\widehat{\alpha}_{jn} \;=\; g^{-1}(\hat{\mu}_{jn}) \;=\; g^{-1}(\overline{X}_j)\,, \tag{4.20}$$

where $\overline{X}_j$ is the sample mean $\sum_{i=1}^n X_{ij}$. Theorem 4 guarantees the consistency of this estimator.

Theorem 4. *Let $\mu_{jn}, g(\cdot)$ and $\widehat{\alpha}_{jn}$ be as in (4.19) and (4.20). If $\mu_{jn} = o(1)$, $\mu_{jn}^{-1} = o(n)$, and $g(\cdot)$ is an invertible function with continuous inverse, then*

$$\left| \frac{\widehat{\alpha}_{jn}}{\alpha_{jn}} - 1 \right| \xrightarrow{p} 0. \tag{4.21}$$

for every $j \in [d_n]$.

Theorem 4, which is proven in Section 4.8.5, provides a procedure for estimating $\boldsymbol{\alpha}_n$ under typical conditions. To estimate Θ, we must also estimate the unobserved values of random variables $(\tau_1, \ldots, \tau_n)$, which we denote by $(\tau_1^0, \ldots, \tau_n^0)$. Consider the posterior distribution of τ_i given $\mathbf{X}_i = (X_{i1}, \ldots, X_{id_n})^t$ and $\boldsymbol{\alpha}_n$, which we denote by $\pi(\cdot \,|\, \mathbf{X}_i, \boldsymbol{\alpha}_n)$. A straightforward estimator for τ_i^0 is the posterior mean,

$$\mathbb{E}\left[\tau_i \,|\, \mathbf{X}_i, \boldsymbol{\alpha}_n\right] \;=\; \int_0^\infty t\; \pi(t \,|\, \mathbf{X}_i, \boldsymbol{\alpha}_n)\; dt\,. \tag{4.22}$$

The following result guarantees consistency of the posterior mean. We appeal directly to Theorem 4.1 of Choi et al. (2008), which requires the following condition on the prior π for τ.

Condition 5.1. *For each $\delta > 0$ there exist sets $S_1, S_2, \ldots$ such that diameter of each set is less than δ, $\cup_{k\geq 1} S_k = \mathbb{R}^+$, and $\sum_{k\geq 1} \sqrt{\pi(S_k)} < \infty$.*

In essence, Condition 5.1 is a concentration condition for π, guaranteeing that the measure is not too spread out over the range of τ.

Theorem 5. (Choi et al. (2008)) *Suppose that Condition 5.1 holds and that $\pi(\cdot \,|\, \mathbf{X}_i, \boldsymbol{\alpha}_n)$ is bounded. Then, for every $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\mathbb{E}\left[\tau_i \,|\, \mathbf{X}_i, \boldsymbol{\alpha}\right] - \tau_i^0\right| > \epsilon\right) \to 0 \tag{4.23}$$

as $n \to \infty$, where the probability is taken over the measure of $\mathbf{X}_i$ from (φ_n, ν_n).

In practice $\boldsymbol{\alpha}_n$ is not known, so we instead estimate τ_i^0 by plugging in consistent estimates $(\widehat{\alpha}_{1n}, \ldots, \widehat{\alpha}_{d_n n})$, i.e.,

$$\widehat{\tau}_i \;=\; \mathbb{E}\left[\tau_i \,|\, \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}_n\right] \;=\; \int_0^\infty t\, \pi(t \,|\, \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}_n)\, dt\,. \tag{4.24}$$

Then, for every n and for $i \in [n], j \in [d_n]$, θ_{ij} is estimated by $\widehat{\theta}_{ij} = \widehat{\tau}_i \widehat{\alpha}_{jn}$. The following example demonstrates a derivation of $\widehat{\theta}_{ij}$ by the process suggested in Theorems 4 and 5.

Example 4.2. *Let τ be exponentially distributed with mean $1/\lambda$. Then, the mgf of τ is $M_\tau(s) = \frac{\lambda}{(\lambda - s)}$, which is continuous and invertible, so the condition of Theorem 4 is satisfied. Note that $\mathbb{E}\, X_j = \alpha_j(\lambda + \alpha_j)^{-1}$. Therefore, the estimator*

$$\widehat{\alpha}_j \;=\; g^{-1}(\overline{X}_j) \;=\; \frac{\lambda \overline{X}_j}{1 - \overline{X}_j} \tag{4.25}$$

is therefore consistent as long as $\alpha_j^{-1} = o(n)$ by Theorem 4. Furthermore, the exponential prior satisfies Condition 4.1 (details may be found in Section 4.8.5), so the posterior means for τ_i^0 are consistent by Theorem 5. $\Diamond$

4.4.2 Implementation

The consistent estimators derived in Theorems 4 and 5 rely on the parent distribution π, and thus the function $g(\cdot)$, being known. In practice, we generally do not know the distribution of τ. Our approach is therefore to approximate the consistent estimators via an empirical distribution function. According to Theorems 4 and 5, estimators $\widehat{\tau}_i$ and $\widehat{\alpha}_j$ will solve

$$\overline{X}_j \;=\; \int_0^\infty \left(1 - e^{-t\alpha_{jn}}\right)\, \pi(t)\, \mathrm{d}t \quad \text{and} \quad \left.\frac{\partial \pi(t \,|\, \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}_n)}{\partial t}\right|_{\widehat{\tau}_i} \;=\; 0\,. \tag{4.26}$$

We circumvent the problem of the prior being unknown by replacing π in the above with the empirical distribution function

$$f_n(t) \;=\; \begin{cases} \frac{1}{n} & \text{if} \quad t \in \{\tau_1^0, \ldots, \tau_n^0\}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.27}$$

where τ_i^0 is the unknown realized value of τ_i. We then define estimators $\tilde{\tau}_i$ and $\tilde{\alpha}_j$ to be the solutions to

$$\bar{X}_j \;=\; \frac{1}{n}\sum_{i=1}^{n} \left(1 - e^{-\tilde{\tau}_i, \tilde{\alpha}_j}\right) \quad \text{and} \quad \left.\frac{\partial \mathcal{L}(t \,|\, \mathbf{X}_t, \tilde{\alpha}_j)}{\partial t}\right|_{\tilde{\tau}_i} \;=\; 0\,. \tag{4.28}$$

Essentially, $\tilde{\alpha}_j$ is the empirical MOM estimator for the fixed parameter α_j, and $\tilde{\tau}_i$ is the frequentist MLE estimator for τ_j^0. Although these equations have no closed form solution, they can be computed to within a desired error. The results of Theorems 4 and 5 provide reassurance, since $\tilde{\alpha}_j, \tilde{\tau}_i$ are empirical analogs to consistent estimators $\widehat{\alpha}_j, \widehat{\tau}_i$.

In the Coherent Set Mining software, we also supply an option to compute $\widehat{\tau}_1, \ldots, \widehat{\tau}_n$ and $\widehat{\boldsymbol{\alpha}}_n$ under an assumed exponential prior, as in Example 4.1. This option should only be used when there is a compelling reason to believe the prior π is known to be exponential with a certain rate λ. Despite the theoretical advantages of the estimators with known prior, we find the flexible empirical approach is more flexible for most settings since π is commonly unknown.

4.5 The Coherent Set Mining Algorithm

We are now prepared to present the full version of the Coherent Set Mining algorithm, which appeals to the results of Sections 4.2 and 4.4. Given observed data $\mathbb{X} \in \{0,1\}^{n \times d}$, the method proceeds a follows.

1. Estimation: Compute $\tilde{\Theta}$, the matrix of estimates of means θ_{ij}, as in Section 4.4.2.
2. Initialization: Set $A_0 = \{j\}$ for some $j \in [d]$.
3. Testing:
   - ▷ Given A_t, for each $j \in [d]$, compute $\widehat{\psi}(j, A_t)$ and $\hat{\sigma}(j, A_t)$ from $\tilde{\mathbb{X}}$ and $\tilde{\Theta}$ as in Section 4.2.
   - ▷ Compute p-values $\{\mathrm{p}_1, \ldots, \mathrm{p}_d\}$ as in (4.17).
   - ▷ Simultaneously test hypotheses

$$H_0(j) : \psi(j, A_\ell) = 0 \quad \text{vs} \quad H_1(j) : \psi(j, A_\ell) > 0$$

     by applying the multiple testing procedure of Benjamini and Yekutieli (2001) to the set of p-values.
4. Update: Set $A_{t+1} = \{\, j \,:\, H_0(j) \text{ was rejected } \}$.
5. Iteration: Repeat steps 3 and 4 until $A_t = A_{t'} := A^*$ for some $t' < t$.
6. Output: If A^* is not empty, select it as an empirical coherent itemset.
7. Repetition: Repeat steps 2-5 as many times as desired, or for every initial $j \in [d]$.

4.5.1 Simulation Study

We first demonstrate the effectiveness of the Coherent Set Mining algorithm via artificial data. Proposition 1 provides us with a convenient generative model for binary data with controlled

strength of association. Given parameters n, d, ρ, k, we simulated data matrix $\mathbb{Z} \in \mathbb{R}^{n \times d}$ by drawing n multivariate Gaussian samples of dimension d, with covariance matrix $I + \Omega$, where $\Omega_{jk} = \rho$ for $j, k \leq m, j \neq k$, and 0 otherwise. We then generate parameters τ and α from hyperparameters λ, a, b by $\tau_i \sim \text{Expo}(\lambda)$, $\alpha_i \sim \text{Beta}(a, b)$. Finally we created Θ as in (4.18), and thresholded $\mathbb{Z}$ as in (4.1) to create binary matrix $\mathbb{X}$.

By Proposition 1, the population coherence of elements of $\mathbb{X}$ is directly related ρ. Thus, by varying our values of ρ, we were able to study the effect of strength of signal on performance of the Coherent Set Mining algorithm. We also studied changes in dimensions $\{n, d, m\}$ and hyperparameters $\{\lambda, a, b\}$. In general, changes to (a, b) did not meaningfully affect the results, since they only alter the values of α, which are considered fixed quantities. Increases in the number of observations (n, d) or the size of the signal block (m) improved algorithm performance, as we would expect. We do not include these results here, since they do not speak to differences in performance between methods. For our study of ρ, remaining parameters were set to $\{n = 101, d = 1000, m = 100, \lambda = 1, a = 1, b = 1\}$, and for our study of λ the same with $\rho = 0.4$.

Remark. In this study, we do not include methods of Frequent Itemset Mining. These procedures are designed for very low dimensional datasets ($d \sim 100$). Since CSM is intended primarily to apply to high dimensional data, our simulation study consists of datasets too large for Frequent Itemset Mining to be computationally feasible.

The success of the compared methods was measured by the false positive rate (FPR), the percentage of variables in a selected set that were not in the seeded coherent set, and the true discovery rate (TDR), the percentage of detected variables from the true coherent set. That is, if B was the output variable set of a procedure and $A = (1, \ldots, m)$ was the embedded correlated set, then

$$\text{FPR} = \frac{|B \setminus A|}{|B|} \quad \text{and} \quad \text{TDR} = \frac{|A \setminus B|}{|A|}.$$

In addition to Coherent Set Mining, we applied four competing methods to generated data. Measures of (dis)association for these methods were:

1. **L1 Dist**: The L1 or "Manhattan" distance between sample vectors,

$$d_{l1}(j, k) = \sum_{i=1}^{n} |X_{ij} - X_{ik}| \tag{4.29}$$

2. **L2 Dist:** The L2 or Euclidean distance between sample vectors.

$$d_{l2}(j,k) \;=\; \left(\sum_{i=1}^{n}(X_{ij}-X_{ik})^2\right)^{1/2} \tag{4.30}$$

3. **Binary Dist**: A distance metric based on treating binary data as on/off bits and comparing the individual frequency of two variables to their joint frequency,

$$d_{bin}(j,k) \;=\; \frac{\left(\sum_{i=1}^{n} X_{ij}\right)\left(\sum_{i=1}^{n} X_{ik}\right)}{\left(\sum_{i=1}^{n} X_{ij}X_{ik}\right)} \tag{4.31}$$

4. **Correlation Distance:** A transformation of the ordinary product-moment correlation between two sample vectors,

$$d_{corr}(j,k) \;=\; \sqrt{2(1-\widehat{\mathrm{cor}}\,(X_j,X_k))}\,. \tag{4.32}$$

For each of the four distance metrics, we applied hierarchical clustering. We selected a cutoff for the dendrogram based on our knowledge of the true embedded set, such that the selected cluster was as close to the correct size as possible.

We also included a an ordinary correlation mining (**CM**, a VSAT procedure adapted from the methods of Chapter 3) to the true underlying data matrix $\mathbb{Z}$, as a benchmark. Of course, we expect this method to naturally perform better than Coherent Set Mining itself, since it is applied to the latent data that is ordinarily not accessible. We include it here to better understand the effects thresholding on the sensitivity of association mining.

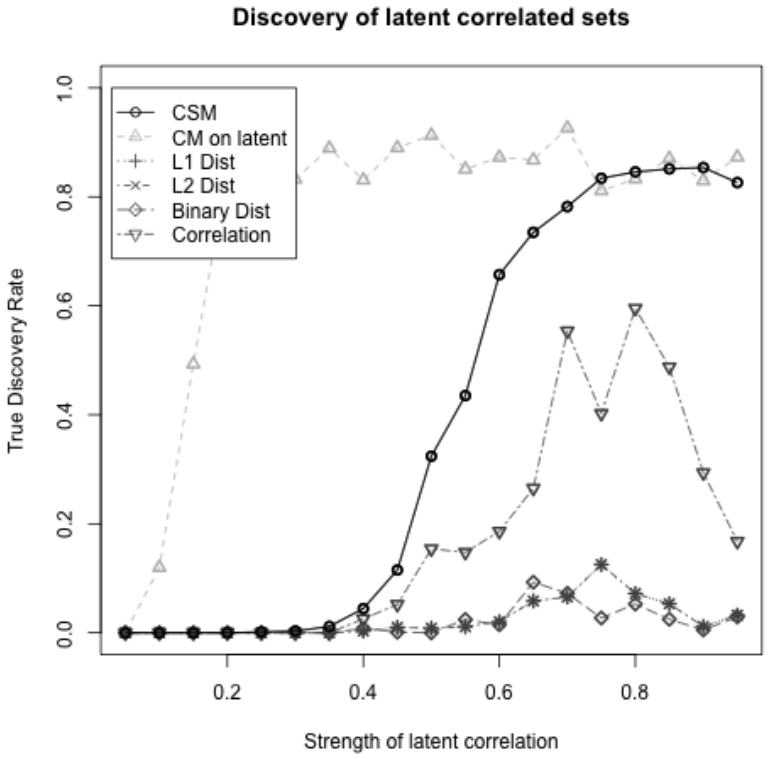
Figure 4.4 shows the True Discovery Rate for all methods as a function of the strength of the true correlation (ρ) in the latent embedded set. Figure 4.4 (a) represents the data setting of interest, where τ is taken to be random (in this case, exponentially distributed with rate 1), while (b) corresponds to the classic setting of non-hierarchical i.i.d. samples. It is clear from the superior performance of the latent **CM** approach that, as one would expect, thresholding continuous data greatly reduces the level at which signal can be detected. However, Coherent Set Mining is able to reliably detect latent correlation at around $\rho = 0.5$ (for the baseline parameter choices of k, n, d). All other methods are unreliable in this setting even for large values of ρ, and only the clustering

based on **correlation** detects signal at all. Figure 4.4 (b) is striking: even when τ is nonrandom, distance-based clustering is unable to detect latent correlation. This is likely because the three distance metrics (**L1, L2** and **Binary**) do not account for differences in mean behavior between *variables*. Not only are samples not appropriately adjusted for randomness in τ, but variables are not treated with proper heterogeneity. Only the correlation distance weights variable behavior appropriately in accordance with mean and variance, and therefore only this clustering was able to detect latent correlation.

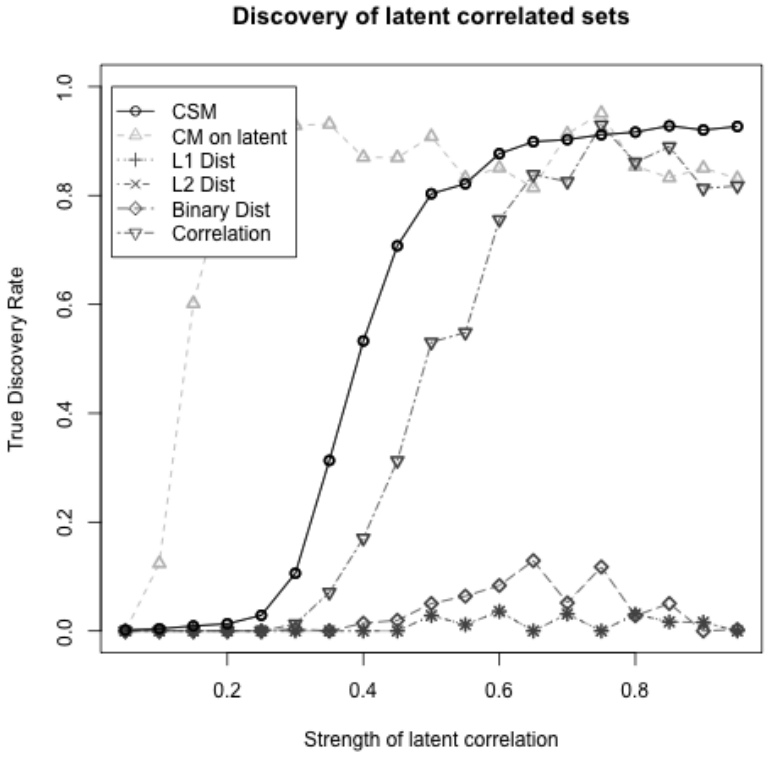
Figure 4.5 also shows True Discovery Rate for all methods, this time as a function of the rate λ for the exponential distribution on τ. (For these simulations, we fix ρ at an intermediate value of 0.6.) Note that $\mathbb{E}[\tau] = 1/\lambda$ and $\mathbb{E}\left[\tau^2\right] = 1/\lambda^2$, so large values of λ correspond to *less* variance in τ. We expect that when τ fluctuates enormously, the induced correlation in $\boldsymbol{\theta}$ will completely drown out the latent correlation of $\mathbf{Z}$, so the detection rate should increase with λ. Indeed, this is the pattern we see for **CSM**. At low values of λ, none of the methods detect structure (except, of course, **CM**, which is not subject to the limitations of the random thresholding). As λ increases, **CSM** and to a lesser extent the **correlation** approach, increase in detection.

Finally, since error control is an important aspect of any VSAT approach, Figure 4.6 shows the raw counts of incorrectly selected variables per set. Figure 4.7 displays the false discovery rate for all tested methods as a percentage of the total identified set, which can be misleading for small set sizes but is nevertheless of interest. In general, **CSM** and the latent clustering **CM** control error and do not identify many false variables. When τ is nonrandom (Figure 4.7(a)) and the latent signal is weak, **CSM** is prone to over-fitting and thus does not control error as a percentage. However, the sets of incorrectly discovered variables are extremely small (Figure 4.6(b)), so this is not cause for too much concern. In cases where it is of scientific importance to control FDR as a percentage of output sets, one could consider disregarding all results below a certain size.

All other methods are susceptible to correlation induced by the randomness in τ, and so they do not control false discovery error in any sense.



(a) τ Exponentially distributed with rate 1



(b) τ nonrandom

Figure 4.4: True discovery rate (when false positive rate < 0.05) by signal latent correlation strength.

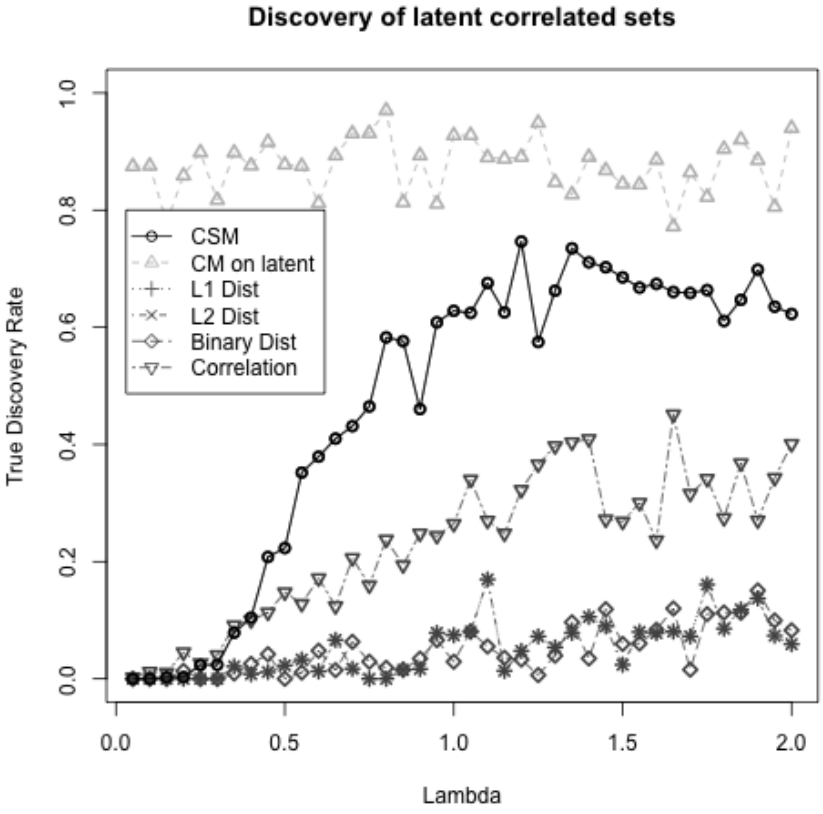
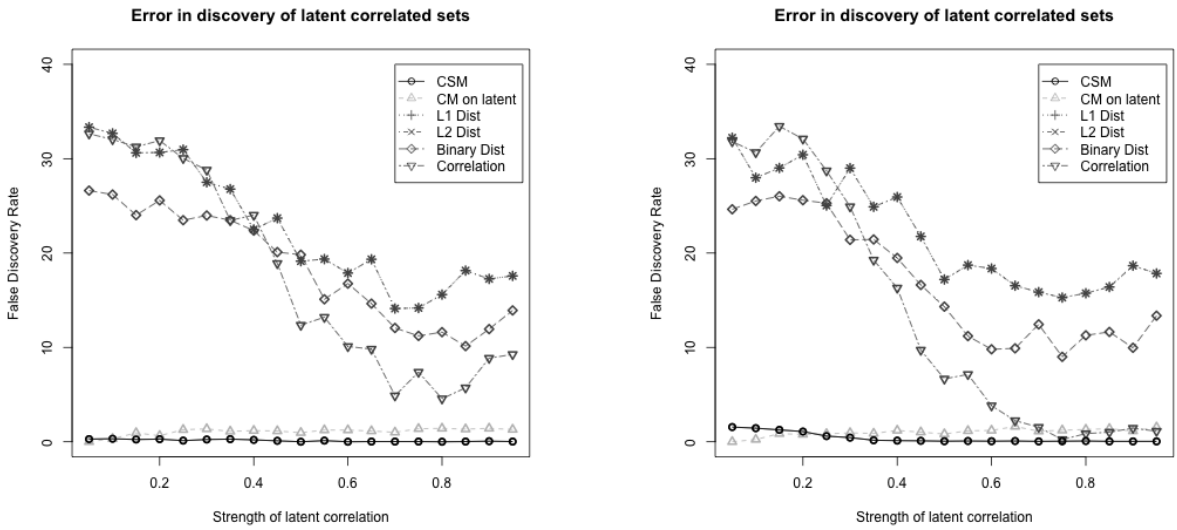
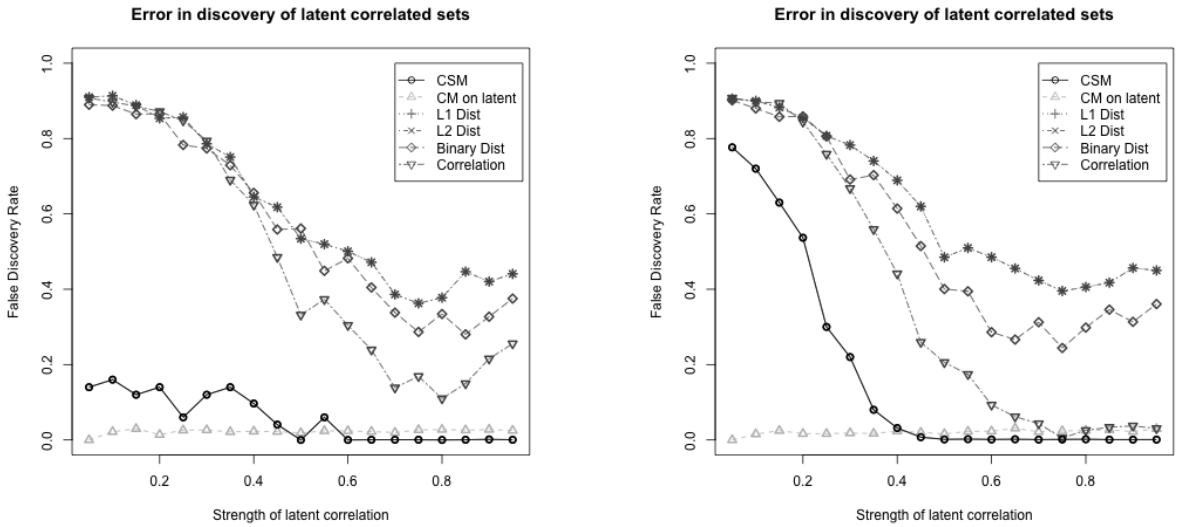


Figure 4.5: True discovery rate (when false positive rate < 0.05) at $\rho = 0.6$ by rate of exponential distribution on τ.



(a) τ Exponentially distributed with rate 1 **(b)** τ nonrandom

Figure 4.6: Number of incorrect variables selected, by signal latent correlation strength.



(a) τ Exponentially distributed with rate 1 **(b)** τ nonrandom

Figure 4.7: False discovery rate by signal latent correlation strength.

4.6 Application: Wordsets in Shakespeare plays

The Coherent Set Mining algorithm is applicable to any binary dataset, and is particularly well-suited to data where the samples may not be identically distributed. Word usage in documents

presents an ideal data source for this paradigm. Text analysts are often interested in finding sets of words that appear together frequently (for an overview of relevant history, see e.g. Salton and McGill (1986)). However, we usually expect documents to vary enormously in length; thus, even if word choice is identical across documents, we expect to observe a non-identical distribution of word presence. By searching for coherent rather than frequent word sets, we are able to extract word groups that are truly associated in a meaningful way, rather than simply appearing frequently together in longer documents.

We used the online database `http://shakespeare.mit.edu/` to download the text of all known Shakespeare plays. We then created a binary dataset for the 1638 unique words that appeared in more than one play and that were used in at least one, but not all, of the 429 acts of Shakespeare's twenty tragedies/histories . That is, a "1" in the data matrix indicated that a particular word appeared at least once in a particular act of a play.

In addition to the Coherent Mining Method, for comparative purposes we also applied a *Text Frequency - Inverse Document Frequency (TF-IDF)* clustering procedure to the Shakespeare data. TF-IDF (Ramos et al., 2003) is a method of standardization for textual data that adjusts observed word frequencies by their importance to differences between documents. Most commonly, TF-IDF adjusted data takes the form of a raw count of a word multiplied by the log ratio of total documents to documents containing that word. That is, let $\mathbb{X} \in \mathbb{N}^{n \times d}$ be a matrix of word counts for d words in n documents. Then, for word j and document i,

$$X_{\text{tf-idf}}(i,j) \;=\; X_{ij} \log\left(\frac{n}{\sum_{i=1}^{n} \mathbb{I}\left\{X_{ij} > 0\right\}}\right).$$

Although TF-IDF can technically be applied to binary observations, it is intended for count data. In this analysis we applied TF-IDF to the word count data for the Shakespeare texts, even though Coherent Set Mining only had access to the binary matrix. Clusters were selected by performing hierarchical clustering on the TF-IDF data matrix by ordinary euclidean distance. The dendrogram was cut at a height that yielded a similar number of clusters as the Coherent Set Mining results, for comparison. (Clusters with more than 50 words were considered "background" and disregarded.)

The Coherent Set Mining software identified 56 coherent word sets from this data, displayed in their entirety in Appendix C. The TF-IDF approach identified 38 associated words sets. On the whole, in both cases these word sets have obvious semantic and/or linguistic themes. For the sake of discussion, Table 4.1 displays five selected coherent word sets, and Table 4.2 displays seven words sets from the TF-IDF clustering that roughly correspond to those in Table 4.1.

Table 4.1: Selected coherent word sets in Shakespearean tragedies

1. earth, heaven
2. thousand, ten, twenty
3. she, her, lady, madam, husband, wife, queen, woman, daughter, shes, marriage, me, tell, sister, herself, sweet
4. hast, dost, art, thy, wilt, thee, thine, thou, death, shalt, canst, didst, ill, sweet, ah, hadst, if, thyself, away, father, eyes, boy, villain, child, mine, mother, kill, wert, me, then, die, o, flesh, am, cheeks, leave, young, sight
5. king, duke, majesty, lords, france, prince, grace, god, princely, unto, liege, sovereign, crown, english, french, highness, uncle, princes, arms, lord, gracious, subjects, cousin, soul, title, now, blood, fathers, then, until, queen, father, traitor, yield, son, right, royal, john, forward, brother, doth, presence, heir, war, sons, embrace, hath, guilty

Table 4.2: Selected word sets in Shakespearean tragedies clustered by TF-IDF adjusted distance

1. arm, arms, base, blood, body, day, doth, earth, eye, farewell, foul, hand, hands, head, heaven, mouth, myself, power, proud, royal, saint, soul, souls, sweet, tale, tongue
2. five, hundred, knight, morrow, today
3. beauty, fair, ladies
4. dead, death, deed, didst, eyes, kill, killd, life, tender, wilt

Set 1 in Table 4.1 is a typical two-word related pair, "earth, heaven". Many such pairs with obvious relationships were selected by both methods. Set 1 in Table 4.2 also joined "earth" and "heaven", but also included many other words in the set. The second set in both analyses captured a numerical relationship, and the third sets are clearly concerned with feminine words. Perhaps most compelling is Set 4 in Table 4.1, which is mostly marked by language rather than meaning - the words are almost entirely from Old English. Set 4 in 4.2 shares some of the same words, is not obviously a linguistically joined word set (nor are any of the further results in Appendix C). Finally, Set 5 in Table 4.1 represents an easily interpretable word set identified by Coherent Set Mining, concerning royalty and titles, that has no equivalent in the TF-IDF results.

The results of Coherent Set Mining on text data are encouraging for several reasons. First, identified word sets have clear interpretation. In a rough sense, this illustrates the notion of "coherence" as a meaningful relationship that is distinct from surface-level association; the word sets in Table 4.1 have clear thematic interpretations. Second, relationships in the resulting word sets may be semantic *or* linguistic. Word sets like "earth, heaven" are validating, but they provide no new information in terms of scientific knowledge. However, the ability to extract word sets like Set 4 that have a deeper linguistic connection may have applications in rigorous studies of language structure. Finally, the comparison between Coherent Set Mining and the popular TF-IDF approach highlights the advantages of CSM. The results of Coherent Set Mining were similar, and perhaps even more nuanced and complete, than those of TF-IDF, even though TF-IDF analyzed full word counts, rather than binary observations. Additionally, the use of the VSAT framework in CSM allowed for overlapping word sets and a selection process that did not require a choice of cut level on a dendrogram. It is worth noting that the hierarchical clustering approach requires the calculation of a full 1638×1638 distance matrix. In larger datasets, such as the one in Section 4.7, this approach would not be computationally convenient.

4.7 Application: Similar Music Artists

Music streaming services such as Pandora, Last.fm, and Spotify offer users the opportunity to discover new musical artists based on existing preferences. These companies have developed complex algorithms for finding similar artists based on era, genre, user ratings, etc. The Coherent

Set Mining framework provides a novel means of artist matching based on coherence. To preserve the directionality of a recommendation approach, instead of seeking coherent sets, we seek *coherent neighborhoods*, consisting of the set of all items that have positive coherence with a chosen target set A. That is, given a set A of preferred artists for an individual, we would like to recommend a neighborhood of similar artists around A. Such neighborhoods are easily estimated by performing only a single iterative step of the Coherent Set Mining algorithm. By considering coherence rather than other similarity measures, we are able to identify related artists (as measured by listener history) without skewing the results towards globally popular music or allowing differences in listener behavior to mask artist associations.

As an example of this approach, we analyzed a dataset provided by Celma (2010) downloaded from the `last.fm` public API. The data consists of listening history for 1893 anonymized users, covering 17,632 unique artists. The data was converted to a binary matrix, where a 1 indicates that a particular user listened at least once to a particular artist. We then applied the single-step Coherent Set Mining algorithm for each individual artist.

Two results of the coherent neighborhood analysis of the `last.fm` data are in Tables 4.3 and 4.4. We also include the top five user-chosen descriptive tags for each artist, to show the type of metadata that might alternatively be used to group artists.[1] Interestingly, although the coherent neighborhoods tend to have clear themes, they do not directly represent the closest artists to the seed based on genre or musical style. For example, the coherent neighborhood in Table 4.3 for "Hannah Montana", a fictional country star from a Disney TV show portrayed by Miley Cyrus, consisted of Cyrus herself and many other singers who got their start on Disney shows (Demi Lovato, Selena Gomez, Ashley Tisdale). Similarly, although many musicians produce similar music to Paul McCartney, the coherent neighborhood in Table 4.4 consists only of the Beatles and fellow Beatles members. This suggests that unsupervised grouping based on coherence may capture links between artists that are not apparent from subjective expert analysis of musical similarities.

[1]Top tags were selected by the percent of times the tag appeared for the artists versus overall in the dataset. Tags were limited to top 100 most popular, to avoid single-user or single-artist tag strings, e.g. "David Bowie" or "Songs for my breakup with Maria."

Table 4.3: Coherent neighborhood for "Hannah Montana"

Artist	Top 5 Tags
Hannah Montana	love at first listen, pop rock, soundtrack, amazing, female vocalist
Miley Cyrus	<3, catchy, love at first listen, amazing, pop rock
Rihanna	rnb, ballad, sexy, love, dance
Katy Perry	pop rock, <3, catchy, love, love at first listen
Britney Spears	catchy, female, sexy, amazing, dance
Ke$ha	love at first listen, dance, <3, pop, catchy
Lady Gaga	dance, female vocalist, love at first listen, catchy, sexy
Demi Lovato	love at first listen, <3, pop rock, catchy, female vocalist
Avril Lavigne	pop rock, canadian, pop punk, female, love at first listen
Taylor Swift	country, <3, catchy, love, amazing
Selena Gomez & the Scene	<3, pop rock, love at first listen, catchy, love
Ashley Tisdale	<3, catchy, pop rock, ballad, awesome
Hilary Duff	favorites, amazing, sexy, pop rock, dance
Christina Aguilera	ballad, sexy, soul, rnb, amazing
Jonas Brothers	pop rock, <3, love, love at first listen, amazing
Beyoncé	rnb, sexy, soul, ballad, female vocalist
Glee Cast	cover, love at first listen, love, catchy, soundtrack

Table 4.4: Coherent neighborhood for "Paul McCartney"

Artist	Top 5 Tags
Paul McCartney	sad, classic rock, cool, british, beautiful
The Beatles	60s, classic rock, british, psychedelic, <3
George Harrison	classic rock, 70s, singer-songwriter, sad, british
John Lennon	classic rock, singer-songwriter, 70s, british, male vocalists

4.8 Proofs and Derivations

4.8.1 Coherence and latent correlation (Proposition 1)

Proposition 1. *Let* $\boldsymbol{\theta} \sim \nu$ *and* $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{u}, \Sigma)$ *for fixed* $\mathbf{u} \in \mathbb{R}^d$ *and* $\Sigma_{jj} = \sigma^2$ *for all* j*. Then, for any* ν*,*

$$\psi(j,k) > 0 \quad \textit{if and only if} \quad \Sigma_{jk} > 0\,. \tag{4.33}$$

Proof. By definition, $\psi(j,k) > 0$ if and only if $\mathbb{E}\left[(X_j - \theta_j)(X_k - \theta_k)\right] > 0$. We proceed by conditioning on $\boldsymbol{\theta}$.

$$\begin{aligned}
\mathbb{E}\left[(X_j - \theta_j)(X_k - \theta_k) \,|\, \boldsymbol{\theta}\right] &= \mathbb{E}\left[X_j X_k \,|\, \boldsymbol{\theta}\right] - \theta_j \theta_k \\
&= \mathbb{P}\left(X_j = 1,\, X_k = 1 \,|\, \boldsymbol{\theta}\right) - \theta_j \theta_k \\
&= \mathbb{P}\left(Z_j < q_j,\, Z_k < q_k \,|\, q_j, q_k\right) - \mathbb{P}\left(Z_j < q_j \,|\, q_j\right) \mathbb{P}\left(Z_k < q_k \,|\, q_k\right) . \quad (4.34)
\end{aligned}$$

When $\Sigma_{jk} = 0$, $\mathbb{P}\left(Z_j < q_j,\, Z_k < q_k\right) = \mathbb{P}\left(Z_j < q_j\right) \mathbb{P}\left(Z_k < q_k\right)$ for any values of q_j, q_k, so (4.34) is 0. Further, by Slepian's Lemma (Slepian, 1962), (4.34) is greater than zero if and only if $\Sigma_{jk} > 0$. Taking the expectation of both sides of (4.34) completes the proof. □

4.8.2 Asymptotic bound on idealized sample coherence (Proposition 2)

Proposition 2. *If* $\sup_{j \le d_n} \theta_j = o_p(1)$ *and* $\mathbb{E}\left[\theta_j^{-1}\right] = o(n)$ *for* $j \in [d_n]$*, then for any* $\epsilon > 0$ *and any* j, k*,*

$$\mathbb{P}\left(\left|\widehat{\psi}(j,k)\right| > 1 + \epsilon\right) \to 0 \quad (4.35)$$

as $n \to \infty$*.*

Proof. We first show that if $\mathbb{E}\left[\theta_j^{-1}\right] = o(n)$ and $\mathbb{E}\left[\theta_k^{-1}\right] = o(n)$, then

$$\mathbb{E}\left[U_j^2 U_k^2\right] = o(n) . \quad (4.36)$$

To prove (4.36), note that it is possible to express U_j^2 in terms of X_j and θ_j as follows,

$$\begin{aligned}
U_j^2 &= X_j \left(\frac{1 - \theta_j}{\theta_j}\right) + (1 - X_j) \left(\frac{\theta_j}{1 - \theta_j}\right) \\
&= X_j \left(\frac{1 - 2\theta_j}{\theta_j}\right) + \left(\frac{\theta_j}{1 - \theta_j}\right) . \quad (4.37)
\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}\left[U_j^2 U_k^2 \,|\, \boldsymbol{\theta}\right] \;&=\; \mathbb{E}\left[X_j X_k \,|\, \boldsymbol{\theta}\right]\left(\frac{(1-2\theta_j)(1-2\theta_k)}{(1-\theta_j)(1-\theta_k)\theta_j\theta_k}\right) + \mathbb{E}\left[X_j \,|\, \boldsymbol{\theta}\right]\left(\frac{(1-\theta_j)\theta_k}{\theta_j(1-\theta_k)}\right) \\ &\quad + \mathbb{E}\left[X_k \,|\, \boldsymbol{\theta}\right]\left(\frac{(1-\theta_k)\theta_j}{\theta_k(1-\theta_j)}\right) + \left(\frac{\theta_j\theta_k}{(1-\theta_j)(1-\theta_k)}\right) \\ &\leq\; \frac{(1-2\theta_j)(1-2\theta_k)}{(1-\theta_j)(1-\theta_k)\sqrt{\theta_j\theta_k}} + \frac{(1-\theta_j)\theta_k}{(1-\theta_k)} + \frac{(1-\theta_k)\theta_j}{(1-\theta_j)} + \frac{\theta_j\theta_k}{(1-\theta_j)(1-\theta_k)}\,,\end{aligned}$$

where the last line follows from the identify for binary random variables $\mathbb{E}\left[X_j X_k\right] \leq \sqrt{\mathbb{E}\left[X_j\right]\mathbb{E}\left[X_k\right]}$. Finally, we note that since $\theta_j, \theta_k = o_p(n)$, for large enough n we have that $\theta_j < (1-\theta_j)$ and similarly for θ_k. Then, $\mathbb{E}\left[U_j^2 U_k^2 \,|\, \boldsymbol{\theta}\right] \leq \sqrt{\theta_j\theta_k}^{-1} \leq \theta_j^{-1} + \theta_k^{-1}$, which proves (4.36).

It then follows that $\mathrm{var}\Big(\widehat{\psi}(j,k)\Big) \to 0$ as $n \to \infty$, and therefore $\Big|\widehat{\psi}(j,k) - \psi(j,k)\Big| \xrightarrow{p} 0$. Since $|\psi(j,k)| < 1$ by construction, this proves the result. □

4.8.3 CLT for idealized sample coherence (Theorem 3)

Theorem 3. *Let* $\mathbf{Z} \sim \varphi_n$, $\boldsymbol{\theta} \sim \nu_n$, *and* $X_j = \mathbb{I}\left\{Z_j > \theta_j\right\}$. *Fix* j *and for each* n *let* $A_n \subset [d_n]\backslash\{j\}$ *be an index set with cardinality* $|A_n| = m_n$. *Let* $\bar{\Psi}_n(A_n)$ *be the average of the coherence matrix for* A_n, *as in* (4.11). *Assume that*

(i) For each n, Z_j *is independent of* $\{Z_k\}_{k\in A_n}$ *under* φ_n;

(ii) $\lim_{n\to\infty}\left(\sup_{k\in\{j\}\cup A_n}\theta_k\right) = o_p(1)$; *and*

(iii) $\left(\frac{1}{m_n}\sum_{k\in A_n} s_{jn}^2 s_{kn}^2\right)\bar{\Psi}_n(A_n)^{-2} \;=\; o(n)\,.$

Then,

$$\sqrt{n}\left(\frac{\widehat{\psi}_n(j,A_n)}{\hat{\sigma}_n(j,A_n)}\right) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad n\to\infty. \tag{4.38}$$

Proof. For convenience, denote

$$U_{ij} \;=\; \frac{X_{ij}-\theta_{ij}}{\sqrt{\theta_{ij}(1-\theta_{ij})}} \quad \text{and} \quad V_{n,i} \;=\; \frac{1}{m_n}\sum_{k\in A_n} U_{ik}\,, \tag{4.39}$$

so that U_{ij} is the standardization of X_{ij} and $V_{n,i}$ is the sample-wise average of $U_{ik} : k \in A_n$. Let $\sigma_n^2 \;=\; \mathrm{var}(\sqrt{n}\psi_n)$. In terms of these, the idealized sample coherence and an estimator for σ_n^2 are given by, respectively,

$$\widehat{\psi}_n(j, A_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} U_{ij}V_{n,i} \quad \text{and} \quad \hat{\sigma}_n^2(j, A_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} U_{ij}^2 V_{n,i}^2\,. \tag{4.40}$$

Since $(\mathbf{X}_i, \boldsymbol{\theta}_i)$ are i.i.d. copies of the pair $(\mathbf{X}, \boldsymbol{\theta})$, we may regard U_{ij} as i.i.d. copies of U_j and $V_{n,i}$ as copies of a random variable V_n. It follows from the definition of $V_{n,i}$ that

$$\mathbb{E}\, V_n^2 \;=\; \frac{1}{n^2}\sum_{k,\ell\in A_n} \mathbb{E}\,[U_k U_\ell] \tag{4.41}$$

which by definition is simply the average coherence of A_n, $\bar{\Psi}_n(A_n)$.

The Lindeberg-Feller condition states that a sufficient condition for (4.38) is that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n^{-1}\,\sigma_n^2}\sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{U_{ij}V_{n,i}}{n}\right)^2 \;\mathbb{I}\left\{\frac{|U_{ij}V_{n,i}|}{n} > \epsilon\, n^{-1/2}\,\sigma_n\right\}\right] = 0\,, \tag{4.42}$$

or equivalently,

$$\lim_{n\to\infty} \frac{1}{n\,\sigma_n^2}\sum_{i=1}^{n} \mathbb{E}\left[\, U_{ij}^2 V_{n,i}^2 \;\mathbb{I}\left\{|U_{ij}V_{n,i}| > \epsilon\, n^{1/2}\,\sigma_n\right\}\right] = 0\,. \tag{4.43}$$

Since $U_{ij}, V_{n,i}$ are i.i.d. copies, the identities in (4.40) give

$$\mathrm{var}\left(\widehat{\psi}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left[U_{ij}^2\, V_{n,i}^2\right] = n^{-1}\mathbb{E}\left[U_j^2\, V_n^2\right] \tag{4.44}$$

Then, (4.43) reduces to

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[U_j^2\, V_n^2 \;\mathbb{I}\left\{|U_j\, V_n| > \epsilon\, n^{1/2}\,\sigma_n\right\}\right]}{\mathbb{E}\, U_j^2 V_n^2} = 0\,. \tag{4.45}$$

By Cauchy-Schwarz, it suffices to show that

$$\lim_{n\to\infty} \left(\frac{\mathbb{E}\, U_j^4 V_n^4}{(\mathbb{E}\, U_j^2 V_n^2)^2} \right)^{1/2} \mathbb{P}\left(n^{-1/2} |U_j V_n| > \epsilon\, \sigma_n \right)^{1/2} = 0. \tag{4.46}$$

Markov's inequality gives

$$\mathbb{P}\left(n^{-1/2} |U_j\, V_n| > \epsilon\, \sigma_n \right) \;\leq\; \frac{\mathbb{E}\left[U_j^2\, V_n^2 \right]}{n\, \epsilon^2\, \sigma_n^2} \;=\; \epsilon^{-2} n^{-1}\,. \tag{4.47}$$

An application of Lemma 1 then completes the proof of (4.43). It remains only to show that $\hat{\sigma}_n^2$ is consistent for σ_n^2. Note that $\hat{\sigma}_n^2$ is unbiased for σ_n^2, since it is an average of i.i.d. copies of $U_{ji}^2 V_{n,i}^2$. The variance of $\hat{\sigma}_n^2$ is given by

$$\mathrm{var}\big(\hat{\sigma}_n^2\big) \;=\; \frac{1}{n} \left(\mathbb{E}\left[U_j^4 V_n^4 \right] - \mathbb{E}\left[U_j^2 V_n^2 \right]^2 \right) \tag{4.48}$$

Recall that U_j^2 and V_n^2 are conditionally independent given $\boldsymbol{\theta}$, and that by the definition of U_j^2, $\mathbb{E}\left[U_j^2 \right] = \mathbb{E}\left[U_j^2 \,|\, \boldsymbol{\theta} \right] = 1$ for all j. Then,

$$\begin{aligned}
\mathbb{E}\left[U_j^2 V_n^2 \right] \;&=\; \mathbb{E}\left[\mathbb{E}\left[U_j^2 V_n^2 \,|\, \boldsymbol{\theta} \right] \right] \\
&=\; \mathbb{E}\left[\mathbb{E}\left[U_j^2 \,|\, \boldsymbol{\theta} \right] \mathbb{E}\left[V_n^2 \,|\, \boldsymbol{\theta} \right] \right] && (4.49) \\
&=\; \mathbb{E}\left[V_n^2 \right] && (4.50) \\
&\leq\; \mathbb{E}\left[\frac{1}{n} \sum_{j \in A} U_j^2 \right] = 1\,, && (4.51)
\end{aligned}$$

where the last line follows from Cauchy-Schwartz. Since $\mathbb{E}\left[U_j^2 V_n^2 \right] \leq 1$, (4.48) holds when (4.52) holds, and so another application of Lemma 1 completes the full proof. □

Lemma 1. *Under the conditions of Theorem 3,*

$$\frac{\mathbb{E}\left[U_j^4 V_n^4 \right]}{(\mathbb{E}\, U_j^2 V_n^2)^2} \;=\; o\,(n)\,. \tag{4.52}$$

Proof. First, we show that for any $k \in [d_n]$,

$$\mathbb{E}\left[U_k^4\right] \,<\, \mathbb{E}\left[\frac{2}{\theta_k(1-\theta_k)}\right]. \tag{4.53}$$

We will show (4.53) using the identity

$$U_k^4 \,=\, \mathbb{I}\{X_k = 1\}\left(\frac{1-\theta_k}{\sqrt{\theta_k(1-\theta_k)}}\right)^4 + \mathbb{I}\{X_k = 0\}\left(\frac{-\theta_k}{\sqrt{\theta_k(1-\theta_k)}}\right)^4. \tag{4.54}$$

Recall that $\mathbb{E}[X_k \,|\, \theta_k] = \theta_k$. Then, using the above identity, a simple rearrangement of terms and conditioning on θ_k gives,

$$\begin{aligned}
\mathbb{E}\left[U_k^4\right] \,&=\, \mathbb{E}\left[\mathbb{E}\left[X_k\left(\frac{(1-\theta_k)^2}{\theta_k^2}\right) + (1-X_k)\left(\frac{\theta_k^2}{(1-\theta_k)^2}\right) \,|\, \boldsymbol{\theta}\right]\right] \\
&=\, \mathbb{E}\left[\left(\frac{(1-\theta_k)^2}{\theta_k}\right) + \left(\frac{\theta_k^2}{(1-\theta_k)}\right)\right] \\
&=\, \mathbb{E}\left[\frac{(1-\theta_k)^3 + \theta_k^3}{\theta_k(1-\theta_k)}\right].
\end{aligned}$$

Finally, $\theta_k \in (0,1)$, $(1-\theta_k)^3 + \theta_k^3 < 2$.

By Holder's inequality, $\mathbb{E}\, V_n^4$ can be expanded and bounded by

$$\mathbb{E}\, V_n^4 \,=\, \frac{1}{m_n^4}\sum_{j,k,\ell,h \in A_n} \mathbb{E}\left[U_j U_k U_\ell U_h\right] \,\leq\, \frac{1}{m_n^4}\sum_{j,k,\ell,h \in A_n} \left(\mathbb{E}\, U_j\,^4 \mathbb{E}\, U_k^4\, \mathbb{E}\, U_\ell^4\, \mathbb{E}\, U_h^4\right)^{1/4} \tag{4.55}$$

From condition (i) of Theorem 4 and the definitions of U_j and V_n, we may conclude that U_j and V_n are conditionally independent given $\boldsymbol{\theta}$. By the argument in (4.49), $\mathbb{E}\left[U_j^2 V_n^2\right] \,=\, \mathbb{E}\, V_n^2 \,\leq\, 1$. Therefore,

$$\begin{aligned}
\frac{\mathbb{E}\left[U_j^4 V_n^4\right]}{\mathbb{E}\left[U_j^2 V_n^2\right]^2} \,&=\, \frac{\mathbb{E}\, U_j^4\; \mathbb{E}\, V_n^4}{\mathbb{E}\left[V_n^2\right]^2} \\
&\leq\, \frac{\mathbb{E}\, U_j^4}{\mathbb{E}\, V_n^2}\left(\frac{1}{m_n}\sum_{k \in A_n} \mathbb{E}\left[U_k^4\right]^{1/4}\right)^4.
\end{aligned} \tag{4.56}$$

A substitution from the bound in (4.53) reduces (4.52) to the condition

$$\mathbb{E}\left[\frac{2}{\theta_j(1-\theta_j)}\right]\left(\frac{1}{m_n}\sum_{k\in A_n}\mathbb{E}\left[\frac{2}{\theta_k(1-\theta_k)}\right]\right)(\mathbb{E}\,V_n^2)^{-2} \;=\; o(n)\,, \tag{4.57}$$

or equivalently, using the definition of s_{jn} from (4.12),

$$4s_{jn}^2\left(\frac{1}{m_n}\sum_{k\in A_n}s_{kn}^2\right)\bar{\Psi}_n(A_n)^{-2} \;=\; o(n)\,, \tag{4.58}$$

which holds by assumption (*iii*) of Theorem 3. □

4.8.4 Parameter estimation (Theorem 4)

Theorem 4. *Let $\mu_{jn}, g(\cdot)$ and $\widehat{\alpha}_{jn}$ be as in* (4.19) *and* (4.20). *If*

(i) $\mu_{jn} = o(1)$,

(ii) $\mu_{jn}^{-1} = o(n)$, and

(iii) $g(\cdot)$ is an invertible function with continuous inverse, then

$$\left|\frac{\widehat{\alpha}_{jn}}{\alpha_{jn}} - 1\right| \xrightarrow{p} 0. \tag{4.59}$$

Proof. For simplicity, we will suppress the dependence on n in the notation of μ_{jn} and α_{jn}. Recall that by definition, $\left|\frac{\hat{\alpha}_{jn}}{\alpha_{jn}} - 1\right| \;=\; \alpha_{jn}^{-1}|g^{-1}(\overline{X}_{\cdot j}) - g^{-1}(\mu_{jn})|$. Because $g^{-1}(\cdot)$ is continuous, the mean value theorem guarantees that for each $j \in [d_n]$,

$$g^{-1}(\overline{X}_{\cdot j}) - g^{-1}(\mu_{jn}) \;=\; (g^{-1})'(\tilde{\mu_{jn}})\left(\overline{X}_{\cdot j} - \mu_{jn}\right) \tag{4.60}$$

where $\tilde{\mu_{jn}}$ is between $\overline{X}_{\cdot j}$ and μ_{jn}, and $(g^{-1})'$ is the derivative of g^{-1}. As $g^{-1}(\cdot)$ is continuous, $(g^{-1})'(\cdot)$ is bounded. We thus reduce (4.59) to the equivalent asymptotic statement,

$$\mu_{jn}^{-1}|\overline{X}_{\cdot j} - \mu_{jn}| \to 0\,. \tag{4.61}$$

Note that $\overline{X}_{\cdot j}$ is a sum of i.i.d. observations of Bernoulli random variables $(X_{1j}, \ldots, X_{nj})$, each of which has finite fourth moment ($\mathbb{E}\, X_{ij}^4 < 1$), and variance $\mu_{jn}(1-\mu_{jn})$. Thus, a basic application of the CLT guarantees

$$\sqrt{n}\frac{|\overline{X}_{\cdot j} - \mu_{jn}|}{\sqrt{\mu_{jn}(1-\mu_{jn})}} = O_p(1) \tag{4.62}$$

Since $(1-\mu_{jn}) \to 1$ by (ii) and $\sqrt{n}\sqrt{\mu_{jn}} \to 0$ by (iii), the proof is complete. □

4.8.5 Example 4.2

Claim. *The prior $\tau \sim$ expo(λ) satisfies Condition 5.1, i.e., for each $\delta > 0$ there exist sets $S_1, S_2, \ldots$ such that diameter of each set is less than δ, $\cup_{k\geq 1} S_k = \mathbb{R}^+$, and $\sum_{k\geq 1}\sqrt{\pi(S_t)} < \infty$.*

Proof. Fix $\delta > 0$ and set $S_k = [(k-1)\delta,\, k\delta]$ for $k = 1, 2, \ldots$. Then,

$$\pi(S_k) \;=\; \int_{(k-1)\delta}^{k\delta} \lambda e^{-\lambda x}\mathrm{d}x \;=\; e^{-\lambda k\delta}\left(1 - e^{-\lambda\delta}\right). \tag{4.63}$$

Since $\lambda\delta > 0$, we conclude that

$$\sum_{k=1}^{\infty}\sqrt{\pi(S_k)} \;\leq\; \sum_{k=1}^{\infty}\exp\left(-\frac{k\lambda\alpha}{2}\right) \;<\; \infty\,, \tag{4.64}$$

so the condition is satisfied. □

CHAPTER 5
Conclusions and Future Work

This dissertation has provided an in-depth discussion of two new methods of statistical association mining, DCM and CSM, both derived from a framework of variable-to-set affinity testing. In addition to having sound foundations in statistical principles, these methods have proven effective in a variety of real data studies. We are particularly encouraged by the ongoing collaborations that have been established with regard to DCM results for statistical genetics (Section 3.7). However, despite our confidence in these methods, some interesting theoretical questions still remain:

- Does the general global error control result for VSAT (Theorem 1) hold in the presence of non-uniform or non-independent p-values? Can a version of this result be shown in a non-null setting?
- Can the central limit theorem for $\hat{\Delta}(j, A)$ used in the DCM method (Corollary 1) be extended to allow the set size $|A| = m$ to grow with the sample size n?
- Do the approximations to consistent estimators given in 4.4 have any similar consistency properties?

In addition to extending existing theory, there are many possible future projects suggested by the work in this dissertation. Below, I detail three major directions for future work.

5.1 Prediction after VSAT

At present, the development of VSAT type methods has focused on unsupervised learning; the objective is to identify ζ-connected sets, not to use ζ to categorize future measurements. In principle, however, I believe the results of VSAT algorithms may lend themselves to prediction and classification tasks. Consider, for example, the DCM data setting. Samples are assumed to come from two known sample conditions, from which we identify differentially correlated variable

sets. Suppose that after applying DCM to a particular dataset, we were to subsequently observe a sample for which the condition is unknown. Could the results of DCM be leveraged to classify the new sample?

A predictive method based on VSAT results could proceed as follows

1. Define ζ such that large values of ζ correspond to differences between a target sample categories and the remaining samples.

2. Apply a VSAT type algorithm to identify ζ-connected sets $A_1, \ldots, A_n$.

3. Given a new sample i, calculate the change in ζ for each of $\{A_1, \ldots, A_n\}$ when the sample is included.

4. Test the change in ζ due to including sample i for significance, to determine if i belongs in the target category.

In essence, this approach makes use of the VSAT framework to determine which variables, out of many, characterize a sample category of interest. We may then use this limited set of variables to test the category membership of a new sample. For example, consider the DCM application to TCGA gene expression data in Section 3.7. This analysis relies on a pre-determined separation between Her-2 and Luminal B type tumor samples. Perhaps the results of this analysis could be used to classify tumor samples with unknown cancer type.

5.2 Correlation mining with continuous response

In the DCM project, we expanded single-dataset correlation mining to a comparative setting. We may also consider expanding this notion further to measure how intracorrelation of a group of variables changes with a response. That is, suppose we have a random variable $\mathbf{X} \in \mathbb{R}^d$ such that the joint distribution of $\mathbf{X}$ depends on a parameter $y \in \mathbb{R}$, i.e. $\mathbf{X} \,|\, y \sim F_y$. The parameter y may be interpreted as a response quantity, such as survival rate for a diseased individual. If one were interested in how the *mean* behavior of $\mathbf{X}$ changes with y, a regression approach would be appropriate. It is also sometimes of interest to understand how the *association* in $\mathbf{X}$ changes with y.

Suppose the correlation between two variables X_j, X_k from $\mathbf{X}$ can be represented as a function of y, i.e.,

$$g_{jk}(y) = \text{cor}\,(X_j, X_k \,|\, y)\ . \tag{5.1}$$

Assume that for each sample i, we observe both $\mathbf{X}_i$ and a corresponding y_i. Then, it may be of interest to identify variable sets in which $g_{jk}(y)$ changes with y. That is, we may consider searching for variable sets A such that

$$\sum_{j,k \in A} g_{jk}(y)$$

is increasing in y. This represents a generalization of the DCM setting, in that if $y \in \{0, 1\}$ the model reduces to looking for correlation differences across sample conditions $y = 1$ and $y = 0$.

A VSAT approach is clearly appropriate to this question. However, the challenge lies in the development and analysis of a test statistic measuring the behavior of $g_{ij}(y)$ when y is nonbinary. The formulation of such a statistic is a complex question; in practice, we would likely observe a response vector $\mathbf{y} = (y_1, ..., y_n)$ and corresponding variable values $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Thus, we only have a single observation of $\mathbf{X}$ for each value of y, and we cannot compute sample correlations $\widehat{\text{cor}}\,(X_j, X_k \,|\, y)$ directly.

A project of this nature would represent a useful contribution to data analysis, especially in the realm of bioinformatics. Researchers are often interested in discovering subsets of genes whose groupwise behavior drives a phenotypic response. This work would open the door to a more nuanced study, as it would allow for continuous phenotypic metrics.

5.3 A VSAT approach to collaborative filtering

Collaborative filtering is a method used by many recommendation algorithms to predict a preferences or rankings of songs, movies, etc. The key idea is to use preference information across many individuals to predict the preferences of a single individual, under the assumption that if two people agree on one topic, they are likely to agree on others. Existing methods tend to take a "nearest neighbors" approach known as collaborative filtering:

1. Calculate the similarity between all users and the target user i. (e.g. correlation of rankings, covariance, etc)

2. Identify the m users with most similarity towards user i

3. Predict user i's rating of item j by some aggregate of the m nearest neighbors. (e.g, average, weighted by similarity average, etc)

Linden et al. (2003) provides and in-depth discussion of the collaborative filtering recommendation approach in the case of amazon.com purchases. Collaborative filtering can also be thought of as a "matrix completion" problem, in the sense that user i's potential rating of item j is a missing data point to be filled in. Candès and Tao (2010) and Candes and Recht (2012) are common methods of matrix completion for recommendation. For surveys of collaborative filtering and matrix completion methods, see e.g. Breese et al. (1998); Sarwar et al. (2001)

As a simple illustration, suppose we observe ratings for 3 people on a scale of 1-10, for 10 movies:

	Movie									
	A	B	C	D	E	F	G	H	I	K
Person 1	1	3	4	9	7	NA	NA	NA	NA	NA
Person 2	NA	2	5	NA	6	9	3	NA	1	1
Person 3	8	NA	6	1	5	NA	5	7	1	4

The goal, then, is be to predict the ratings of Person 1 from 2 and 3. (In practice, we would likely have datasets for tens of thousands of people and possibly hundreds of movies.) A very basic approach would be to note that Person 2 has very similar ratings to Person 1. Then, we would take Person 1 to be the "nearest neighbor" of Person 2, and we would predict that Person 2 has the same ratings as Person 1 for Movies E-K (excluding Movie H, for which Person 2 has no ratings). There are a few notable weaknesses with this type of approach:

1. The identification of the N nearest neighbors can be computationally costly.

2. One must decide how best to aggregate the data from the N nearest neighbors to predict an individual.

3. A decision must be made about which size N to use.

4. Methods must be able to handle missing data (i.e. unrated items).

5. Correlation may be induced by patterns in items. That is, if one movie is naturally more likely to be rated highly than another, we will see users have similar rating patterns even if they aren't actually similar in preference.

The collaborative filtering framework is an ideal setting for a VSAT algorithm. As in Coherent Set Mining, a careful choice of ζ and $\mathbb{P}_0$ can account for overall differences in user behavior and in item popularity, as well as properly handle missing data points. Estimates of ζ-neighborhoods around an individual are then a good choice of "neighbors" from which to predict that individual's unobserved preferences.

An advantage of this approach is that we can use *dissimilarity* data as well as similarities. That is, we can find neighborhoods around a user j such that $\zeta(j, A)$ is positive *and* neighborhoods in which $\zeta(j, A)$ is negative. For example, in the above sample data, we may also note that Person 3 has very different ratings from Person 1, especially about movies A and E. We might expect, then, that Persons 1 and 3 will also have opposite opinions about other movies. If we properly weight the data from Persons 2 and 3, in accordance with their (dis)similarity to Person 1 (and the uncertainty thereof), we bring more information to bear in our estimate of Person 1's ratings.

The VSAT approach to collaborative filtering also comes equipped with a natural method of prediction from a neighborhood. Instead of directly averaging the preferences of the users in the ζ-neighborhood of a user j, we can make use of a weighted average, with weights given by individual associations $\zeta(j, k)$. I believe that the results of such a method would be useful in many recommendation contexts.

"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know."

-Donald Rumsfeld, U.S. Department of Defense briefing (2002)

APPENDIX A: PSEUDOCODE FOR DCM

Algorithm 1 Initial Search Procedure

procedure INITDCM($\mathbf{X}_1, \mathbf{X}_2, k$) ▷ Target output size k.

  $\mathbf{F}_1, \mathbf{F}_2 \leftarrow$ Fisher transformed correlation matrices of $\mathbf{X}_1, \mathbf{X}_2$.

  $B \leftarrow$ index set of size k chosen uniformly at random

  repeat

    $S = \sum_{i,j \in B} (\mathbf{F}_1 - \mathbf{F}_2)_{ij}$

    for a in B, r in B^C **do** ▷ Possible swaps

      $S_{ar} = \sum_{i,j \in B \cup \{r\} \setminus \{a\}} (\mathbf{F}_1 - \mathbf{F}_2)_{ij}$

    end for

    $a^*, r^* \leftarrow$ maximizers of S_{ar} subject to $S_{a^*r^*} > S$.

    $B \leftarrow B \cup \{r^*\} \setminus \{a^*\}$ ▷ Best swap

  until no such a^*, r^* exist

return B

end procedure

Algorithm 2 Core Search Algorithm

procedure DCM($\mathbf{X}_1, \mathbf{X}_2, A$) ▷ Initial index set A
  $A_{prev} \leftarrow \emptyset$
  $cycle \leftarrow 0$
  repeat
    $t = 1$ or 2 ▷ Sample class label
    $m_t \leftarrow \text{mean}\left(\{\mathbf{X}_t\}_A\right)$
    for i in $1, \cdots, p$ **do**
      $r_{ti} \leftarrow \widehat{\text{cor}}\left(X_{ti}, m_t\right)$
      $\hat{T}_i \leftarrow m_1 r_{1i} - m_2 r_{2i}$ ▷ Sample test statistic.
      $H_{0,i} : T_i = 0$
      $p_i \leftarrow P\left(T_i \geq \hat{T}_i \,|\, H_{0,i}\right)$ ▷ DC variable p-values
    end for
    $A_{next} = \{i : H_{0,i}$ rejected by FDR controlled multiple testing$\}$
    if $A_{prev} = A_{next}$ **then** ▷ Check for cycles.
      $A_{prev} \leftarrow A$
      $A \leftarrow A \cap A_{next}$
      $cycle+ = 1$
    else ▷ Update sets.
      $A_{prev} \leftarrow A$
      $A \leftarrow A_{next}$
    end if
  until $A_{next} = A_{prev}$ or $A_{next} = \emptyset$ or $cycle = 2$
return A
end procedure

APPENDIX B: ADDITIONAL TCGA GENE LISTS

Table B.1: Gene lists for TCGA data

1. PDE3A, SPTBN2, FMOD, SLC26A2, FAM84A, RHOBTB2, CTSF, NAT10, S100P, CCS, PARM1, ALG8, KCTD21, ITGA10, CD59, C11orf80, NARS2, CGREF1, USP35, GALNT4, PTPRN2, CAPRIN1, ATP8B5P, FBXO3, EIF3M, PAIP2B, RCE1, AVPI1, TFF3, LOXL4, PPAPDC1B

2. SELL, IL2RB, RAMP3, CLIC5, PLA1A, LEF1, TMEM176A, PTGER2, CST7, SASH3, CD2, CD4, MYO1F, RASGRP1, CXCR3, FMNL1, RSPO3, FERMT3, LAPTM5, CD3D, CLIC2, RASAL3, ARHGAP9, ACAP1, TRAF3IP3, GZMA, FAM20A, PTPN7, GPRIN3, SERPINF2, TMEM176B, CD37, CSF1, CARD11, CD5, LRRC8C, GIMAP4, NKG7, DOK2, STX11, CD7, INPP5D, CD6, JAK3, ICAM2, CCL5, RAB37, MAP4K1, LCK, KLRK1, SEPT1, PRF1, AIF1, AMICA1, MFNG, ITM2A, LCP2, CD3E, SPI1, SLA2, GIMAP5, CD96, IL2RG, CXCL13, TBC1D10C, WAS, GIMAP6, HCK, SNRK, TNFRSF1B, SELPLG, CCR5, CYTH4, SNX20, RGS18, CD52, IKZF1, PLEK, CD247, ZDHHC2, CSF2RA, CSF2RB, ARHGAP25, CD83, TIGIT, CSF1R, GMFG, PRCP, CD8A, PIK3R5, HCST, ITGAL, PIK3CD, SRGN, ITGB2, ZAP70, GGTA1, FLI1, DOCK10, NCKAP1L, PLEKHO2, EBF1

3. AGER, Ube2l6, Irf1, echdc1, ARPC4, ETV7, amt, LOC400759, IDO1, HLA-E, PILRB, HLA-F, GJD3, GBP4, STAT1, BATF2, Rufy4, FBXO6, GBP1, calml4, SAMD9L, SEC31B, CCDC146, HLA-H, APOL1, EXOSC10, Myo15b, LOC115110, OASL, HLA-A, LOC91316, Tapbp, B2M, HLA-B, tap1, TTLL3, TXNDC6, IL15, BTN3A2 ,BTN3A3, micB, Rec8, C19orf38, Zbp1, CHKB-CPT1B, HSH2D, gnb3, HLA-C

4. STAG3, BVES, MAP3K7, RRAGD, C6orf170, LYRM2, MDN1, UBE2J1, CASP8AP2, TRMT11, POP7, PILRB, EPHB4, ZCWPW1, GNB2, GIGYF1, ANKRD6, UFSP1, CNPY4, MCM7, HSPA4L, LRCH4

5. CCDC78, C2orf27A, COLQ, CASC1, SPATA17, FAM154B, C2orf77, CCDC19, C10orf79, ZMYND10, IQCK, WDR54, UNC5CL, TMEM121, WDR66, HOXC6, COL9A2, PIH1D2, EVL, FER1L4, ALKBH3, C11orf74, NAT1, CCDC30, PRICKLE4, MORN1, C1orf88, OSCP1, SPA17, KCNJ8, MESDC1, C14orf79, MYL9, EYA2, CCDC74B, AHNAK2, CADM1, C10orf116, MTL5, SEMA3F, C1orf192, ZNF137, C5orf49, C14orf174, GAS6, DNAH7, HOXC9, CCNL2, CCDC103, GATA3, MGST3, CXCL14, C2orf81, C9orf116, ZNF239, PRR7, RSPH1, BAI2, CCDC114, C19orf51, KCNK1, CROCC, RIPK3, RPP30, RARA, IGFBP4, FZD7, FAM176B, TPPP3, RHOT2, LRRC49, NEK11

6. HPGD, HERC3, SRD5A1, ZNF518B, SC5DL, SEPP1, FKBP5, ALOX15B, APOD, ZNF689, LACTB, ADHFE1, MPV17L, ACPP, SLC41A2, AFMID, IDH1, GALC, CROT, LIMS1, STEAP4, AADAT, PXMP4, ANXA4, AACS, CASP10, GPRC5B, SCP2, SMPDL3A, LACTB2, NSUN2, KYNU, CYP1B1, CYP2R1, APLP2, UBE2G1, DHRS2, HIBADH, MAOA, SLC5A3, KLHL8, AMACR, SGK1, HERC4, OPHN1, ALDH1A3, CLCN7, NPC1, DBI, ABHD6, FITM2, MAML2, PKIB, AK3L1, RMND5A, GPR109A, CTH, AGFG1, CTBS

7. C6orf97, LRRC34, FLOT1, NUCKS1, MDM4, STK19, EXOC2, ZBTB9, CCDC39, SYNGAP1, CMYA5, TCEAL6, ZBTB12, KIAA1529, CREB3L4, FAR1, WDR52, HSD17B4, GAMT, PABPC1L, RERG, CHRD, GLI3, ABAT, PCSK6, DCAF10, TMEM231, RBM39, TRERF1, PRRT2, ZNF692, RDBP, NPHP1, ZNF83, ZNF516, GPSM3, GALNT6, POLN, CCDC14, IFT140, ESR1, DCLRE1A, BBC3, POLL, WASH7P, UQCC, CHST15, SLC7A2, TCEAL3, C10orf78, KITLG, EFHC1, RHBDL1, MPP2, C6orf154, RCOR3, GTF2H4, TPRN, NEK2, ZDHHC6, EHMT2, ZNF525, ZNF37B, TRAF2, GADD45G, TMPRSS3, NSL1, SFI1, C3orf52, SCAND2, ARL3, MGEA5, POLH, BAT1, TAF9B, SEPT8, MAP3K12, KIAA1407, RMND1, TRPV1, GATA3, IGF1R, KIAA0040, LOC143188, LRRC56, LOC678655, ZNF187, C1orf203, LOC729375, LOC100129550, RAB40C, PRR3, BRD2, LOC283050, DACH1, RGS11, GPR77, C5orf30, LRDD, HMGN4, ANXA6, ANKRD10, BAT4, VPS52, AGAP11, RPS6KC1, PAAF1, SHROOM2, PARP10, NUAK2, RANGRF, TTC30B, ZNF137, CLSTN2, ABCC5, TTC30A, C1orf113, DTX3, ANXA9, PLCD4, FNBP4, LZTS2, C1orf226

8. C19orf66, HCP5, HLA-A, IFIT5, ATHL1, UBE2L6, IRF7, TRIM21, HSH2D, OASL, IFIT3, DDX58, HLA-F, IFIT2, RTP4, PSMB8, IFIT1, HLA-B, SAMD9, IFI27, UBA7, PFDN6, HERC6, PSMB9, HLA-H, GBP1, XAF1, RPF1, C19orf38, TREX1, MX2, C3orf62, ZNF404, IFI35, C10orf4, RIBC2, DDX60L, GBP4, B2M, RING1, IRF9, IFITM1, PARP14, IFI27L1, MX1, SIGIRR, LOC115110, PARP12, CCDC18, LOC339047, REC8, PPP1R11, DDAH2, EXOSC10, CCDC101, MESDC1, FAM193B, SYNC, ZC3H11A, CARD16, ZBTB22, PPM1K, ZSCAN16

9. PHF10, PHC2, WDTC1, SBK1, ZNF362, RCC2, CCDC23, MMP25, MAD2L2, ZNRF1, HNRNPA0, PTMA, PARD6G, HIST1H4J, FAM54B, MARCKSL1, TMEM50A, DYNLT1, TMEM88, CDK5R1, FOXP4, H3F3A, SYTL3, PATZ1, CMIP, KPTN, DCLK1, C1orf144, HMGN3, CAPS

10. SIK2, CREB3L2, GPRIN3, CAPN9, RALGAPA2, DOCK4, CDKL5, ERN1, NAGLU, KIAA1147, ANKRD36BP1, BAG4, MLL3, LMTK2, ADAM9, PLEKHA2, FAM63B, MPP7, DENND1B, PRKAA1, SERINC5, SGK196, TRMT2B, SEC24A, UGGT1, AVL9, GOLGA4, PRKAR2A, PARD3B, LOC283922, XYLT1, HIPK3

11. PPP2R3A, SLC23A2, MLL, CBL, MGA, NBEAL1, RC3H2, MAP3K2, TTF2, ZFP91, CNTN1, DOCK10, RASA2, NPHP3, ZNF318, C10orf18, DSP, HIPK1, MLX, RNF214, FAM168A, NOTCH2, ARL10, PPARA, SAMD8

12. MYO10, FAM105B, BEST1, PDZD2, PAPD7, NLRX1, CLPTM1L, ELF4, SMURF1, CCT5, KIAA0947, IFRD1, GRB10, PLEKHA8, CXorf56

13. PLAT, LEF1, C12orf68, TBX2, METRNL, MSX1, PSD, THBD, FBXL15, MIF, ZNF628, C7orf50, CHST12

14. ZMYND17, GALNTL1, C9orf46, CNIH2, HMGN2, CIRBP, HOXD9, CCL5

APPENDIX C: ADDITIONAL SHAKESPEARE TEXT RESULTS

Table C.1: Coherent word sets in Shakespearean tragedies

1. them, cheer
2. young, boy
3. spirit, now
4. prithee, yet
5. our, whence
6. fury, on
7. image, hand
8. think, opinion
9. than, more
10. small, much
11. courage, on
12. he, his
13. going, go
14. than, rather
15. my, fight
16. noble, present
17. on, deed
18. did, wast
19. when, past
20. move, our
21. earth, heaven
22. well, fare
23. hate, do
24. war, people
25. speak, speaks
26. our, cries
27. cruel, most
28. being, against
29. our, hearing
30. mother, make
31. hath, spend
32. else, pleasure
33. your, welcome
34. which, parts
35. man, wits
36. go, humour
37. much, health
38. met, no
39. least, our
40. less, would
41. vile, out
42. our, affairs
43. above, from
44. there, brow
45. ours, our, us
46. too, indeed, time
47. he, his, powers
48. soon, hath, ere
49. thousand, ten, twenty
50. us, are, labour
51. had, twas, been
52. brothers, if, brother, were
53. she, her, lady, madam, husband, wife, queen, woman, daughter, shes, marriage, me, tell, sister, herself, sweet
54. hast, dost, art, thy, wilt, thee, thine, thou, death, shalt, canst, didst, ill, sweet, ah, hadst, if, thyself, away, father, eyes, boy, villain, child, mine, mother, kill, wert, me, then, die, o, flesh, am, cheeks, leave, young, sight
55. king, duke, majesty, lords, france, prince, grace, god, princely, unto, liege, sovereign, crown, english, french, highness, uncle, princes, arms, lord, gracious, subjects, cousin, soul, title, now, blood, fathers, then, until, queen, father, traitor, yield, son, right, royal, john, forward, brother, doth, presence, heir, war, sons, embrace, hath, guilty
56. pray, has, sir, fellow, good, know, indeed, theres, ha, does, knave, hes, whats, matter, said, honest, think, would, some, time, go, can, faith, nay, am, thats, most, lady, beseech, ont, ay, say, marry, could, ist, prithee, you, yourself, well, tis, fit, ones, yes, such, so, thing, em, no, how, very, fool, wife, tell, put, sorry, are, what, there, better, her, never, sirrah, out, shes, sure, tot, one, your, much, ho

Table C.2: Word sets in Shakespearean tragedies clustered by TF-IDF adjusted distance

1. ah, tears
2. child, lady
3. city, enter
4. english, french
5. father, son
6. gates, town
7. god, grace
8. letter, villain
9. mad, sing
10. oath, swear
11. thou, thy
12. you, your
13. anon, falstaff, four
14. art, dost, thee
15. beauty, fair, ladies
16. husband, wife, woman
17. peace, right, whose
18. plague, rascal, rogue
19. brothers, children, mighty, suit
20. drunk, tonight, watch, wine
21. fellow, says, sirrah, whoreson
22. gracious, heir, sovereign, unto
23. feast, murder, revenge, sent, witness
24. five, hundred, knight, morrow, today
25. to, awake, dream, gentlemen, sleep, tomorrow
26. hear, honourable, mark, read, speak, wrong
27. age, cause, honour, most, nature, noble, poor
28. die, fight, fly, hurt, quarrel, slain, soldiers, sword
29. against, church, forsworn, hang, holy, law, mayst, need, priest
30. court, ha, marry, old, said, shadow, silence, very, yea
31. dead, death, deed, didst, eyes, kill, killd, life, tender, wilt
32. within, call, coward, devil, door, faith, fat, hast, prithee, seven, two
33. appear, bears, denied, durst, endure, faults, gown, humour, justice, letters, roman, yourselves
34. ancient, bravely, damned, discharge, mistress, neighbour, quiet, receive, stuff, troth, warrant, whether, wicked
35. does, drink, gods, ho, indeed, ist, madness, matter, nay, pray, sense, t, theres, think, tis, whats
36. ass, beast, beggar, below, dog, fools, forgot, hadst, hate, ild, mend, mere, misery, shouldst, thief, thine, thyself, want, wealth, wert, wouldst
37. arm, arms, base, blood, body, day, doth, earth, eye, farewell, foul, hand, hands, head, heaven, mouth, myself, power, proud, royal, saint, soul, souls, sweet, tale, tongue
38. aside, between, business, confess, conscience, drown, face, fie, free, hes, home, lost, marriage, methought, please, pluck, pound, presence, purse, red, sea, sentence, side, thousand, truth, wear, white, wit, work, worthy, youre

APPENDIX D: ADDITIONAL LAST.FM RESULTS

Table D.1: Coherent neighborhood for "Slayer"

Artist	Top 5 Tags
Slayer	thrash metal, heavy metal, metal, power metal, death metal
Iron Maiden	heavy metal, metal, power metal, hard rock, seen live
Metallica	thrash metal, heavy metal, metal, hard rock, awesome
Megadeth	thrash metal, heavy metal, metal, cool, power metal
Motrhead	heavy metal, hard rock, metal, thrash metal, uk
Black Sabbath	heavy metal, hard rock, metal, classic rock, 70s
Pantera	thrash metal, heavy metal, power metal, metal, 90s
Judas Priest	heavy metal, hard rock, metal, classic rock, thrash metal
Sepultura	thrash metal, death metal, brazilian, heavy metal, metal
Kreator	thrash metal, metal, heavy metal, power metal, german
Anthrax	thrash metal, heavy metal, metal, cool, american
AC/DC	hard rock, heavy metal, classic rock, 70s, metal
Children of Bodom	melodic death metal, death metal, power metal, metal, gothic
Death	death metal, progressive metal, melodic death metal, thrash metal, metal
Exodus	thrash metal, heavy metal, metal, 80s, rock
Led Zeppelin	70s, classic rock, progressive rock, hard rock, blues
Testament	thrash metal, heavy metal, death metal, metal, seen live
Deep Purple	hard rock, progressive rock, classic rock, heavy metal, 70s

Table D.2: Coherent neighborhood for "Brandy"

Artist	Top 5 Tags
Brandy	ballad, rnb, sexy, soul, hip-hop
Rihanna	rnb, ballad, sexy, love, dance
Mariah Carey	rnb, soul, love, ballad, female
Beyonce	rnb, sexy, soul, ballad, female vocalist
Christina Aguilera	ballad, sexy, soul, rnb, amazing
The Pussycat Dolls	rnb, sexy, favorites, dance, pop
Jennifer Lopez	female, rnb, female vocalist, dance, sexy
Ciara	rnb, hip hop, hip-hop, sexy, amazing
Janet Jackson	rnb, female, sexy, soul, female vocalist

Table D.3: Coherent neighborhood for "Creedence Clearwater Revival"

Artist	Top 5 Tags
Creedence Clearwater Revival	60s, classic rock, 70s, folk, blues
Led Zeppelin	70s, classic rock, progressive rock, hard rock, blues
The Doors	psychedelic, 60s, classic rock, blues, rock
The Rolling Stones	60s, classic rock, blues, 70s, british
The Beatles	60s, classic rock, british, psychedelic, <3
Pink Floyd	progressive rock, psychedelic, classic rock, 70s, 60s
AC/DC	hard rock, heavy metal, classic rock, 70s, metal
Deep Purple	hard rock, progressive rock, classic rock, heavy metal, 70s
Queen	classic rock, 70s, hard rock, 80s, progressive rock
Black Sabbath	heavy metal, hard rock, metal, classic rock, 70s
The Who	60s, classic rock, uk, hard rock, 70s
Jimi Hendrix	blues, psychedelic, classic rock, 60s, funk

BIBLIOGRAPHY

Aggarwal, C. C., Li, Y., Wang, J., and Wang, J. (2009). Frequent pattern mining with uncertain data. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 29–38, New York, NY, USA. ACM.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I., et al. (1996). Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328.

Anderson, T. (1959). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience.

Anderson, T. W. (1958). *An Introduction to Multivariate Analysis*. Wiley.

Bassi, F. and Hero, A. (2012). Large scale correlation detection. *Proc. of the IEEE International Symposium on Information Theory*, pages 2591–2595.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188.

Bickel, P. and Levina, E. (2008). Covariance regularization via thresholding. *Ann. Statist.*, 34(6):2577–2604.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science and Business Media, LLC., Oxford, England.

Bockmayr, M., Klauschen, F., Gyrffy, B., Denkert, C., and Budczies, J. (2013). New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Systems Biology*, 7(1):78.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc.

Browne, M. and Shapiro, A. (1986). The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra and its Applications*, 82:169–176.

Burdick, D., Calimlim, M., and Gehrke, J. (2001). Mafia: A maximal frequent itemset algorithm for transactional databases. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 443–452. IEEE.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and constructions of compressed sensing materials. *Ann. Statist.*, 39(3):1496–1525.

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journ. of the Am. Stat. Ass.*, 108:265–277.

Cai, T. T. and Zhang, A. (2014). Inference on high-dimensional differential correlation matrix. *Technical Report*.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

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

Celma, O. (2010). *Music Recommendation and Discovery in the Long Tail*. Springer.

Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pages 43–48.

Choi, T., Ramamoorthi, R., et al. (2008). Remarks on consistency of posterior distributions. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 170–186. Institute of Mathematical Statistics.

Choi, Y. and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21):2780–2786.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cdna microarray experiments. *Genome Biology*, 4(4):210.

Datta, S. and Datta, S. (2002). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4).

Derado, G., Bowman, F. D., and Kilts, C. D. (2010). Modeling the spatial and temporal dependence in fmri data. *Biometrics*, 66(3):949–957.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. *Twenty-first international conference on Machine learning - ICML '04*.

Donner, A. and Zou, G. (2014). Testing the equality of dependent intraclass correlation coefficients. *J. R. Statist. Soc. D*, 51(3):367–379.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3):75–174.

Friguet, C., Kloareg, M., and Causeur, D. (2012). A factor model approach to multiple testing under dependence. *J. Am. Statist. Ass.*, 104(488):1406–1415.

Fukushima, A. (2013). Diffcorr: An r package to analyze and visualize differential correlations in biological networks. *Gene 518*, pages 209–214.

Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11(95):1471–2105.

Gopalan, P., Ruiz, F. J., Ranganath, R., and Blei, D. M. (2014). Bayesian nonparametric poisson factorization for recommendation systems. In *AISTATS*, pages 275–283.

Grahne, G. and Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In *FIMI*, volume 90.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2002). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *P.N.A.S*, 100(1):253–258.

Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. (2012). *arules: Mining Association Rules and Frequent Itemsets*. R package version 1.0-12.

Harman, H. H. (1960). *Modern factor analysis*. Univ. of Chicago Press, New York.

Hu, R., Qiu, X., and Glazko, G. (2010). A new gene selection procedure based on the covariance distance. *Bioinformatics*, 26(3):348–354.

Iglesia, M. D., Vincent, B. G., Parker, J. S., Hoadley, K. A., Carey, L. A., Perou, C. M., and Serody, J. S. (2014). Prognostic b-cell signatures using mrna-seq in patients with subtype-specific breast and ovarian cancer. *Clinical Cancer Research*, 20(14):3818–3829.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Jakobsson, M. and Rosenberg, N. A. (2007). Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions*, 16.11:1370–1386.

Kriegel, H.-P., Krger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. on Knowledge Disc. from Data (TKDD)*, 3(1).

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, 6(4):e18961.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330–342.

Li, T. A unified view on clustering binary data. In *Machine Learning*, page 2006.

Li, T. (2005). A general model for clustering binary data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 188–197, New York, NY, USA. ACM.

Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.

Liu, B.-H., Yu, H., Tu, K., Li, C., Li, Y.-X., and Li, Y.-Y. (2010). Dcgl: an r package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinformatics*, 26(20):2637–2638.

Liu, H., Han, J., Xin, D., and Shao, Z. (2006). Mining frequent patterns from very high dimensional data: A top-down row enumeration approach. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 282–293. SIAM.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

MacMahon, M. and Garlaschelli, D. (2015). Community detection for correlation matrices. *Physical Review X*, 5(2).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. of the fifth Berkeley Symp. on math. stat. and prob.*, 1(14).

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, Inc.

Neuhaus, J., Kalbfleisch, J., and Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59(1):25–35.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Palowitch, J., Bhamidi, S., and Nobel, A. B. (2016). The continuous configuration model: A null for community detection on weighted networks. *arXiv preprint arXiv:1601.05630*.

Pan, F., Tung, A. K., Cong, G., and Xu, X. (2004). Cobbler: combining column and row enumeration for closed pattern discovery. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, pages 21–30. IEEE.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2008). Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Ass.*, 104(486).

Perou, C. M., Srlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–752.

Phan, K. L., Wager, T., Taylor, S. F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in pet and fmri. *NeuroImage*, 16:331–348.

Prabha, S., Shanmugapriya, S., and Duraiswamy, K. (2013). A survey on closed frequent pattern mining. *International Journal of Computer Applications*, 63(14):47–52.

Rajaratnam, B., Massam, H., and Carvalho, C. (2008). Flexible covariance estimation in graphical models. *Ann. Statist.*, 36:2818–2849.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM.

Sheng, E., Witten, D., and Zhou, X.-H. (2016). Hypothesis testing for differentially correlated features. *Biostatistics*, 17(4):677?691.

Slepian, D. (1962). The one-sided barrier problem for gaussian noise. *Bell Labs Technical Journal*, 41(2):463–501.

Sohrabi, M. K. and Barforoush, A. A. (2012). Efficient colossal pattern mining in high dimensional datasets. *Know.-Based Syst.*, 33:41–52.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(91).

Stamatatos, E. (2009). A comparison of methods for differential expression analysis of rna-seq data. *Journ. of the Am. Soc. for Info. Science and Tech.*, 60(3):538–556.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psych. Bulletin*, 87(2):245–251.

Steiger, J. H. and Hakstian, A. R. (1982). The asymptotic distribution of elements of a correlation matrix: Theory and application. *Brit. Journ. of Math. and Stat. Psych.*, 35:208–215.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11(1):497.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tong, Y., Chen, L., Cheng, Y., and Yu, P. S. (2012). Mining frequent itemsets over uncertain databases. *Proceedings of the VLDB Endowment*, 5(11):1650?1661.

Voets, N., Adcock, J., Flitney, D., Behrens, T., Hart, Y., Stacey, R., Carpenter, K., and Matthews, P. (2006). Distinct right frontal lobe activation in language processing following left hemisphere injury. *Brain*, 129(3):754–766.

Wainer, H. and Braun, H. I. (2013). *Test Validity*. Routledge.

Wang, J., Fan, L., Wang, Y., Xu, W., Jiang, T., Fox, P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2015). Determination of the posterior boundary of wernicke's area based on multimodal connectivity profiles. *Human Brain Mapping*, 36:1908–1924.

Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2016). Community extraction in multilayer networks with heterogeneous community structure. *arXiv preprint arXiv:1610.06511*.

Wilson, J. D., Wang, S., Mucha, P. J., Bhamidi, S., and Nobel, A. B. (2014). A testing based extraction algorithm for identifying signicant communities in networks. *The Annals of Applied Statistics*, 8(3):1853–1891.

Xia, Y., Cai, T., and Cai, T. T. (2014). Testing differential networks with applications to detecting gene-by-gene interactions. *Biometrika (to appear)*.

Zaki, M. J. et al. (1999). Parallel and distributed association mining: A survey. *IEEE concurrency*, 7(4):14–25.

Zaki, M. J. and Hsiao, C.-J. (2002). Charm: An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining*, pages 457–473. SIAM.

Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W. (1997a). Parallel algorithms for discovery of association rules. *Data mining and knowledge discovery*, 1(4):343–373.

Zaki, M. J., Parthasarathy, S., Ogihara, M., Li, W., et al. (1997b). New algorithms for fast discovery of association rules. In *KDD*, volume 97, pages 283–286.

Zhang, K. (2017). Bet on independence. *pre-print*.

Zhang, Q., Li, F., and Yi, K. (2008). Finding frequent items in probabilistic data. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 819–832.

Zhou, C., Han, F., Zhang, X., and Liu, H. (2015). An extreme-value approach for testing the equality of large u-statistic based correlation matrices. *arXiv:1502.03211*.

Zhu, F., Yan, X., Han, J., Philip, S. Y., and Cheng, H. (2007). Mining colossal frequent patterns by core pattern fusion. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 706–715. IEEE.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.