UNIVERSITY OF CALIFORNIA
RIVERSIDE

Clustering: Algorithm, Optimization and Inference

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Zhanpan Zhang

December 2011

Dissertation Committee:

Dr. Xinping Cui, Co-Chairperson
Dr. Daniel R. Jeske, Co-Chairperson
Dr. Subir Ghosh
Dr. James Borneman

The Dissertation of Zhanpan Zhang is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to my advisor, Dr. Xinping Cui, whose encouragement, support and understanding have provided a good basis to this dissertation. I appreciate her time and ideas to make my Ph.D. experience productive.

I am deeply grateful to my co-advisor, Dr. Daniel R. Jeske. His wide knowledge and logic way of thinking have been of great value for me. This dissertation would not have been possible without his guidance and persistent help.

I wish to express my warm and sincere thanks to Dr. Subir Ghosh and Dr. James Borneman for their continuous assistance and encouragement.

The text of this dissertation, in part, is a reprint of the material as it appears in "Clustering Scatter Plots Using Data Depth Measures" at the 6th International Conference on Data Mining (DMIN'10). The co-authors, Dr. Xinping Cui and Dr. Daniel R. Jeske listed in that publication directed and supervised the research which forms the basis for this dissertation. The co-authors, Dr. James Borneman, Xiaoxiao Li and Dr. Jonathan Braun, contributed effort to data analysis and provided valuable comments. I would also like to thank Dr. Mark S. Hoddle for his expertise and constructive discussion.

DEDICATIONS

This dissertation is dedicated to my family for their unconditional love and support. I could not have achieved this goal without them.

ABSTRACT OF THE DISSERTATION

Clustering: Algorithm, Optimization and Inference

by

Zhanpan Zhang

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2011
Dr. Xinping Cui, Co-Chairperson
Dr. Daniel R. Jeske, Co-Chairperson

Clustering is rapidly becoming a powerful data mining technique, and has been broadly applied to many domains. Usually data are arranged in a matrix with rows and columns, and each cell of this matrix is a real number. This dissertation aims at developing clustering algorithms with statistical inference incorporated in the following two scenarios.

First, when each cell of the data matrix is not represented by a single numerical value and instead contains a scatter plot, the existing clustering methods are not applicable any more. In this dissertation, we develop both hierarchical clustering and co-clustering procedure to handle a data matrix of scatter plots. To more accurately reflect the nature of data, we introduce a dissimilarity statistic based on “data depth” to measure the

discrepancy between two bivariate distributions without oversimplifying the nature of the underlying pattern. We also propose novel painting metrics and construct heat maps to allow visualization of the clusters. We demonstrate the utility and power of our proposed clustering methods through simulation studies and application to a microbe-host-interaction study.

Second, when spatial information is embedded in the data matrix, the order of rows and columns can not be changed. Model-based spatial co-clustering has not been well studied. In this dissertation, we develop a co-clustering method using a Generalized Linear Mixed Model (GLMM) for spatial data. To avoid the high computational intensity associated with global optimization, we propose a heuristic optimization algorithm to search for a near optimal co-clustering. A sampling strategy is introduced to capture as much of the spatial information that is available from the sparse data as possible. For an application pertinent to Integrated Pest Management (IPM), we combine the spatial co-clustering technique with a statistical inference method to make assessment of pest density more accurate. We demonstrate the utility and power of our proposed pest assessment procedure through simulation studies and apply the procedure to a study of the persea mite (*Oligonychus perseae*), a pest of avocado trees.

TALBE OF CONTENTS

Chapter 1 Introduction 1

1.1 ONE-DIMENSIONAL CLUSTERING 1

1.2 CO-CLUSTERING 3

1.3 MULTI-DIMENSIONAL CLUSTERING 12

1.4 DATA VISUALIZATION 14

1.5 OPEN ISSUES 15

Chapter 2 Clustering Scatter Plots Using Data Depth Measures 18

2.1 INTRODUCTION 18

2.2 METHODOLOGY 20

2.2.1 Clustering Procedure 20

2.2.2 Hypothesis Testing 22

2.2.3 Data Depth 25

2.2.4 $Q(F,G)$ vs. $Q(G,F)$ 27

2.3 SIMULATION STUDY 28

2.4 DATA VISUALIZATION 32

2.5 APPLICATION 37

2.5.1 Motivation 37

2.5.2 Results 38

2.6 CONCLUSION 41

Chapter 3 Co-clustering Scatter Plots Using Data Depth Measures 43

3.1 INTRODUCTION 43

3.2 METHODOLOGY 44

3.2.1 Co-clustering Procedure 44

3.2.2 Hypothesis Testing 49

3.3 SIMULATION STUDY 51

3.4 DATA VISUALIZATION 54

3.5 APPLICATION 58

3.5.1 Results 58

3.5.2 Biological Hypothesis 64

3.6 CONCLUSION 66

Chapter 4 Co-clustering Spatial Data Using a Generalized Linear Mixed Model With Application to the Integrated Pest Management 67

4.1 INTRODUCTION 67

4.2 METHODOLOGY 69

4.2.1 Spatial GLMM 69

4.2.1.1 Model Definition 69

4.2.1.2 Likelihood and Parameter Estimation 72

4.2.2 Model-based Co-clustering 73

4.2.2.1 Global Optimization 73

4.2.2.2 Heuristic Optimization 76

4.2.2.3 Efficiency of Heuristic Optimization Algorithm 77

4.2.2.4 Heuristic Optimization with Non-Exhaustive Samples 79

4.3 APPLICATION TO PEST DENSITY ASSESSMENT 84

4.3.1 Proposed Methodology 84

4.3.2 Coverage Probability 89

4.3.3 Simulation Study 96

4.4 EXAMPLE 99

4.5 DISCUSSION 102

Chapter 5 Discussion 105

5.1 VECTOR-BASED CLUSTERING 105

5.2 EXTENSION TO MULTI-DIMENSIONAL CLUSTERING 109

Bibliography 113

LIST OF FIGURES

Figure 1.1	Co-cluster Type	4
Figure 1.2	Co-cluster Structure	4
Figure 2.1	Data Structure: A Data Matrix of Scatter Plots	21
Figure 2.2	Three Data Pattern Settings	29
Figure 2.3	Success Rate versus Location Shift	31
Figure 2.4	Painting Metrics	34
Figure 2.5	Painting Example 1	34
Figure 2.6	Painting Example 2	35
Figure 2.7	Painting Example 3	35
Figure 2.8	Painting Example 4	36
Figure 2.9	Heat Map with the OQI Painting Metric	40
Figure 3.1	Data Structure: A Data Matrix of Scatter Plots	45
Figure 3.2	A Co-clustering Example	49
Figure 3.3	Co-cluster Specification	53
Figure 3.4	Probability of Consistency	53
Figure 3.5	Painting Example 1	55
Figure 3.6	Painting Example 2	55

Figure 3.7 Painting Example 3 56

Figure 3.8 Painting Example 4 56

Figure 3.9 Heat Map with the OQI Painting Metric 61

Figure 4.1 “Checkerboard” Co-cluster Structure 71

Figure 4.2 Heuristic Optimization vs. Global Optimization 79

Figure 4.3 Sampling Strategy 80

Figure 4.4 Success Rate of Design vs. Sample Size 83

Figure 4.5 Histogram of Conditional Prediction Interval Width for $\kappa = 1$ 94

Figure 4.6 Histogram of Conditional Prediction Interval Width for $\kappa = 5$ 95

Figure 4.7 Probabilities of Correct Decision 99

Figure 4.8 Pest Assessment for Orchard A 100

Figure 4.9 Pest Assessment for Orchards B and C 101

Figure 4.10 Pest Assessment for Orchard D (Integrated Orchard) 102

Figure 4.11 “Tree” Co-cluster Structure 103

Figure 5.1 Polynomial Regression Model 107

Figure 5.2 A Nonparametric Approach 108

Figure 5.3 A Three-dimensional Example 111

Figure 5.4 Information-theoretic Measures 112

LIST OF TABLES

Table 3.1	Proteins and Microbes in the Identified Co-clusters	62
Table 4.1	Number of Co-clusterings for Global Optimization	75
Table 4.2	Coverage Probability for $\kappa = 1$	93
Table 4.3	Coverage Probability for $\kappa = 5$	93
Table 4.4	Confusion Matrix	97

Chapter 1

Introduction

Clustering is rapidly becoming a powerful data mining technique, and has been broadly applied to many domains such as bioinformatics and text mining. Usually data are arranged in a matrix with rows and columns, and each cell of this matrix is a real number. A large number of clustering methods have been studied in the literature, which include one-dimensional clustering, co-clustering, and multi-dimensional clustering (for multi-dimensional data). In addition to summarize the clustering methods, we also briefly review the methods of data visualization for the clustering results in this chapter.

1.1. ONE-DIMENSIONAL CLUSTERING

One-dimensional clustering is to divide rows, or columns, into a number of groups. For simplicity, only the row clustering is discussed in this section. Here we review two commonly used one-dimensional clustering methods: hierarchical clustering and partitioning clustering. One may see Andreopoulos et al. (2009) and Jiang et al. (2004) for a survey.

Hierarchical clustering builds a hierarchy of clusters based on the dissimilarity (distance) measures among rows, such as Euclidean distance and Pearson's correlation

coefficient, whose result can be graphically presented in a tree structure, called dendrogram. The clustering algorithm initially regards each row as an individual cluster, and at each step, merges the closest pair of clusters until all the rows are merged into one cluster. The distance between two clusters may be determined by different criteria. Single linkage defines the distance between two clusters to be the minimum distance between any pair of rows, one row from a cluster and the other row from the other cluster. Complete linkage defines the distance between two clusters to be the maximum distance between any pair of rows, one row from a cluster and the other row from the other cluster. Average linkage defines the distance between two clusters to be the average distance between all pairs of rows, one row from a cluster and the other row from the other cluster. Some applications of hierarchical clustering can be found in Eisen et al. (1998), Kaplan et al. (2004), Baehrecke et al. (2004), and Loewenstein et al. (2008).

Partitioning clustering, such as K-means (MacQueen 1967), divides rows into a pre-specified number of clusters, say K clusters, in which each row belongs to the cluster with the nearest mean. First K initial cluster mean vectors (centroids) are selected (Huang 1998), which represent K clusters. Each row is assigned to the closest cluster whose centroid has the smallest distance from this row. For each cluster, the corresponding centroid is updated with the mean of all rows that belong to this cluster. Then each row is reassigned to a new cluster based on the updated centroids, and the

above procedure repeats. The iteration continues until the number of rows changing clusters is below a user-specified threshold. More partitioning clustering methods and applications can be found in Hochbaum & Shmoys (1985), Kaufman & Rousseeuw (1987), Gasch & Eisen (2002), and Chopra et al. (2008).

1.2. CO-CLUSTERING

Co-clustering, also called biclustering, bivariate clustering, or two-mode clustering, has been an active area of research in recent years, resulting in the development of a wide variety of approaches and algorithms. Different from the one-dimensional clustering methods that seek to identify similar rows or columns independently, co-clustering simultaneously clusters rows and columns to identify “blocks” (or “co-clusters”) of rows and columns that show highly inter-related coherence. For example, in gene expression analysis, co-clustering can be used to solve the dual problem of identifying a set of genes and conditions simultaneously involved in a metabolic process, a problem that traditional one-dimensional clustering methods cannot handle. Moreover, co-clustering is desirable over traditional one-dimensional clustering as it is more informative and easily interpretable while preserving most of the information contained in the original data; and it allows dimensionality reduction along both axes simultaneously and hence leads to a much more compact representation for subsequent analysis.

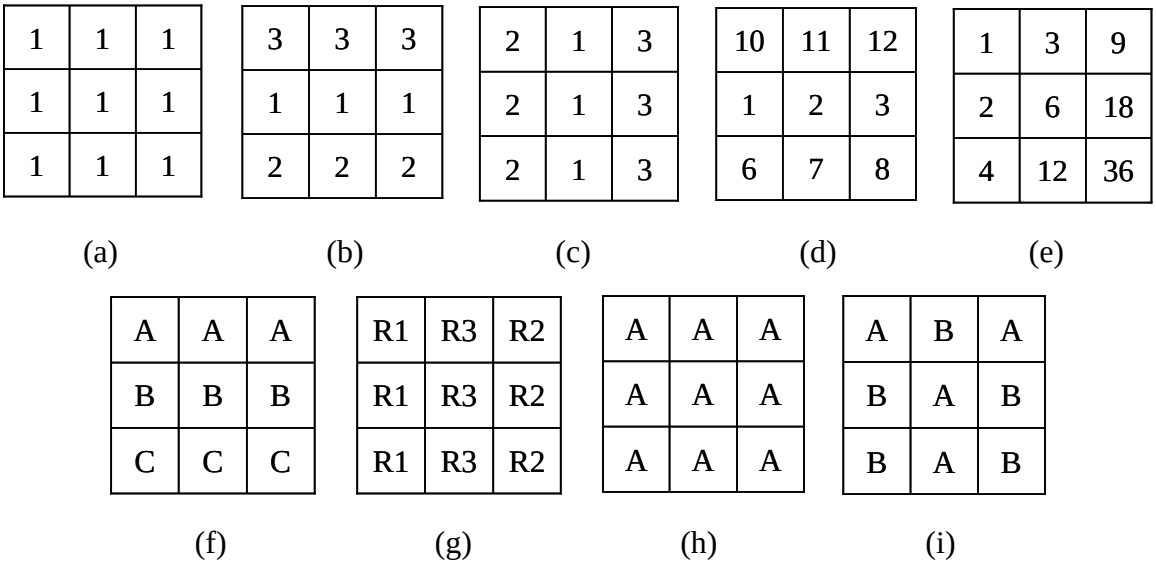


Figure 1.1. Co-cluster Type: (a) constant; (b) constant rows; (c) constant columns; (d) additive model; (e) multiplicative model; (f) common status rows; (g) common order rows; (h) common status; (i) simultaneous status change along rows and columns.

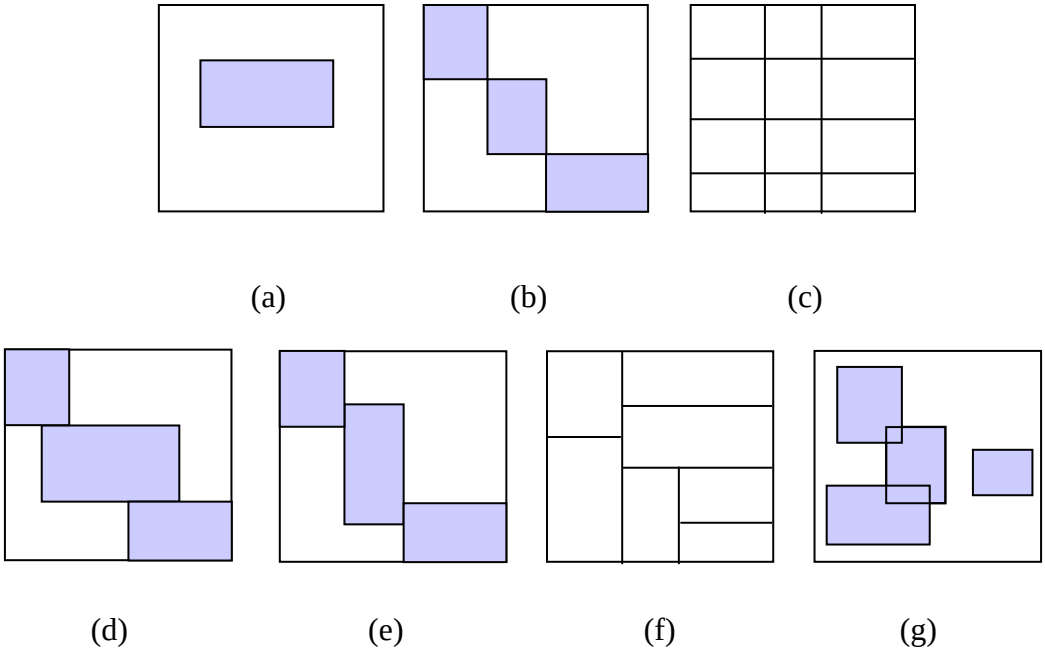


Figure 1.2. Co-cluster Structure: (a) one co-cluster; (b) exclusive clustered rows and columns; (c) checkerboard structure; (d) exclusive clustered rows; (e) exclusive clustered columns; (f) tree structure; (g) overlapping co-clusters.

The most common framework for co-clustering methods is to first define a meaningful objective function to evaluate the quality of co-clusters, and then develop an algorithm to find the co-clusters that optimizes the objective function. Figure 1.1 summarizes a number of co-cluster types that have been defined in the literature, and Figure 1.2 shows the possible co-cluster structures that may exist in the data matrix.

Hartigan (1972), also known as block clustering, has been considered as one of the earliest co-clustering papers, which used a version of squared Euclidean distance as the objective function minimized by a "divide and conquer" direct clustering algorithm. To avoid the situation that each cell of the data matrix forms a co-cluster, the number of co-clusters is usually pre-specified. The algorithm splits the original data matrix into a number of non-overlapping co-clusters as shown in Figure 1.2(f), each of which follows the co-cluster type of Figure 1.1(a).

Knowing that overlapping clusters are very natural in biology, Cheng & Church (2000) used a model-based squared Euclidean distance as the objective function minimized by a greedy iterative search. The resulting co-clusters can be overlapping as shown in Figure 1.2(g), and the type of co-clusters can be any one of Figures 1.1(b)-(d). However, the iterative insertion and deletion based algorithm causes random perturbations to the data that may mask previously discovered co-clusters. Also, the algorithm identifies co-clusters one at a time sequentially rather than all at once.

Califano et al. (2000) introduced a pattern discovery algorithm to discover statistically significant patterns in which the values of each gene (row) are consistent across a subset of columns, as shown in Figure 1.1(b). An optimal set of patterns is then chosen among the statistically significant ones using a greedy set covering algorithm.

Getz et al. (2000) proposed a coupled two-way clustering (CTWC) analysis based on the iterative row and column clustering combination. Any reasonable one-dimensional clustering method can be used within the framework of CTWC. The authors used a hierarchical clustering algorithm to generate stable clusters of rows and columns at each iteration, and consequently discover a set of co-clusters at a time. Due to the normalization step, the type of co-clusters can be any one of Figures 1.1(a)-(e).

Lazzeroni & Owen (2000) introduced the plaid model with each cell in the data matrix viewed as a sum of terms called layers (corresponding to co-clusters), which incorporates additive two-way ANOVA models within co-clusters. The co-cluster type may be any one of Figures 1.1(a)-(d). The overlapping co-cluster structure is directly modeled in the plaid model approach, and multiple co-clusters can be identified sequentially instead of simultaneously.

Using the similar objective function to that in Cheng & Church (2000), Yang et al. (2002) and Yang et al. (2003) introduced a greedy move-based optimization algorithm FLOC (FLexible Overlapping biClustering) that can simultaneously discover a set of

possibly overlapping co-clusters when dealing with the data with missing values. Likewise, the co-cluster structure and the co-cluster types identified by FLOC are same as Cheng & Church (2000).

Ben-Dor et al. (2002) suggested looking for order-preserving submatrices (OPSMs), in which the values of all the rows induce the same linear ordering of columns, as shown in Figure 1.1(g). This approach focuses on the uniformity of the relative order of columns rather than on the uniformity of the actual values, therefore is potentially more robust to the stochastic nature of the observed values and to the variation caused by the measurement process.

Busygin et al. (2002) proposed Double Conjugated Clustering (DCC) that implements a coupled conjugated node-driven clustering method processing the rows and columns of the data matrix and synchronizing the two spaces by means of a projection between row-space and column-space. In this framework, Self-Organizing Maps (SOM) is recommended, and the angle metric is used as similarity measure. The co-cluster structure is shown in Figure 1.2(b).

Murali & Kasif (2003) introduced a greedy algorithm to find the conserved gene expression motifs (xMOTIFs) in gene expression analysis, in which a subset of genes (rows) are simultaneously conserved for a subset of samples (columns), as shown in Figure 1.1(f). A gene (row) is conserved across a set of samples (columns) if the value of

this row for each column is within the same range, denoted by A, B and C in Figure 1.1(f). Although many xMOTIFs may exist, the authors were only interested in the largest xMOTIF, the one that contains the maximum number of conserved genes (rows), as shown in Figure 1.2(a).

For the discretized data, Sheng et al. (2003) tackled the co-clustering problem in the Bayesian framework by presenting a co-clustering strategy based on a simple frequency model for the pattern of a co-cluster and on Gibbs sampling for parameter estimation. To enable the detection of multiple co-clusters, the authors mask the rows (columns) that belong to the previously discovered co-clusters, and rerun the algorithm on the rest of the data. The algorithm discovers one co-cluster at a time, and is iterated until no co-cluster can be found for the unmasked part of the data matrix. The co-cluster type is shown in Figures 1.1(b)-(c), and the co-cluster structure is shown in Figures 1.2(d)-(e).

Cho et al. (2004) and Cho & Dhillon (2008) considered two versions of squared Euclidean distance that are similar to those used by Hartigan (1972) and Cheng & Church (2000). Two fast K-means like co-clustering algorithms were proposed to identify a checkerboard co-cluster structure as shown in Figure 1.2(c), and therefore simultaneously discover a number of co-clusters as opposed to one co-cluster at a time like Cheng & Church (2000). The co-cluster type may be any one of Figures 1.1(a)-(d).

Rogers and Kulkarni (2005) extended the mixed-integer linear programming model from one-dimension clustering to co-clustering. A genetic algorithm was developed to optimize the objective function that is a sum of dissimilarity measures from the family of Minkowski metrics. The genetic algorithm enables clustering large-sized data, which is a formidable challenge for many conventional algorithms. The co-cluster structure is shown in Figure 1.2(b).

Reiss et al. (2006) reformulated Yang et al. (2003)'s δ-cluster model with a Markov Chain model, which enables integration of additional information as well as a prior distribution for constraining co-cluster size and redundancy. The authors developed an algorithm called cMonkey with the iterative optimization conducted using Markov Chain Monte Carlo methods. The procedure constructs one co-cluster at a time, and will stop when a given number of co-clusters have been generated, or significant optimization is no longer possible.

Pensa & Boulicaut (2008) and Pensa et al. (2010) considered the same co-cluster types and the same checkerboard co-cluster structure as Cho et al. (2004), and also used the same objective functions as those in Cho et al. (2004). An iterative constraint-based co-clustering algorithm was introduced to exploit user-defined constraints such as the case that the selected rows and/or columns must (not) be in the same co-cluster.

Another type of objective function used in the literature is the Kullback-Leibler (KL) divergence, pioneered by Dhillon et al. (2003), which considered a checkerboard co-cluster structure, and proposed an information-theoretic co-clustering method. The data matrix of nonnegative values is treated as a joint probability distribution between two discrete random variables. With pre-specifying the number of clusters for each dimension, the authors aimed at finding the optimal co-clustering that leads to the largest mutual information between the clustered random variables, or equivalently, the one that minimizes the difference (loss) between the mutual information of the original random variables and the mutual information of the clustered random variables.

Banerjee et al. (2007) extended Dhillon et al. (2003)'s work by introducing a more general objective function "Bregman divergence" that includes both squared Euclidean distance and KL-divergence as special cases. Multiple structurally different co-clustering schemes are allowed that preserve various linear statistics of the original data matrix. The authors introduced a minimum Bregman information (MBI) principle that simultaneously generalizes the well-known maximum entropy and standard least squares principles to all Bregman loss functions, and leads to a matrix approximation that is optimal among all generalized additive models in a certain natural parameter space.

A connection between data matrices and graph theory has been also established in the literature. Dhillon (2001) introduced a bipartite graph model to pose the co-clustering

problem as a graph partitioning problem. An undirected bipartite graph is a triple $G = (D, W, E)$ with D corresponding to the set of rows, W the set of columns, and E the undirected edges between rows and columns. The association of a row with a column cluster is measured by the sum of the edge-weights of this row to all columns in the column cluster, and similarly the association of a column with a row cluster can be measured. Thus each row cluster is determined by the column clustering, and in turn the row clustering determines each column cluster. The author presented a spectral algorithm to find the optimal co-clustering that corresponds to a partitioning of the graph such that the crossing edges between partitions have minimum weight.

Tanay et al. (2002) introduced SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) that combines graph theoretic and statistical considerations. The data matrix is modeled as a bipartite graph with two sets of vertices corresponding to rows and columns, and edges representing significant value changes. The authors presented two statistical models of the resulting graph, and showed how to assign weights to the vertex pairs of the bipartite graph so that heavy sub-graphs correspond to significant co-clusters. The co-cluster type is shown in Figures 1.1(h)-(i).

More co-clustering methods and applications can be found in Kluger et al. (2003), Ihmels et al. (2004), Aguilar-Ruiz & Divina (2005), Gao et al. (2005), Kung et al. (2005), Long et al. (2005), Madeira & Oliveira (2005), Pensa et al. (2005), Abdullah & Hussain

(2006), Lonardi et al. (2006), Deodhar & Ghosh (2007), Divina & Aguilar-Ruiz (2007), Yoon et al. (2007), Cai et al. (2008), Kerr et al. (2008), Puolamaki et al. (2008), and Rocci & Vichi (2008). Furthermore, Madeira & Oliveira (2004), Mechelen et al. (2004), Prelic et al. (2006), Busygin et al. (2008), and Kriegel et al. (2009) provided detailed reviews on co-clustering.

1.3. MULTI-DIMENSIONAL CLUSTERING

When researchers are interested in measurements over more than two dimensions, the data can be arranged in a multi-way contingency table with each cell being a real number. Some recent literature reflects efforts to generalize co-clustering methods to multi-dimensional contexts so that all the dimensions can be clustered simultaneously.

Bekkerman et al. (2005) extended Dhillon (2003)'s information-theoretic co-clustering to multi-dimensional clustering, and established a connection between multi-way contingency tables and undirected graphs with pairwise interaction. By treating each dimension as a random variable, the objective function to be maximized is defined as the sum of the weighted pairwise mutual information between the clustered random variables, in which the prior knowledge is incorporated by adjusting the corresponding weights. An algorithm was developed to discover the optimal multi-dimensional clustering, which interleaves conglomerative (top-down) clustering of

some variables and agglomerative (bottom-up) clustering of the other variables, with a local optimization correction routine. Taking into account that top-down clustering is efficient and bottom-up clustering leads to meaningful results, the authors argued the benefit from combining both clustering procedures.

Chiaravalloti et al. (2006) pointed out that there may not be enough knowledge to precisely set the weights in the objective function that is a linear combination of losses in the pairwise mutual information. Instead of using a pre-fixed weighting scheme, the authors introduced a notion of agreement to represent a sort of optimal “compromise” among minimizing all the losses in the pairwise mutual information. A specific data structure, called “star-structure”, is considered, in which one dimension is treated as the central dimension, and the other dimensions as the auxiliary dimensions that are pairwise independent and are all correlated with the central dimension. The authors proposed the AD-HOCC (to solve the High-Order Co-Clustering by computing Agreements for contrasting Domain objective functions) algorithm to compute the optimal agreement.

Sun et al. (2006) extended Dhillon (2003)’s information-theoretic co-clustering to the three-way contingency table and proposed cube-clustering. By using multi-information, a multivariate generalization of the mutual information between two random variables, the objective function is defined to be the loss between the multi-information of the original random variables and the multi-information of the clustered random variables, therefore

minimizing the objective function leads to the optimal cube-clustering. The authors also applied cube-clustering to the clickthrough data to improve the web search performance in a collaborative manner.

1.4. DATA VISUALIZATION

The result obtained from the hierarchical clustering of rows can be displayed in a tree structure, called dendrogram, based on which the rows of the original data matrix can be reordered. Eisen et al. (1998) introduced a graphical representation method to color the reordered data matrix as a heat map, in which large contiguous patches of color represent groups of rows that share similar patterns over columns.

To visualize the result obtained from the partitioning clustering of rows, dimension reduction techniques such as principal component analysis (PCA) and multidimensional scaling are needed to display the rows in a low dimensional space.

Pison et al. (1999) developed the CLUSPLOT package to denote rows by a set of points in a two-dimensional space, which is composed of the first principal component and the second principal component from PCA. Each cluster is then denoted by an ellipse that covers all the rows belonging to this cluster. In addition, a segment between any pair of ellipse centers can be drawn and its length designates the dissimilarity between the corresponding pair of clusters.

Rasmussen & Karypis (2004) provided a 3D mountain visualization, which is based on multidimensional scaling and produces a colored mountain-like terrain. Each cluster is denoted by a peak with the peak height being proportional to the internal cluster similarity (for example, the average pairwise similarity between rows that belong to this cluster), and the distance between a pair of peaks on the plane representing the relative dissimilarity between the corresponding pair of clusters.

To visualize the co-clustering results, Barkow et al. (2006) developed the BicAT package to display the co-clusters obtained from a number of co-clustering methods, which provides both heat map and profile visualization. In profile visualization, each row within the co-cluster is denoted by a colored line that connects the values corresponding to different columns, with columns included in the co-cluster marked with upright bars.

More data visualization methods can be found in Li (2004), Ultsch & Morchen (2005), and Zhou et al. (2008).

1.5. OPEN ISSUES

Co-clustering has proved successful in various application domains such as simultaneous clustering of genes and experimental conditions (or tissue samples) in bioinformatics, words and documents in text mining, and image or video features. Despite its success in the above domains, especially in analyzing gene expression data,

co-clustering has not found its way into medical biology applications until recent work on large-scale data sets where it has been demonstrated that it can be a very powerful tool for mining medical data (Yoon et al. 2007).

In system biology, one may be interested in not only biological variables themselves, but also the interactions between these biological variables. For example, consider a set of row variables and a set of column variables. For each pair of row and column, a number of observations may be obtained that simultaneously measure different levels of row variable and column variable, which leads to a scatter plot characterizing the relationship between them. It is of interest to cluster rows and/or columns to identify groups of individual relationships that have similar patterns because large assemblages of individual relationships with similar patterns may point toward those that have increased importance. However, when each cell of the data matrix is not represented by a single numerical value and instead contains a scatter plot, the existing clustering methods are not applicable any more. In Chapter 2 and 3, we develop both hierarchical clustering and co-clustering procedure to handle a data matrix of scatter plots. To more accurately reflect the nature of data, we introduce a dissimilarity statistic based on “data depth” to measure the discrepancy between two bivariate distributions without oversimplifying the nature of the underlying pattern. We also propose novel painting metrics and construct heat maps to allow visualization of the clusters. We demonstrate the utility and power of

our proposed clustering methods through simulation studies and application to a microbe-host-interaction study.

Another situation is that spatial information is embedded in the data matrix. In this case, the order of rows and columns of a data matrix can not be changed. None of the literature has proposed a spatial co-clustering technique that co-clusters data such that any co-cluster only contains a set of spatially consecutive rows and columns. Furthermore, there is very little literature about model-based co-clustering. In Chapter 4, we develop a co-clustering method using a Generalized Linear Mixed Model (GLMM) for spatial data. Specifically, to avoid the high computational intensity associated with global optimization, we propose a heuristic optimization algorithm to search for a near optimal co-clustering. A sampling strategy is introduced to capture as much of the spatial information that is available from the sparse data as possible. For an application pertinent to Integrated Pest Management (IPM), we combine the spatial co-clustering technique with a statistical inference method to make assessment of pest density more accurate. We demonstrate the utility and power of our proposed pest assessment procedure through simulation studies and apply the procedure to a study of the persea mite (*Oligonychus perseae*), a pest of avocado trees.

Chapter 2

Clustering Scatter Plots Using Data Depth Measures

2.1. INTRODUCTION

Microorganisms play a variety of important roles in human biology. They are involved in critical aspects of normal host (e.g., human being) physiology and development, and have been associated with a wide range of disease processes including obesity, autoimmunity, gastric ulcers and cancers (Turnbaugh et al. 2007). Despite these findings, the nature and breadth of interactions between microorganisms and humans is not well understood, and attempting to clarify these relationships is an ongoing challenge in system biology. Commonly used ordination methods such as principal component analysis (PCA) can only assess microbial and/or host variables independently for their ability to group hosts by their physiological or disease status. While canonical correlation analysis (CCA) attempts to identify relationships between microbial and host variables, its drawbacks lie in the difficulty interpreting the meaning of the results and the inherent restriction to identifying linear relationships. For our research, ordination methods are not appropriate because we are interested in identifying groups of similar associations between microbial and host variables, rather than building disease discriminators from the combined set of microbial and host variables. System biologists hold the point of

view that larger groups resulting from this process have increased importance in the sense that the constituent microbes and host variables are more likely to play important roles in the disease process.

In cluster analysis, usually data are arranged in a matrix with each cell being a real number. To avoid confusion, we call this matrix “the data matrix of scalars”. Two one-dimensional clustering methods are commonly used. For the row clustering, hierarchical clustering builds a hierarchy of clusters based on the dissimilarity measures among rows whose results can be graphically presented in a tree structure, called dendrogram. Partitioning clustering, such as K-means, divides rows into a pre-specified number of clusters in which each row belongs to the cluster with the nearest mean. One may see Andreopoulos et al. (2009) and Jiang et al. (2004) for a survey.

However, when each cell of the data matrix is not represented by a single numerical value and instead contains a scatter plot, the existing clustering methods are not applicable any more. One may think of using a single measure, say Pearson correlation coefficients, to characterize the scatter plots, which then reduces the data matrix of scatter plots to a data matrix of scalars. Current clustering methods can then be applied to analyze the associations between row variables and column variables. However, the use of Pearson correlation coefficients is not always sufficient since it is only a measure of linear association and is very sensitive to outliers. Therefore, similarity measurements

among scatter plots based on such coefficients will hinder the power of discovering clusters of scatter plots with nonlinear patterns and/or clusters with outliers.

In this chapter we introduce a hierarchical clustering procedure that is able to handle a data matrix of scatter plots. In Section 2.2, to more accurately reflect the nature of data, we introduce a dissimilarity statistic based on “data depth” to measure the discrepancy between two bivariate distributions without oversimplifying the nature of the underlying pattern. We then combine hypothesis testing with hierarchical clustering to cluster rows and columns of the data matrix of scatter plots. The power of our proposed hierarchical clustering method is demonstrated through simulation studies in Section 2.3. In Section 2.4, we propose novel painting metrics and construct heat maps to allow visualization of clusters. In Section 2.5, we apply our proposed hierarchical clustering method to a microbe-host-interaction study.

2.2. METHODOLOGY

2.2.1. Clustering Procedure

Consider M row variables $\{X_1, X_2, \cdots, X_M\}$ and N column variables $\{Y_1, Y_2, \cdots, Y_N\}$. For each pair of row variable and column variable, a random sample of observations are taken that can be drawn as a scatter plot in the Cartesian plane as shown

in Figure 2.1, in which each square contains a scatter plot. Our goal is to cluster both rows and columns based on these $M \times N$ independent scatter plots.

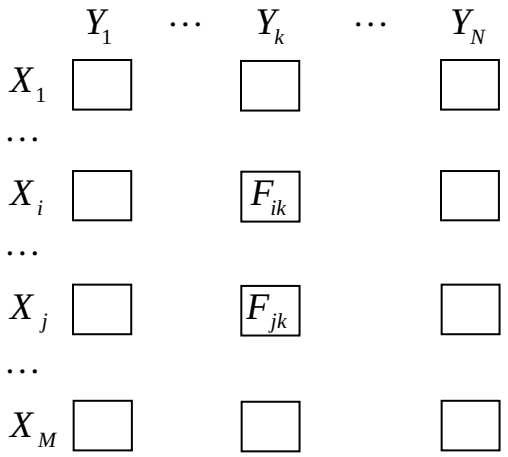


Figure 2.1. Data Structure: A Data Matrix of Scatter Plots

To obtain the distance matrix for performing the row hierarchical clustering, we have to calculate the distance between any two rows. Consider the i^{th} row and the j^{th} row, we would like to measure how similar these two rows are to each other based on comparing the corresponding N pairs of scatter plots. For each column, say the k^{th} column, the pair of scatter plots can be thought of as the samples taken from two independent bivariate distributions F_{ik} and F_{jk}, respectively, as shown in Figure 2.1. As a result, the problem of comparing the pair of scatter plots can be formulated as testing the following hypothesis:

$$H_0 : F_{ik} = F_{jk} \text{ vs. } H_a : F_{ik} \neq F_{jk}. \tag{2.1}$$

Denote by $p_{ij(k)}$ the p-value for testing the above hypothesis. The smaller the p-value, the less similar the pair of scatter plots to each other. By testing the same kind of hypotheses for all the N columns, we define the dissimilarity (distance) between the i^{th} row and the j^{th} row as

$$dist_{ij} = \sum_{k=1}^{N} \left(1 - p_{ij(k)}\right). \tag{2.2}$$

Then the distance matrix for rows $\{dist_{ij}\}$ $(i, j = 1, 2, \cdots, M,$ and $i \neq j)$ is inputted to the regular hierarchical clustering algorithm, which initially regards each row as an individual cluster, and at each step, merges the closest pair of clusters until all the rows are merged into one cluster. In doing this, hierarchical clustering creates a hierarchy of row clusters that can be represented in a tree structure called dendrogram.

The same clustering procedure can be applied to columns as well. Therefore, rows and columns in the original data matrix of scatter plots (Figure 2.1) are reordered according to the row dendrogram and the column dendrogram, respectively, which produces a new data matrix of scatter plots that acts as the output of our proposed clustering procedure.

2.2.2 Hypothesis Testing

Liu & Singh (1993) proposed a multivariate rank sum test for the hypothesis $H_0 : F_{ik} = F_{jk}$ vs. $H_a : F_{ik} \neq F_{jk}$ where F_{ik} and F_{jk} are the distribution functions of

two independent populations. Specifically, the test statistic is based on a quality index that measures the overall "outlyingness" of population F_{jk} relative to population F_{ik},

$$Q(F_{ik}, F_{jk}) = P\left(D(F_{ik};\vec{U}) \le D(F_{ik};\vec{V}) \,|\, \vec{U} \sim F_{ik}, \vec{V} \sim F_{jk}\right), \tag{2.3}$$

where $D(F_{ik};\cdot)$ is an affine-invariant data depth function with respect to F_{ik} that could be Mahalanobis depth, Tukey (Half-space) depth, and Simplicial depth, etc., as shown in Section 2.2.3.

Given two samples $\{\vec{U}_1,\cdots,\vec{U}_S\}$ from F_{ik} and $\{\vec{V}_1,\cdots,\vec{V}_T\}$ from F_{jk}, $Q(F_{ik}, F_{jk})$ can be estimated by

$$Q(F_{ik}^S, F_{jk}^T) = \frac{1}{T}\sum_{t=1}^{T} R(F_{ik}^S;\vec{V}_t), \tag{2.4}$$

where F_{ik}^S and F_{jk}^T are the empirical distributions, $R(F_{ik}^S;\vec{V}_t)$ is the proportion of $\vec{U}_s$'s with $D(F_{ik}^S;\vec{U}_s) \le D(F_{ik}^S;\vec{V}_t)$, and $D(F_{ik}^S;\cdot)$ is the empirical data depth with respect to F_{ik}^S. From Liu & Singh (1993) and Zuo & He (2006), we have

$$Q(F_{ik}^S, F_{jk}^T) - 1/2 \sim AN(0, (1/S + 1/T)/12) \tag{2.5}$$

under $H_0 : F_{ik} = F_{jk}$ for many commonly used data depth functions (under general regularity conditions).

Notice that the overall "outlyingness" of F_{ik} relative to F_{jk} can be also measured by a quality index

$$Q(F_{jk}, F_{ik}) = P\left(D(F_{jk};\vec{V}) \le D(F_{jk};\vec{U}) \,|\, \vec{V} \sim F_{jk}, \vec{U} \sim F_{ik}\right), \tag{2.6}$$

where $D(F_{jk};\cdot)$ is an affine-invariant data depth function with respect to F_{jk}. Likewise, $Q(F_{jk},F_{ik})$ may be estimated by

$$Q(F_{jk}^T,F_{ik}^S)=\frac{1}{S}\sum_{s=1}^{S}R(F_{jk}^T;\vec{U}_s), \tag{2.7}$$

where $R(F_{jk}^T;\vec{U}_s)$ is the proportion of $\vec{V}_t$'s with $D(F_{jk}^T;\vec{V}_t)\le D(F_{jk}^T;\vec{U}_s)$, and $D(F_{jk}^T;\cdot)$ is the empirical data depth with respect to F_{jk}^T.

As Section 2.2.4 shows, $Q(F_{jk},F_{ik})$ is not directly related to $Q(F_{ik},F_{jk})$. However, to obtain the p-value for testing hypothesis (2.1), we would like to have a unique parameter to measure the difference between two distributions, either comparing F_{ik} to F_{jk}, or F_{jk} to F_{ik}. Under $H_0:F_{ik}=F_{jk}$, $Q(F_{ik},F_{jk})=Q(F_{jk},F_{ik})=1/2$. With the location shift and/or scale change between F_{ik} and F_{jk}, either $Q(F_{ik},F_{jk})$ or $Q(F_{jk},F_{ik})$, or both, would deviate from 1/2 significantly. Therefore, to avoid having one distribution as the reference distribution, we propose a new quality index, called TS, to measure the overall "difference" between F_{ik} and F_{jk},

$$TS=\begin{cases}Q(F_{ik},F_{jk}),\text{if } \left|Q(F_{ik},F_{jk})-1/2\right|>\left|Q(F_{jk},F_{ik})-1/2\right|;\\ Q(F_{jk},F_{ik}),\text{if } \left|Q(F_{ik},F_{jk})-1/2\right|<\left|Q(F_{jk},F_{ik})-1/2\right|.\end{cases} \tag{2.8}$$

The test statistic for testing hypothesis (2.1) is the estimate of TS,

$$\widehat{TS}=\begin{cases}Q(F_{ik}^S,F_{jk}^T),\text{if } \left|Q(F_{ik}^S,F_{jk}^T)-1/2\right|>\left|Q(F_{jk}^T,F_{ik}^S)-1/2\right|;\\ Q(F_{jk}^T,F_{ik}^S),\text{if } \left|Q(F_{ik}^S,F_{jk}^T)-1/2\right|<\left|Q(F_{jk}^T,F_{ik}^S)-1/2\right|.\end{cases} \tag{2.9}$$

Then $p_{ij(k)}$ is calculated by the following permutation test procedure:

1) Pool two samples $\{\vec{U}_1,\cdots,\vec{U}_S\}$ and $\{\vec{V}_1,\cdots,\vec{V}_T\}$.
2) Take a sample of size S without replacement $\{\vec{U}^*_1,\cdots,\vec{U}^*_S\}$ from the pooled sample, and the remaining is $\{\vec{V}^*_1,\cdots,\vec{V}^*_T\}$, which are called two permutation samples.
3) Estimate $Q(F_{ik},F_{jk})$ and $Q(F_{jk},F_{ik})$ by $Q^*(F^S_{ik},F^T_{jk})$ and $Q^*(F^T_{jk},F^S_{ik})$, respectively, based on the permutation samples obtained in Step 2.
4) Set $\widehat{TS}^*$ to be equal to $Q^*(F^S_{ik},F^T_{jk})$ if $\left|Q^*(F^S_{ik},F^T_{jk})-1/2\right|>\left|Q^*(F^T_{jk},F^S_{ik})-1/2\right|$; and equal to $Q^*(F^T_{jk},F^S_{ik})$ otherwise.
5) Repeat the above steps (Step 2 - Step 4) B times to yield B values of $\widehat{TS}^*$, denoted by $\widehat{TS}^*_b$ ($b=1,2,\cdots B$), whose distribution estimates the sampling distribution of the test statistic $\widehat{TS}$ under $H_0: F_{ik}=F_{jk}$.
6) Let p_{lower} be the proportion of $\{\widehat{TS}^*_b\}_{b=1}^B$ with $\widehat{TS}^*_b<\widehat{TS}$, and p_{upper} the proportion of $\{\widehat{TS}^*_b\}_{b=1}^B$ with $\widehat{TS}^*_b>\widehat{TS}$. Hence $p_{ij(k)}=2\times min(p_{lower},p_{upper})$.

2.2.3. Data Depth

Let F be a probability distribution in $\Re^p$ with $p\geq 1$ and $\vec{x}$ a point in $\Re^p$. The data depth at $\vec{x}$ with respect to F is denoted by $D(F;\vec{x})$, which measures how deep (or central) the point $\vec{x}$ is with respect to F. The larger $D(F;\vec{x})$, the deeper (or more

central) the point $\vec{x}$ with respect to F. Some commonly used data depth functions are listed as follows.

1) Mahalanobis depth (Mahalanobis 1936):

$$M_hD(F;\vec{x}) = 1/[1+(\vec{x}-\vec{\mu}_F)'\sum_F^{-1}(\vec{x}-\vec{\mu}_F)],$$

where $\vec{\mu}_F$ and $\sum_F$ are the mean and variance-covariance matrix of F, respectively. The sample version of $M_hD(F;\vec{x})$ is obtained by replacing $\vec{\mu}_F$ and $\sum_F$ with their sample estimates.

2) Tukey depth / Half-space depth (Tukey 1974):

$$TD(F;\vec{x}) = \inf_{\mathrm{H}}\{P_F(\mathrm{H}) : \mathrm{H} \text{ is a closed half-space in } \Re^p \text{ containing } \vec{x}\}.$$

The sample version of $TD(F;\vec{x})$ is $TD(F_n;\vec{x})$ where F_n is the empirical distribution.

3) Simplicial depth (Liu 1990):

$$SD(F;\vec{x}) = P_F(\vec{x} \text{ is inside the closed simplex whose vertices are } \{\vec{X}_1,\cdots,\vec{X}_{p+1}\}),$$

where $\{\vec{X}_1,\cdots,\vec{X}_{p+1}\}$ is a random sample from F. The sample version of $SD(F;\vec{x})$ is the fraction of the sample random simplexes containing the point $\vec{x}$.

It is easy to compute Mahalanobis depth that studies the elliptical structure of a multivariate distribution. Rousseeuw & Ruts (1996) addressed the computation issues for Tukey depth and Simplicial depth that are more robust than Mahalanobis depth. More data depths can be found in Liu et al. (1999).

2.2.4. $Q(F,G)$ **vs.** $Q(G,F)$

Consider two independent distributions F and G, and two variables $X \sim F$ and $Y \sim G$. We present three examples to show the relationship between $Q(F,G)$ and $Q(G,F)$. For simplicity, univariate normal distributions and Mahalanobis depth are adopted here.

Example 1: For $F = N(\mu_0, \sigma_0^2)$, $G = N(\mu_0, \sigma_1^2)$, and $\sigma_1^2 > \sigma_0^2$, we have

$$Q(F,G) = P((X-\mu_0)^2/\sigma_0^2 \geq (Y-\mu_0)^2/\sigma_0^2) < 1/2,$$

$$Q(G,F) = P((Y-\mu_0)^2/\sigma_1^2 \geq (X-\mu_0)^2/\sigma_1^2) > 1/2,$$

$$\text{and} \quad Q(F,G) + Q(G,F) = 1.$$

Example 2: For $F = N(\mu_0, \sigma_0^2)$, $G = N(\mu_1, \sigma_0^2)$, and $\mu_0 \neq \mu_1$, we have

$$Q(F,G) = P((X-\mu_0)^2/\sigma_0^2 \geq (Y-\mu_0)^2/\sigma_0^2) < 1/2,$$

$$Q(G,F) = P((Y-\mu_1)^2/\sigma_0^2 \geq (X-\mu_1)^2/\sigma_0^2) < 1/2,$$

$$\text{and} \quad Q(F,G) = Q(G,F).$$

Example 3: For $F = N(\mu_0, \sigma_0^2)$, $G = N(\mu_1, \sigma_1^2)$, $\mu_0 \neq \mu_1$, and $\sigma_1^2 > \sigma_0^2$, we have

$$Q(F,G) = P((X-\mu_0)^2/\sigma_0^2 \geq (Y-\mu_0)^2/\sigma_0^2) < 1/2,$$

$$\text{and } Q(G,F) = P((Y-\mu_1)^2/\sigma_1^2 \geq (X-\mu_1)^2/\sigma_1^2) \begin{cases} < 1/2, \\ = 1/2, \text{ or} \\ > 1/2. \end{cases}$$

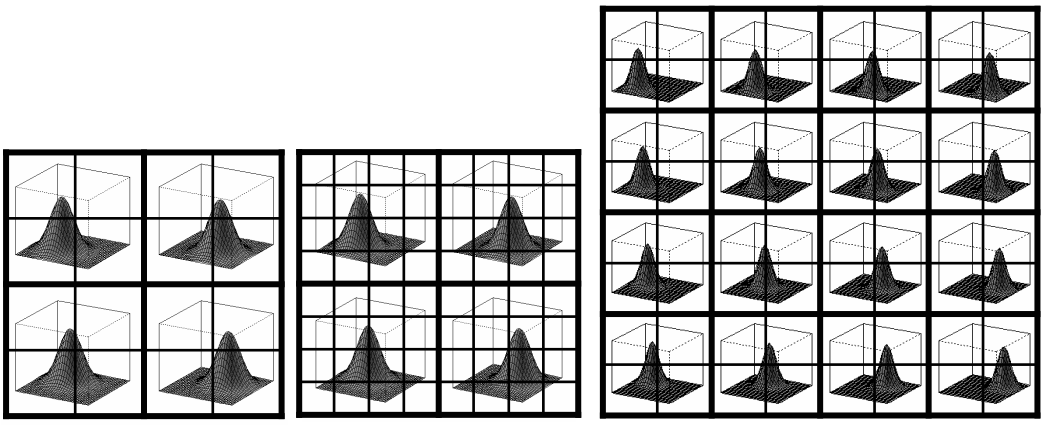
2.3. SIMULATION STUDY

We performed a simulation study to investigate the power of our proposed clustering method. The basic procedure is as follows:

1) Specify a “checkerboard” data pattern with a set of row clusters and column clusters, and specify a bivariate distribution for the cells within each block.
2) Generate random samples based on the given bivariate distributions in Step 1, which creates a data matrix of scatter plots.
3) Apply our proposed clustering method to this data matrix of scatter plots, and check whether the original data pattern can be retrieved or not. That is, we check whether rows within the same block are still close to each other compared to other rows in the row dendrogram, and columns as well; or equivalently, whether there exists a cutting of row dendrogram such that the generated branch set are exactly same as the original set of row clusters, and columns as well;
4) Repeat Step 2 - Step 3 a number of times, and record the success rate, the proportion of times that we succeed in retrieving the original data pattern, which acts as the power measurement for our proposed clustering method.

Intuitively, the total number of rows and columns (the size of the data matrix of scatter plots, or the data size), the number of rows and columns within each block (the

block size), and the number of blocks would affect the success rate. Therefore, we considered three data pattern settings shown in Figure 2.2.



(a) “R2C2” (b) “2*R2C2” (c) “R4C4”

Figure 2.2. Three Data Pattern Settings: (a) “R2C2”: there are 2×2 blocks (2 row clusters and 2 column clusters), each of which contains 2×2 cells, thus the data size is 4×4. (b) “2*R2C2”: the block size is doubled in the “R2C2” setting, thus the data size is 8×8. (c) “R4C4”: there are 4×4 blocks (4 row clusters and 4 column clusters), each of which contains 2×2 cells, thus the data size is 8×8.

For each setting, we specified a class of bivariate normal distributions for blocks, which only differ in location. Specifically, the $x-$coordinate of the mean increases equidistantly along the row direction ranging from 0 with the $y-$coordinate of the mean remaining same; whereas the $y-$coordinate of the mean increases equidistantly along the column direction ranging from 0 with the $x-$coordinate of the mean remaining same. For example, with a location shift of 1, the mean of the top left bivariate

normal distribution in “R2C2” and “2*R2C2” is $\begin{pmatrix}0\\0\end{pmatrix}$, the top right $\begin{pmatrix}1\\0\end{pmatrix}$, the bottom left $\begin{pmatrix}0\\1\end{pmatrix}$, and the bottom right $\begin{pmatrix}1\\1\end{pmatrix}$. Furthermore, 50 data points were generated for each scatter plot, Mahalanobis depth was adopted, 500 resampling times were taken for the permutation test, and the average linkage method was chosen for the hierarchical clustering procedure. We performed 500 simulations for each setting. The relationship between the success rate and the location shift is summarized in Figure 2.3, where the solid lines stand for the variance-covariance matrix $\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}$ specified for the bivariate normal distributions, and the dashed lines for $\begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}$ with the correlation coefficient $\rho = 0.5$.

From Figure 2.3, we may observe the following:

1) By comparing the solid line with the dashed line for each setting, the correlation in the bivariate normal distribution improves the success rate.
2) By comparing “R2C2” with “2*R2C2” both having a fixed number of blocks, with a relatively large location shift, the larger the block size, the higher the success rate; with a relatively small location shift, the smaller the block size, the higher the success rate. That is, more scatter plots with larger distance between blocks improves the

chance of capturing the pattern. However, more scatter plots with smaller distance between blocks introduces a higher chance for noise in the clustering.

3) By comparing “2*R2C2” with “R4C4” both having a fixed data size, the smaller the number of blocks, the higher the success rate, which means it is harder to do a more delicate job (more row clusters and column clusters).

4) By comparing “R2C2” with “R4C4” both having a fixed block size, with a relatively small location shift, the smaller the number of blocks, the higher the success rate; with a relatively large location shift, the larger the number of blocks, the higher the success rate. The reason is similar to what we previously discussed in the comparison of “R2C2” with “2*R2C2”.

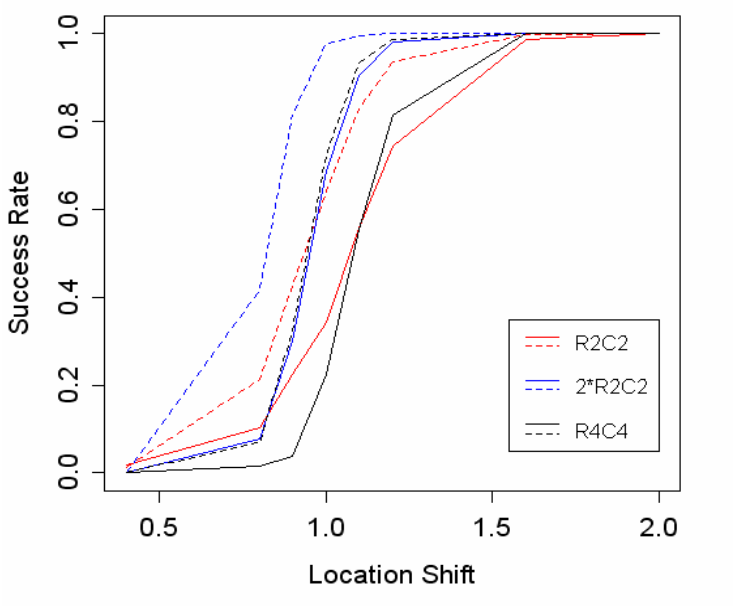


Figure 2.3. Success Rate versus Location Shift

2.4. DATA VISUALIZATION

Data visualization is an important aspect in the clustering technique. In the traditional hierarchical clustering application in which cells of a data matrix are scalars, the original data can be rearranged according to the dissimilarity scores between rows (or columns). The smaller the dissimilarity score between two rows (or columns), the closer the two rows (or columns). A graphical representation of the rearranged data matrix, called heat map, can be created where cells are painted with different colors based on their scalar values. Obviously, we would expect cells in close proximity to each other to have a similar color.

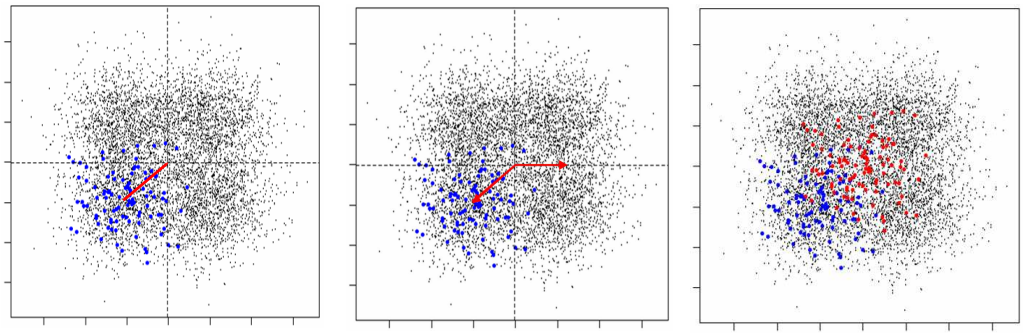
To apply the above painting strategy to a data matrix of scatter plots, we introduce a painting metric, called Overall Quality Index (OQI), to graphically represent the scatter plots so that similar scatter plots are painted with a similar color whereas dissimilar scatter plots are painted with different colors. All the MN scatter plots are pooled as a single scatter plot that is thought of as a sample from the bivariate distribution F_{pool}. Consider a scatter plot that is regarded as a sample from the bivariate distribution F. The OQI value is the estimated value of the quality index $Q(F_{pool}, F)$. However, as we discuss below, it is unlikely that using a single painting metric for the scatter plot is sufficient. Therefore, we introduce three additional finer painting metrics to further characterize the scatter plot.

1) Center Deviation Index (CDI): For any scatter plot, we define its center to be the point that maximizes the empirical data depth for $D(F;\cdot)$. Then the CDI for a scatter plot is the distance between its center and the center of the pooled scatter plot. For example, in Figure 2.4(a), the length of red segment is the CDI measuring the deviation of the scatter plot consisting of blue points from the pooled scatter plot consisting of black points.

2) Center Deviation Direction Index (CDDI): By taking the center of the pooled scatter plot as the origin of a new Cartesian coordinate system, the CDDI for a scatter plot is the magnitude of the angle formed by the vector from the origin to its center and the positive $x-\text{axis}$, which ranges from $-\pi$ to π. The CDDI depicts the relative location of a scatter plot with respect to the pooled scatter plot, and then the relative locations among the scatter plots. For example, in Figure 2.4(b), the CDDI for the blue scatter plot is the degree of the angle formed by two red vectors.

3) Dispersion Index (DI): Moving a scatter plot such that its center and the center of the pooled scatter plot coincide produces a shifted scatter plot that is regarded as a sample from a new bivariate distribution F'. The DI for the original scatter plot is the estimation of the quality index $Q(F_{pool}, F')$, which accounts for the difference between the original scatter plot and the pooled scatter plot excluding the effect due to the location shift. For example, in Figure 2.4(c), the DI for the blue scatter plot is

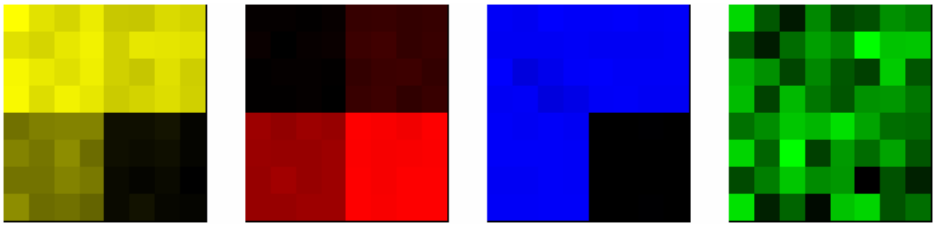
the estimated quality index of the red scatter plot (obtained from moving the blue scatter plot) with respect to the pooled scatter plot.



(a) CDI (b) CDDI (c) DI

Figure 2.4. Painting Metrics

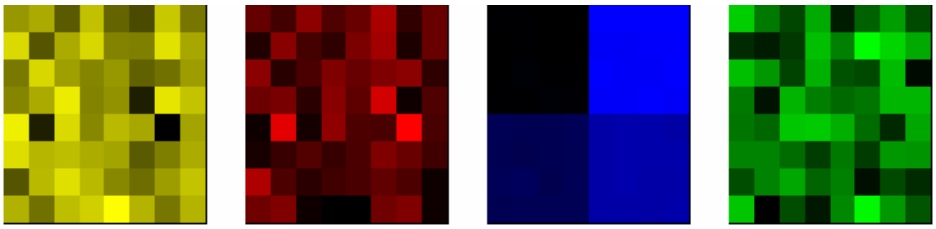
$N^{(2)}\left(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}0\\3\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$
$N^{(2)}\left(\begin{pmatrix}0\\7\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}0\\-11\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 2.5. Painting Example 1: OQI can reveal clusters. Also, the bivariate normal distributions only differ by location and are asymmetric about the origin, therefore CDI can reveal clusters whereas CDDI and DI can not.

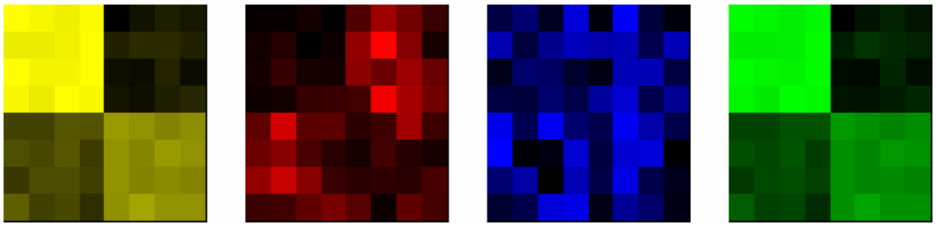
$N^{(2)}\left(\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}-2\\2\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$
$N^{(2)}\left(\begin{pmatrix}2\\-2\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}2&0\\0&2\end{pmatrix}\right)$



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 2.6. Painting Example 2: OQI can not reveal clusters. Also, the bivariate normal distributions only differ by location and are symmetric about the origin, therefore CDDI can reveal clusters whereas CDI and DI can not.

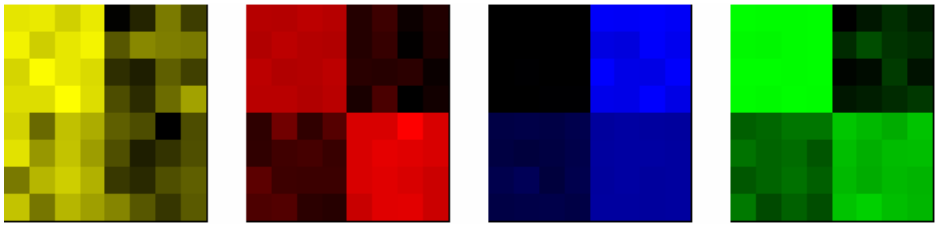
$N^{(2)}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}16&0\\0&16\end{pmatrix}\right)$
$N^{(2)}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}9&0\\0&9\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}4&0\\0&4\end{pmatrix}\right)$



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 2.7. Painting Example 3: OQI can reveal clusters. Also, the bivariate normal distributions only differ by scale, therefore DI can reveal clusters whereas CDI and CDDI can not.

$N^{(2)}\left(\begin{pmatrix}-3\\-3\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}-1\\1\end{pmatrix},\begin{pmatrix}16&0\\0&16\end{pmatrix}\right)$
$N^{(2)}\left(\begin{pmatrix}2\\-2\end{pmatrix},\begin{pmatrix}9&0\\0&9\end{pmatrix}\right)$	$N^{(2)}\left(\begin{pmatrix}4\\4\end{pmatrix},\begin{pmatrix}4&0\\0&4\end{pmatrix}\right)$



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 2.8. Painting Example 4: OQI has poor performance to reveal clusters. Also, the bivariate normal distributions differ by both location and scale and asymmetric about the origin, therefore CDI, CDDI and DI can all reveal clusters.

To illustrate the utility of the above painting metrics, we present four painting examples. In each example, an 8×8 matrix of scatter plots (each scatter plot contains 100 data points) was generated with the top left 4×4, top right 4×4, bottom left 4×4 and bottom right 4×4 scatter plots following the four different distributions specified in the top panel of Figures 2.5-2.8. We then obtained 8×8 matrices of OQI, CDI, CDDI and DI, based upon which four heat maps can be generated as shown in the bottom panel of Figures 2.5-2.8, where the “yellow” heat map is based on OQI, the “red” heat map on CDI, the “blue” heat map on CDDI, the “green” heat map on DI, and the “black” color stands for the minimum index value in all the four heat maps. For simplicity, we used Mahalanobis depth in all the examples discussed here.

The painting examples illustrate that OQI captures the overall effect due to both location shift and scale change of a scatter plot. When OQI can distinguishes two scatter plots from each other, one may further investigate CDI, CDDI, and DI to see the details of how these two scatter plots differ. Also, when OQI can not distinguish two scatter plots from each other, one may want to see if any of CDI, CDDI, and DI can distinguish them.

2.5. APPLICATION

2.5.1. Motivation

Identifying causative microbial and host variables in multi-factorial diseases remains a considerable challenge. For example, consider the case of inflammatory bowel disease (IBD). IBD etiology appears to involve several factors, including genetics, lifestyle and intestinal bacteria. Traditionally, investigations attempting to identify variables associated with complex diseases such as IBD have used ordination methods such as principle component analysis (PCA) to define host phenotypes by levels of the microorganisms and/or host variables (proteome, transcriptome, etc). A shortcoming of this approach is that it does not account for upstream events such as the physical or chemical interactions between the microorganisms and the host, nor the cascade of events that likely connect these interactions to host phenotype. Here, we describe an alternative experimental approach that begins to address this shortcoming, which is to first analyze upstream

events – the physical or chemical interactions between the microorganisms and the host, which are represented by the relationships in the scatter plots – and then assess if and how those relationships (and/or the variables involved in those relationships) are linked to host phenotype. More biological hypothesis will be discussed in Chapter 3.

2.5.2. Results

To identify putatively important microbe-host interactions, Li et al. (2010) recently examined the amounts of bacteria and proteins in mucosal luminal interface samples from IBD and healthy subjects. Two datasets were generated from the experiment. “Microbe” data were arranged as a data matrix with 81 rows (3 rows containing missing values are excluded) standing for samples, 15 columns for microbes, and each cell being a single numerical value recording the level of a microbe in a sample. “Protein” data were also arranged as a data matrix with 81 rows standing for the same set of samples, 440 columns for proteins, and each cell being a single numerical value recording the level of a protein in a sample. To identify associations between levels of the microbes and proteins, we combined the above two data matrices of scalars by pairing up the columns (one from “Microbe” data, the other from “Protein” data) and treating each 81×2 array of data as bivariate data with the $x-$axis being microbe level and the $y-$axis being protein

level. This process leads to a data matrix of scatter plots as shown in Figure 2.1 where $M = 440$, $N = 15$, and each scatter plot contains 81 data points.

Considering the scatter plots as independent samples, we applied our proposed clustering method to the 440×15 data matrix of scatter plots, and cluster both proteins (rows) and microbes (columns). We used Mahalanobis depth as the data depth measure, $B = 500$ resampling times for the permutation test, and the average linkage method to perform the hierarchical clustering.

We then cut the "Protein" dendrogram at the height of 6, which generates eighty protein branches/clusters. The proteins within the same branch are more similar to each other, or show more similar microbe-protein patterns, than those in other branches. From the eighty protein clusters, we only selected those containing at least twenty proteins, which leads to five protein clusters. We also generated four microbe clusters by cutting the "Microbe" dendrogram at the height of 430, and selected those containing at least five microbes. One pair of the selected protein cluster and microbe cluster is depicted in Figure 2.9, where the heat map with the OQI painting metric is shown. The promise of these results is demonstrated by the fact that most of the identified proteins have been previously associated with IBD as in Ahrenstedt et al. (1990), Broedl et al. (2007), Foell et al. (2003), Greenstein et al. (1992), Hansen et al. (2009), and Larsson et al. (2006).

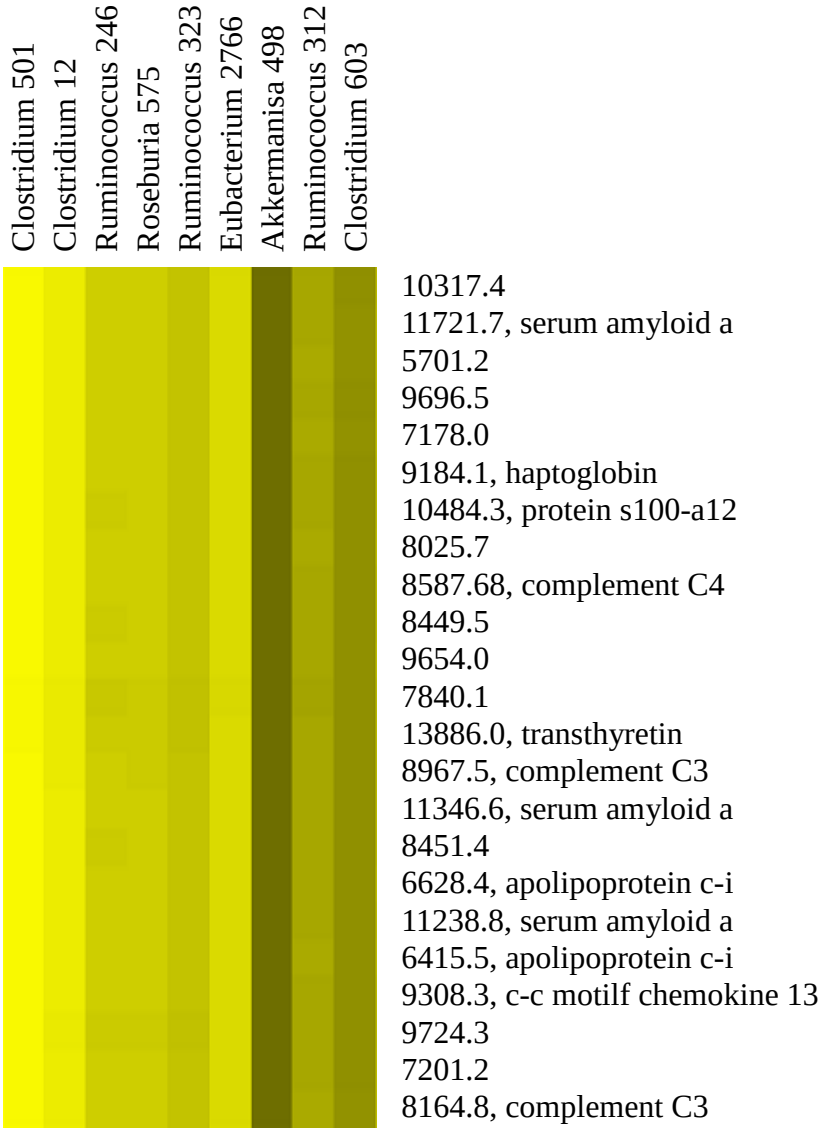


Figure 2.9. Heat Map with the OQI Painting Metric

Examining such relationships will have utility for several purposes. First, by clustering relationships of various microbial and host variables, one can identify groups of relationships that have similar and/or dissimilar associations by visually examining the

heat maps. Large assemblages of individual relationships with similar associations may point toward those that have increased importance, because they indicate organisms having a greater impact on the host, or vice versa. Assemblages with similar associations might also be used to identify different taxa with similar functions as well as direct decisions concerning which of the myriad of unidentified variables should be examined further. This latter feature addresses the nature of data generated in this “omics era,” where most of the variables cannot be identified by simple database searches, but instead require procedures consuming considerable amounts of time and effort. Lastly, dissimilar relationships could provide key information, for example, in identifying relationships between host defense molecules and the bacteria they target.

2.6. CONCLUSION

Our proposed method showed a significant utility and power in handling a data matrix of scatter plots. More importantly, this clustering procedure can be easily extended to the high dimensional case when one or more sets of variables needs to be analyzed. Moreover, the novel painting metrics we proposed can be easily extended to multi-dimensional clusters of multivariate plots.

Co-clustering is desirable over traditional one-dimensional clustering as it is more informative and easily interpretable while preserving most of the information contained

in the original data; and it allows dimension reduction along both axes simultaneously and hence leads to a much more compact representation of the original data for subsequent analysis. In Chapter 3, we will develop a co-clustering method to deal with a data matrix of scatter plots.

Finally, although these methods were developed to analyze microbe-host interactions, we anticipate that this general approach will have utility for a wide range of investigations, including those examining relationships among gene expression profiles, metabolites, genes and epigenetic parameters.

Chapter 3

Co-clustering Scatter Plots Using Data Depth Measures

3.1. INTRODUCTION

Co-clustering, also called biclustering, bivariate clustering, or two-mode clustering, has been broadly applied to many domains such as bioinformatics and text mining. Different from the one-dimensional clustering methods that seek to identify similar rows or columns independently, co-clustering simultaneously clusters rows and columns to identify “blocks” (or “co-clusters”) of rows and columns that show highly inter-related coherence. For example, in gene expression analysis, co-clustering can be used to solve the dual problem of identifying a set of genes and conditions simultaneously involved in a metabolic process, a problem that traditional one-dimensional clustering methods can not handle. Madeira & Oliveira (2004), Mechelen et al. (2004), Prelic et al. (2006), and Busygin et al. (2008) provided detailed reviews on co-clustering.

In this chapter we introduce a co-clustering procedure that is able to handle a data matrix of scatter plots. In Section 3.2, to more accurately reflect the nature of data, we introduce a dissimilarity statistic based on “data depth” to measure the discrepancy between two bivariate distributions without oversimplifying the nature of the underlying pattern. We then combine hypothesis testing with a searching algorithm to simultaneously

cluster rows and columns of the data matrix of scatter plots. The power of our proposed co-clustering method is demonstrated through simulation studies in Section 3.3. In Section 3.4, we propose novel painting metrics and construct heat maps to allow visualization of co-clusters. In Section 3.5, we apply our proposed co-clustering method to a microbe-host-interaction study.

3.2. METHODOLOGY

3.2.1. Co-clustering Procedure

Consider M row variables $\{X_1, X_2, \cdots, X_M\}$ and N column variables $\{Y_1, Y_2, \cdots, Y_N\}$. For each pair of row variable and column variable, a random sample of observations are taken that can be drawn as a scatter plot in the Cartesian plane as shown in Figure 3.1, in which each square contains a scatter plot. By regarding a scatter plot as a sample from a bivariate distribution, a co-cluster is defined to be the union of a subset of row variables and a subset of column variables, $\{X_{i_1}, X_{i_2}, \cdots, X_{i_r}\} \bigcup \{Y_{j_1}, Y_{j_2}, \cdots, Y_{j_c}\}$ with $\{i_1, \cdots, i_r\} \subset \{1, \cdots, M\}$ and $\{j_1, \cdots, j_c\} \subset \{1, \cdots, N\}$, within which each pair of row variable and column variable follows the common bivariate distribution. Our goal is to identify all the co-clusters based on these $M \times N$ independent scatter plots.

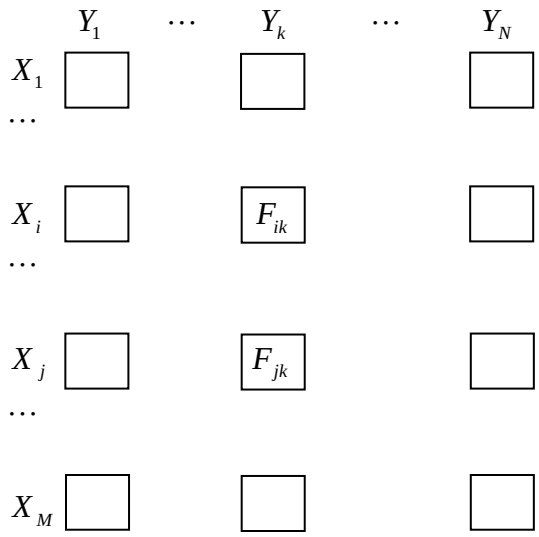


Figure 3.1. Data Structure: A Data Matrix of Scatter Plots

In many cases, one would not expect all the rows or columns being investigated to be involved in the obtained co-clusters since some rows or columns may not share the common pattern with other rows or columns and therefore do not belong to any co-cluster. Also, some rows or columns may belong to two or more co-clusters simultaneously. Moreover, one would allow co-clusters to be overlapping, which means scatter plots may belong to two or more different co-clusters simultaneously. We next propose a co-clustering procedure to identify all the hidden co-clusters that satisfy the above properties. Notice that a single scatter plot itself may be a co-cluster. To avoid this, we specify the minimum co-cluster size to be $r_0 \times c_0$, in which r_0 is the minimum number of rows and c_0 the minimum number of columns.

Starting with one column, say the k^{th} column, we apply the one-dimensional hierarchical row clustering (Zhang et al. 2010) to the corresponding M scatter plots, for which we have to calculate the distance between any pair of scatter plots as a measure of how similar they are to each other. Consider the i^{th} scatter plot (row) and the j^{th} scatter plot (row) that can be thought of as the samples taken from two independent bivariate distributions F_{ik} and F_{jk}, respectively. The problem of comparing the pair of scatter plots is then formulated as testing the following hypothesis:

$$H_0 : F_{ik} = F_{jk} \quad \text{vs.} \quad H_a : F_{ik} \neq F_{jk}. \tag{3.1}$$

Denote by $p_{ij(k)}$ the p-value for testing the above hypothesis. The smaller the p-value, the less similar the pair of scatter plots to each other. We define the dissimilarity (distance) between the i^{th} scatter plot (row) and the j^{th} scatter plot (row) as

$$dist_{ij(k)} = 1 - p_{ij(k)}. \tag{3.2}$$

Then the distance matrix $\{dist_{ij(k)}\}$ ($i, j = 1, 2, \cdots, M$, and $i \neq j$) is inputted to the regular hierarchical clustering algorithm, which initially regards each scatter plot as an individual cluster, and at each step, merges the closest pair of clusters until all the scatter plots are merged into one cluster. In doing this, hierarchical clustering creates a hierarchy of clusters that can be represented in a tree structure called dendrogram. We cut the dendrogram at a prespecified height δ_{seed} and select those branches containing at least r_0 scatter plots.

The scatter plots within a particular branch are more similar to each other than the scatter plots between different branches. Specifically, if the complete linkage method was used when constructing the dendrogram, then the p-value for testing any pair of scatter plots within that branch is greater than $1-\delta_{seed}$. For example, suppose we cut the dendrogram at the height $\delta_{seed}=0.95$. Within each branch, the maximum distance between any pair of scatter plots is less than 0.95. Hence, the minimum p-value for testing any pair of scatter plots is greater than $1-\delta_{seed}=0.05$.

Defining the subset of row variables corresponding to a selected branch as a "seed", which acts as the row part in a potential co-cluster, we next identify the column part of this co-cluster. For each seed, say $\{X_{i_1}, X_{i_2}, \cdots, X_{i_r}\}$ of size r with $\{i_1, i_2, \cdots, i_r\} \subset \{1, 2, \cdots, M\}$, we pool all the data points contained in the corresponding r scatter plots, which can be regarded as the sample from a common bivariate distribution F_{seed}. Suppose all the other $N-1$ columns are potentially included in the co-cluster. Moving to another column, say the $(k')^{th}$ column, we test the hypothesis

$$H_0 : F_{i'k'} = F_{seed} \quad \text{vs.} \quad H_a : F_{i'k'} \neq F_{seed} \tag{3.3}$$

for each $i' \in \{i_1, i_2, \cdots, i_r\}$, where $F_{i'k'}$ is the distribution that the scatter plot for the $(i')^{th}$ row and the $(k')^{th}$ column follows. The $(k')^{th}$ column will be excluded from the potential co-cluster if any of the above r null hypotheses is rejected (using a suitable multiple comparisons adjusted test procedure as discussed in Section 3.2.2). We continue

this exclusion process to reduce the column part of the potential co-cluster until we finish scanning all the columns. The resulting co-cluster will be reported if it contains at least c_0 column variables. This may be implemented for $k = 1, 2, \cdots, N$. Therefore, starting with each of the original columns, we can identify all the co-clusters hidden in the data matrix of scatter plots.

We illustrate the proposed co-clustering procedure by presenting an example with a set of row variables $\{X_1, X_2, \cdots, X_{10}\}$ and column variables $\{Y_1, Y_2, \cdots, Y_8\}$ as shown in Figure 3.2, in which each square represents a scatter plot. The minimum co-cluster size is set to be 3×3. Starting with the column Y_1, we apply the hierarchical row clustering to the corresponding 10 scatter plots across the rows. Two seeds are obtained by cutting the dendrogram at δ_{seed}, $\{X_2, X_3, X_4, X_5\}$ denoted by the blue solid line and $\{X_7, X_8, X_9\}$ by the red solid line. For the seed $\{X_2, X_3, X_4, X_5\}$, we pool all the observations contained in the four scatter plots, (X_2, Y_1), (X_3, Y_1), (X_4, Y_1), and (X_5, Y_1), which leads to a pooled scatter plot. Moving to another column, say Y_2, we compare each of the four scatter plots corresponding to the seed, (X_2, Y_2), (X_3, Y_2), (X_4, Y_2), and (X_5, Y_2), with the pooled scatter plot. The column Y_2 will be excluded from the potential co-cluster if any of the above scatter plots does not share the same bivariate distribution as the pooled scatter plot. After scanning all the columns, we finish building the potential co-cluster and check if it satisfies the minimum co-cluster size. For example,

the reported co-cluster is $\{X_2, X_3, X_4, X_5\} \bigcup \{Y_1, Y_4, Y_6\}$ as shown in Figure 3.2. Likewise, the reported co-cluster is $\{X_7, X_8, X_9\} \bigcup \{Y_1, Y_3, Y_4, Y_8\}$ for the seed $\{X_7, X_8, X_9\}$. Notice that these two co-clusters are not overlapping, however, the columns Y_1 and Y_4 belong to both co-clusters simultaneously. Starting with each of the other columns, we repeat the above procedure and may identify all the hidden co-clusters.

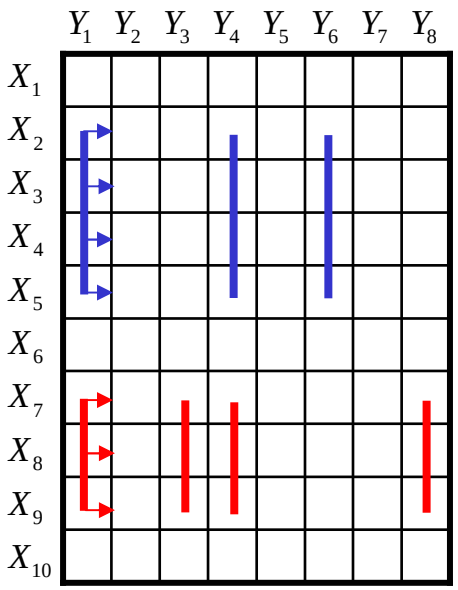


Figure 3.2. A Co-clustering Example

3.2.2. Hypothesis Testing

In this section, we introduce hypothesis testing procedures to test (3.1), which is needed to compute the distance matrix, and to test (3.3) which is used to determine which

columns are included in a co-cluster. In regards to testing hypothesis (3.1), we adopt the permutation test procedure discussed in Chapter 2 to obtain $p_{ij(k)}$.

In regards to testing hypothesis (3.3), we treat F_{seed} as a reference distribution since the seed is the starting point for building a co-cluster. Hence, we only focus on the quality index $Q(F_{seed}, F_{i'k'})$ that measures the overall "outlyingness" of $F_{i'k'}$ relative to F_{seed}. Assuming the scatter plot corresponding to $F_{i'k'}$ contains $N_{i'k'}$ data points and the pooled scatter plot corresponding to F_{seed} contains N_{seed} data points, we estimate $Q(F_{seed}, F_{i'k'})$ by $Q(F_{seed}^{N_{seed}}, F_{i'k'}^{N_{i'k'}})$. From Liu & Singh (1993) and Zuo & He (2006), we have

$$Q(F_{seed}^{N_{seed}}, F_{i'k'}^{N_{i'k'}}) - 1/2 \sim AN(0, (1/N_{seed} + 1/N_{i'k'})/12) \tag{3.4}$$

under $H_0 : F_{i'k'} = F_{seed}$ for many commonly used data depth functions (under general regularity conditions). Therefore, the p-value for testing hypothesis (3.3) is equal to $2 \times P\left(Z > \left|Q(F_{seed}^{N_{seed}}, F_{i'k'}^{N_{i'k'}}) - 1/2\right| / \sqrt{(1/N_{seed} + 1/N_{i'k'})/12}\right)$ where $Z \sim N(0,1)$.

For a seed $\{X_{i_1}, X_{i_2}, \cdots, X_{i_r}\}$ of size r with $\{i_1, i_2, \cdots, i_r\} \subset \{1, 2, \cdots, M\}$ that is generated from the hierarchical clustering of the k^{th} column, we test hypothesis (3.3) for each $i' \in \{i_1, i_2, \cdots, i_r\}$ and all the columns other than k to identify the column part of a potential co-cluster. Here, Holm's method (Holm 1979) is used to adjust these $r(N-1)$ p-values. Specifically, we sort $r(N-1)$ p-values in ascending order. If the smallest p-value is less than $\alpha_0 / [r(N-1)]$ with α_0 being a pre-specified overall type I

error, we reject the corresponding null hypothesis, and check whether the smallest p-value among the remaining $r(N-1)-1$ ones is less than $\alpha_0 / [r(N-1)-1]$ or not. We continue the above sequential comparison until the null hypothesis with the smallest p-value among the remaining ones is not rejected, and at that point, all the remaining null hypotheses are not rejected.

3.3. SIMULATION STUDY

To evaluate our proposed co-clustering method in Section 3.2, we performed the following simulation study:

1) For a set of rows and columns, specify a number of co-clusters and a bivariate distribution for the cells within each of the co-clusters. Additionally, specify a bivariate distribution for each of the remaining cells that are not contained in a co-cluster.
2) Generate random samples based on the given bivariate distributions in Step 1, which creates a data matrix of scatter plots. Apply our proposed co-clustering method to this data matrix of scatter plots, and identify the co-clusters.
3) For each sub-block of size $m \times n$ ($m = 1, \cdots, M$, $n = 1, \cdots, N$, and $mn > 1$), we check whether it belongs to any of the true co-clusters and any of the identified co-clusters (Since co-clustering aims at grouping rows and columns, we check

sub-blocks with at least two rows or two columns). This sub-block is defined to be "consistent" if it simultaneously belongs to some one of the true co-clusters and some one of the identified co-clusters, or it neither belongs to any of the true co-clusters nor any of the identified co-clusters. For each (m, n), the number of consistent sub-blocks is denoted by d_{mn}.

4) Repeat Step 2-3 a number of times, say L times, and accumulate the values of d_{mn}. The probability of consistency, $\sum_{\substack{m=1,\ n=1 \\ mn>1}}^{M}\sum^{N} d_{mn} / \{L\,[(2^M-1)(2^N-1)-MN]\}$, acts as the measure to evaluate the proposed co-clustering method.

We considered a data pattern setting as shown in Figure 3.3, in which two overlapping co-clusters were specified in a 8×6 data matrix. The two co-clusters are:

Co-cluster 1: $\{X_2, X_3, X_4, X_5\} \bigcup \{Y_2, Y_3, Y_4\}$;

Co-cluster 2: $\{X_4, X_5, X_6\} \bigcup \{Y_3, Y_4, Y_5\}$.

The scatter plots within each of the two co-clusters follow the bivariate normal distribution $N^{(2)}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$. The grey cells in Figure 3.3 represent cells that do not belong to the two co-clusters. The scatter plots for those cells follow a hierarchical bivariate distribution specified as follows: i) conditional on (X_L, X_U, Y_L, Y_U), X and Y are independent with distributions $Unif(X_L, X_U)$ and $Unif(Y_L, Y_U)$, respectively, and ii) (X_L, X_U, Y_L, Y_U) are independent with distributions $Unif(-4, 0)$, $Unif(0, 4)$,

$Unif(-4,0)$, $Unif(0,4)$, respectively. Furthermore, 100 data points were generated for each scatter plot, Mahalanobis depth was adopted, 500 resampling times were taken for the permutation test, and the complete linkage method was chosen for the hierarchical clustering.

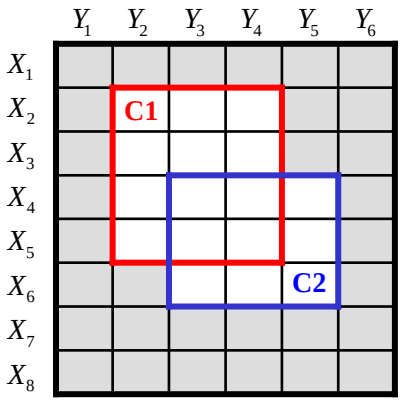


Figure 3.3. Co-cluster Specification

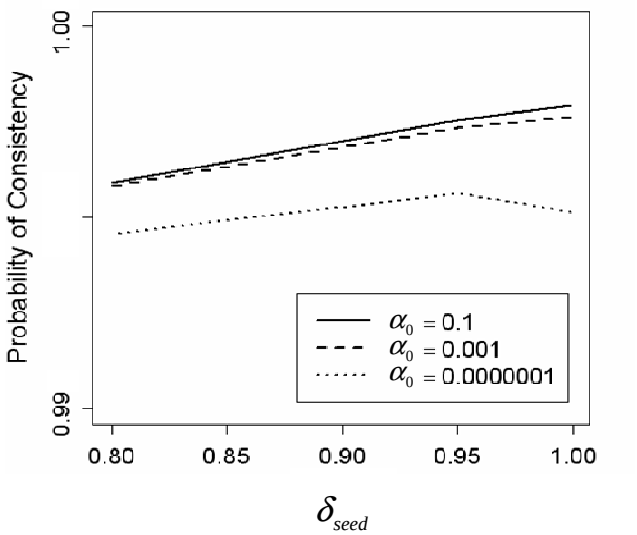


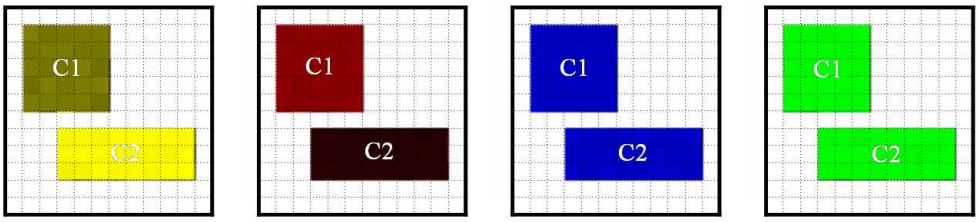
Figure 3.4. Probability of Consistency

We specified the minimum co-cluster size to be 3×3 and performed 500 simulations. The results are summarized in Figure 3.4, from which we notice the probability of consistency is pretty high for different scenarios of (δ_{seed},α_0), which demonstrates the power of our proposed co-clustering method.

3.4. DATA VISUALIZATION

Data visualization is an important aspect in the clustering technique. In the traditional co-clustering application in which cells of a data matrix are scalars, a graphical representation of the data matrix, called heat map, can be created where cells are painted with different colors based on their scalar values. Painting provides a visualization of the relative homogeneity within co-clusters. Obviously, we would expect cells in close proximity to each other to have a similar color. Although it is usually impossible to display all the co-clusters in a single heat map, each co-cluster would still show a color pattern when we investigate all the co-clusters one by one. For example, by defining a co-cluster to be the union of a subset of rows and a subset of columns within which all the scalars are similar to each other, each co-cluster would be represented by a block of cells that have similar colors.

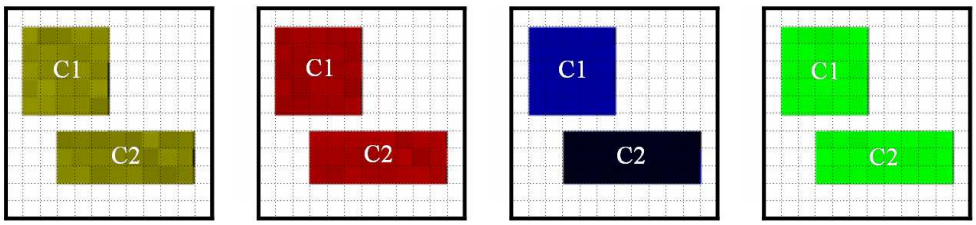
Co-cluster 1: $N^{(2)}\left(\begin{pmatrix}0\\4\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	Co-cluster 2: $N^{(2)}\left(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$
Noise: $\begin{pmatrix}X\\Y\end{pmatrix},\begin{cases}X\sim Unif(X_L,X_U),X_L\sim Unif(-12,0),X_U\sim Unif(0,12);\\Y\sim Unif(Y_L,Y_U),Y_L\sim Unif(-16,0),Y_U\sim Unif(0,8),X\perp Y.\end{cases}$	



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 3.5. Painting Example 1: OQI can distinguish two co-clusters from each other. Also, the bivariate normal distributions followed by two co-clusters only differ by location, and both means are located above the origin, therefore CDI can distinguish them from each other whereas CDDI and DI can not.

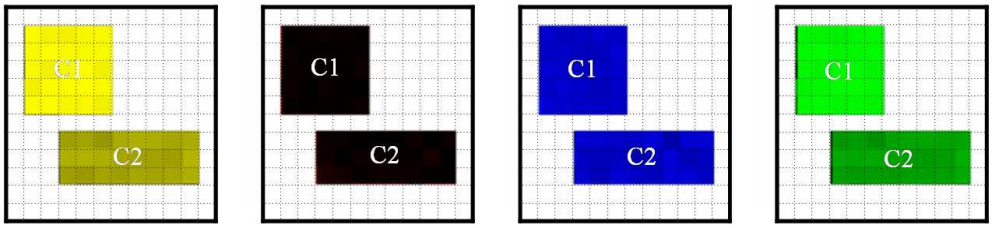
Co-cluster 1: $N^{(2)}\left(\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	Co-cluster 2: $N^{(2)}\left(\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$
Noise: $\begin{pmatrix}X\\Y\end{pmatrix},\begin{cases}X\sim Unif(X_L,X_U),X_L\sim Unif(-8,0),X_U\sim Unif(0,8);\\Y\sim Unif(Y_L,Y_U),Y_L\sim Unif(-8,0),Y_U\sim Unif(0,8),X\perp Y.\end{cases}$	



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 3.6. Painting Example 2: OQI can not distinguish two co-clusters from each other. Also, the bivariate normal distributions followed by two co-clusters only differ by location, and are symmetric about the origin, therefore CDDI can distinguish them from each other whereas CDI and DI can not.

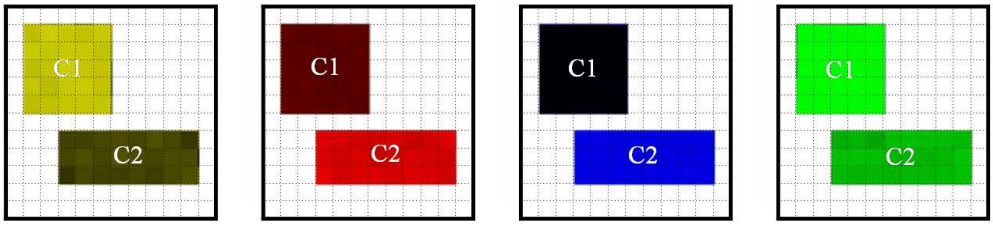
Co-cluster 1: $N^{(2)}\left(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	Co-cluster 2: $N^{(2)}\left(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}5&0\\0&5\end{pmatrix}\right)$
Noise: $\begin{pmatrix}X\\Y\end{pmatrix},\begin{cases}X\sim Unif(X_L,X_U),X_L\sim Unif(-12,0),X_U\sim Unif(0,12);\\Y\sim Unif(Y_L,Y_U),Y_L\sim Unif(-12,0),Y_U\sim Unif(0,12),X\perp Y.\end{cases}$	



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 3.7. Painting Example 3: OQI can distinguish two co-clusters from each other. Also, the bivariate normal distributions followed by two co-clusters only differ by scale, therefore DI can distinguish them from each other whereas CDI and CDDI can not.

Co-cluster 1: $N^{(2)}\left(\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$	Co-cluster 2: $N^{(2)}\left(\begin{pmatrix}-4\\4\end{pmatrix},\begin{pmatrix}5&0\\0&5\end{pmatrix}\right)$
Noise: $\begin{pmatrix}X\\Y\end{pmatrix},\begin{cases}X\sim Unif(X_L,X_U),X_L\sim Unif(-8,0),X_U\sim Unif(0,16);\\Y\sim Unif(Y_L,Y_U),Y_L\sim Unif(-14,0),Y_U\sim Unif(0,10),X\perp Y.\end{cases}$	



(a) OQI (b) CDI (c) CDDI (d) DI

Figure 3.8. Painting Example 4: OQI can distinguish two co-clusters from each other. Also, the bivariate normal distributions followed by two co-clusters differ by both location and scale, and are asymmetric about the origin, therefore CDI, CDDI and DI can all distinguish them from each other.

We use the same painting metrics discussed in Chapter 2 to graphically represent the co-clusters of scatter plots so that similar scatter plots are painted with a similar color whereas dissimilar scatter plots are painted with different colors. To illustrate the utility of painting metrics, we present four painting examples. In each example, a 12×12 matrix of scatter plots (each scatter plot contains 100 data points) was generated, and two non-overlapping co-clusters were specified as follows:

$$\text{Co-cluster 1: } \{X_2, X_3, X_4, X_5, X_6\} \bigcup \{Y_2, Y_3, Y_4, Y_5, Y_6\};$$

$$\text{Co-cluster 2: } \{X_8, X_9, X_{10}\} \bigcup \{Y_4, Y_5, Y_6, Y_7, Y_8, Y_9, Y_{10}, Y_{11}\}.$$

Data was generated for each scatter plot in the co-clusters by following the distributions shown in the first row of the top panel of Figures 3.5-3.8. Data for the remaining scatter plots in the matrix of scatter plots was generated from the noise distributions shown in the second row of the top panel in these figures. We then obtained 12×12 matrices of OQI, CDI, CDDI and DI, based upon which four heat maps can be generated as shown in the bottom panel of Figures 3.5-3.8, where the “yellow” heat map is based on OQI, the “red” heat map on CDI, the “blue” heat map on CDDI, the “green” heat map on DI, and the “black” color stands for the minimum index value in all the four heat maps. For simplicity, we used Mahalanobis depth in all the examples discussed here.

The painting examples illustrate that OQI captures the overall effect due to both location shift and scale change of a scatter plot. When OQI can distinguishes two scatter

plots from each other, one may further investigate CDI, CDDI, and DI to see the details of how these two scatter plots differ. Also, when OQI can not distinguish two scatter plots from each other, one may want to see if any of CDI, CDDI, and DI can distinguish them.

3.5. APPLICATION

3.5.1. Results

We revisit the microbe-host-interaction study discussed in Chapter 2. Two datasets were generated from the experiment (Li et al. 2010) to identify putatively important microbe-host interactions. “Microbe” data were arranged as a data matrix with 81 rows (3 rows containing missing values are excluded) standing for samples, 15 columns for microbes, and each cell being a single numerical value recording the level of a microbe in a sample. “Protein” data were also arranged as a data matrix with 81 rows standing for the same set of samples, 590 columns for proteins, and each cell being a single numerical value recording the level of a protein in a sample. To identify associations between levels of the microbes and proteins, we combined the above two data matrices of scalars by pairing up the columns (one from “Microbe” data, the other from “Protein” data) and treating each 81×2 array of data as bivariate data with the $x-$axis being microbe level and the $y-$axis being protein level. This process leads to a data matrix of scatter

plots as shown in Figure 3.1 where $M = 590$, $N = 15$, and each scatter plot contains 81 data points.

Considering the scatter plots as independent samples, we applied our proposed co-clustering method to the 590×15 data matrix of scatter plots. We used Mahalanobis depth as the data depth measure, $B = 500$ resampling times for the permutation test, and the complete linkage method in the hierarchical clustering to generate the seeds. Furthermore, we assumed the minimum co-cluster size is 20×5 (20 proteins and 5 microbes), and prespecified $\delta_{seed} = 0.8$ and $\alpha_0 = 0.2$.

Nine co-clusters were obtained, one of which is depicted in Figure 3.9, where the heat map with the OQI painting metric is shown. The promise of these results is demonstrated by the fact that many of the identified proteins have been previously associated with IBD as in Ahrenstedt et al. (1990), Larsson et al. (2006), Ripollés Piquer et al. (2006), and Fagerberg et al. (2007).

The proteins and microbes in the other eight identified co-clusters that have been previously associated with IBD are listed in Table 3.1. Apolipoprotein levels (c-ii and c-iii but not c-i) in blood have been shown to be useful biomarkers of IBD disease activity (Ripollés Piquer et al. 2006). S100A12, a calcium binding protein produced by granulocytes, has been associated with IBD (Foell et al. 2003, Foell et al. 2009). Increased levels of both complement C3 and C4 have been detected in IBD patients

(Ahrenstedt et al. 1990, Halstensen & Brandtzaeg 1991, Halstensen et al. 1992, Laufer et al. 2000, Ueki et al. 1996). *In vitro* studies of epithelial cells have also shown that complement factors open tight junctions (Conyers et al. 1990), which is consistent with various IBD theories involving barrier dysfunction. IBD patients often exhibit elevated levels of serum amyloid a (Ripollés Piquer et al. 2006), which is a protein involved in systemic AA amyloidosis (Lachmann et al. 2007). Transthyretin levels in serum were lower in Crohn's disease subjects than healthy controls (Reimund et al. 2005). Haptoglobin has been shown to be a marker for colitis in mouse models (Larsson et al. 2006, Torrence et al. 2008) and its precursor was shown to increase intestinal permeability in mice (Tripathi et al. 2009) and genetic variants have been associated with Crohn's disease (Papp et al. 2007). Chromogranin A levels were higher in IBD subjects than controls (Sciola et al. 2009, Yamaguchi et al. 2009). Inter-alpha trypsin inhibitor has been associated with human IBD (de la Motte et al. 2003) as well as a mouse model of IBD (Bandyopadhyay et al. 2008). Beta-2-microglobulin, osteopontin, and platelet basic protein have been shown to be reliable markers in IBD (Agnholt et al. 2007, Kruidenier et al. 2006, Zissis et al. 2001). PubMed (NCBI) searches did not identify reports linking the following proteins with IBD: amyloid beta a4, hemoglobin subunit beta (beta-globin), neurosecretory protein vgf, c-c motif chemokine 13, secretogranin-1, and proactivator polypeptide. Finally, based on analyses using BLAST (NCBI) (Altschul et al. 1997), only

two of the 13 bacterial phylotypes have been previously associated with IBD: Clostridium 12 (Presley et al. 2011) and Faecalibacterium 2994 phylotype (Baumgart et al. 2007, Martinez-Medina et al. 2006, Sokol et al. 2009, Swidsinski et al. 2008, Willing et al. 2009).

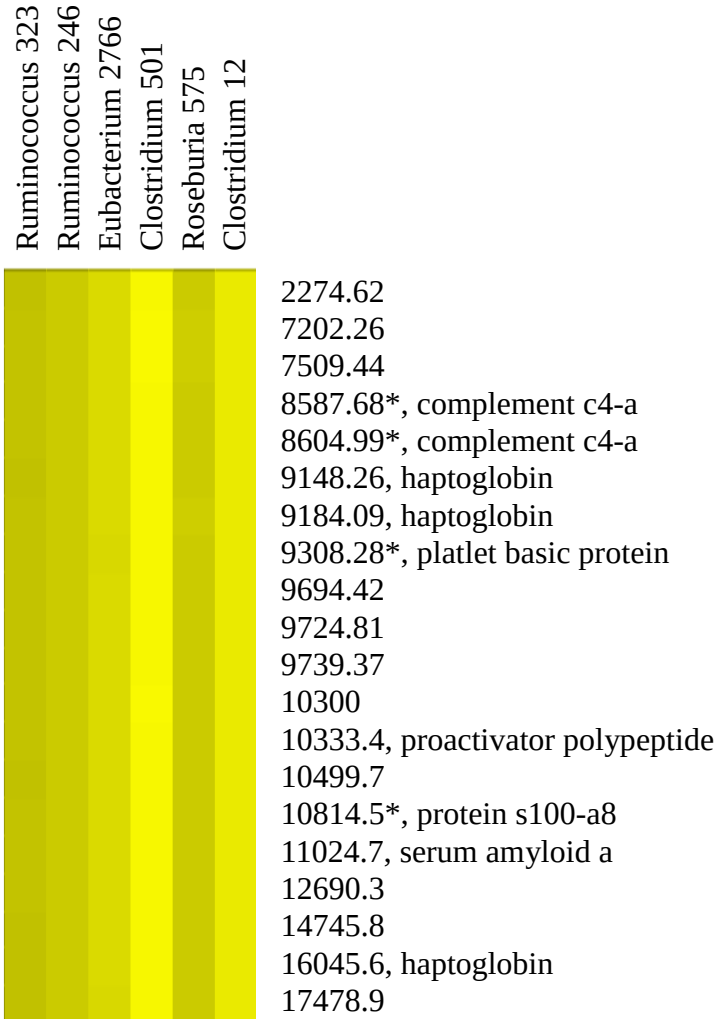


Figure 3.9. Heat Map with the OQI Painting Metric
(* indicates the immune system molecules.)

Table 3.1. Proteins and Microbes in the Identified Co-clusters

Protein	Microbe	Protein	Microbe
Co-cluster 1		Co-cluster 2	
8165.06*, complement c3 frag 7933.18, hemoglobin subunit beta 14315.2, transthyretin 9294.05*, platelet basic protein 7040.59, transthyretin 11710.9, serum amyloid a protein 10484.3*, protein s100-a12 8587.68*, complement c4-a frag 14272.2, transthyretin 8604.99*, complement c4-a frag, complement c4-a 8949.0, apolipoprotein a-ii 6231.41, secretogranin-1 frag 7741.71*, osteopontin frag	Clostridium 12 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Clostridium 501 Roseburia 575 Clostridium 603 Eubacterium 2766	7572.22, hemoglobin subunit alpha 9456.62, apolipoprotein c-iii 14315.2, transthyretin 10441.2*, protein s100-a12 8320.13*, complement c3 frag 9166.42, haptoglobin 10484.3*, protein s100-a12 11346.6, serum amyloid a protein 8604.99*, complement c4-a frag, complement c4-a 8134.06*, complement c3 frag 6231.41, secretogranin-1 frag 10333.4, proactivator polypeptide	AllBac Escherichia 8 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Roseburia 575 Eubacterium 2766
Co-cluster 3		Co-cluster 4	
11259.5, serum amyloid a protein 15861.0, hemoglobin subunit beta 9308.28*, c-c motif chemokine 13 13887.0, transthyretin 11979.7, beta-2-microglobulin 10892.5, serum amyloid a protein 15780.6, haptoglobin 8604.99*, complement c4-a frag, complement c4-a 10831.0*, protein s100-a8	Ruminococcus 3 Clostridium 12 Bacteroides 832 Eubacterium 2766 Faecalibacterium 2994	3895.27, chromogranin-a frag 8182.49, apolipoprotein c-ii frag 9166.42, haptoglobin 3961.47, neurosecretory protein vgf frag 15846.2, hemoglobin subunit beta 8967.25, apolipoprotein a-ii 3699.01, neurosecretory protein vgf frag	Ruminococcus 3 Clostridium 12 Bacteroides 832 Eubacterium 2766 Faecalibacterium 2994

Table 3.1 (Continued). Proteins and Microbes in the Identified Co-clusters

Protein	Microbe	Protein	Microbe
Co-cluster 5		Co-cluster 6	
3979.21, inter-alpha-trypsin inhibitor heavy chain h4 frag 2166.26, amyloid beta a4 protein 6614.87, apolipoprotein c-i 14272.2, transthyretin 15846.2, hemoglobin subunit beta 14240.7, transthyretin 11607.3, serum amyloid a protein 14040.3, transthyretin	Clostridium 12 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Roseburia 575 Clostridium 603 Eubacterium 2766	4793.06, neurosecretory protein vgf frag 11677.0, serum amyloid a protein 8182.49, apolipoprotein c-ii frag 6646.61, apolipoprotein c-i 6898.5, transthyretin 4807.93, neurosecretory protein vgf frag 3699.01, neurosecretory protein vgf frag	Clostridium 12 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Roseburia 575 Clostridium 603 Eubacterium 2766
Co-cluster 7		Co-cluster 8	
11259.5, serum amyloid a protein 7933.18, hemoglobin subunit beta 11979.7, beta-2-microglobulin 9294.05*, platelet basic protein 6417.37, apolipoprotein c-i frag, apolipoprotein c-i 8949.0, apolipoprotein a-ii 6628.66, apolipoprotein c-i	Escherichia 8 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Roseburia 575 Eubacterium 2766 Faecalibacterium 2994	15861.0, hemoglobin subunit beta 10814.5*, protein s100-a8 9456.62, apolipoprotein c-iii 11979.7, beta-2-microglobulin 8134.06*, complement c3 frag 10831.0*, protein s100-a8 10333.4, proactivator polypeptide	Clostridium 12 Ruminococcus 246 Ruminococcus 312 Ruminococcus 323 Clostridium 501 Roseburia 575 Clostridium 603 Eubacterium 2766

* indicates the immune system molecules.

Using our proposed co-clustering method, 22.6% of the proteins with a database match were immune system molecules, compared to only 14%, 12% and 9% for the other

methods (nearest shrunken centroids, RCCA, and ANOVA), respectively (see Presley et al. 2011 for other method values). These molecules are induced by a host immune response initiated by contact with microorganisms and their products, and as such are indicators of intimate host-microbe interplay occurring in the habitat under investigation. In contrast, the most predominant proteins identified by the other statistical methods (e.g., transthyretin, hemoglobin and serum amyloid) were high abundance proteins commonly and non-specifically associated with many settings of tissue injury. We therefore anticipate that our proposed co-clustering method may yield new and important clues regarding upstream host-microbe interplay associated with IBD, enhancing investigations of causal relationships in IBD and other multi-factorial disease etiologies.

3.5.2. Biological Hypothesis

We hypothesize that this approach will provide a more effective strategy for identifying causative variables associated with multi-factorial diseases such as IBD. For example, consider a multi-factorial disease in which one important factor is the increased production of a host immune molecule in response to a growing population of a particular microorganism. Because a direct physical or chemical interaction between the microbe and the immune molecule exists, the relationship between these variables will be tightly linked and relatively easy to detect. However, depending on the etiological complexity of

the disease, the levels of the microbe or immune molecule might not be strongly correlated to disease status. In general, as the number of factors contributing to an etiology increases, the strength of the linkage between the levels of any one specific variable and disease status decreases. Moreover, the linkage between the levels of these variables and disease status will probably fluctuate through cycles of remission and disease activity. However, in both cases, the relationships between the microbial and host variables will likely remain the same, and therefore detectable using our approach.

In a further attempt to elucidate cause from effect, our approach will also enable analysis of the strength and numbers of microbe-host relationships. New technologies have provided the ability to measure and analyze large numbers of variables, but most of these variables are not likely contributing to causation. Instead, differences in their levels are simply a response to environmental changes driven by the causative factors. Using our approach, one can identify and focus on those microorganisms having the strongest and/or the most numerous relationships with the host proteins. We theorize that such microorganisms are more likely to be involved in a direct physical or chemical interaction with the host, and therefore have a higher probability of being causative agents.

3.6. CONCLUSION

Our proposed co-clustering method showed a significant utility and power in handling a data matrix of scatter plots. The idea behind this co-clustering procedure can be applied to the higher dimensional clustering when one or more sets of variables needs to be analyzed. Moreover, the novel painting metrics we proposed can be easily extended to multidimensional clusters of multivariate plots.

Tukey depth and Simplicial depth are more robust to outliers than Mahalanobis depth. The computation complexity associated with these depths under the framework of our proposed co-clustering method will be addressed in future work.

Finally, although these methods were developed to analyze microbe-host interactions, we anticipate that this general approach will have utility for a wide range of investigations, including those examining relationships among gene expression profiles, metabolites, genes and epigenetic parameters.

Chapter 4

Co-clustering Spatial Data Using a Generalized Linear Mixed Model With Application to the Integrated Pest Management

4.1. INTRODUCTION

Integrated Pest Management (IPM) is a sustainable approach to managing pests by combining biological, cultural, physical and chemical tools in a way that minimizes economic losses, while simultaneously reducing human health and environmental risks. An important characteristic of an IPM program, which we focus on in this chapter, is the ability to accurately assess pest density levels. Recent literature has shown that pest density levels are influenced by spatial population dynamics. For example, spatial analyses have been applied in studies of agricultural pests of attacking lentils (Schotzko & O'Keeffe 1989), corn and alfalfa (Williams et al. 1992), cotton (Gozé et al. 2003), and grapes (Ifoulis & Savopoulou-Soultani 2006, Ramírez-Dávila & Porcayo-Camargo 2008). However, spatial analyses were usually conducted by transforming the count data to approximately satisfy the normality assumption (Gotway & Stroup 1997). Generalized Linear Mixed Models (GLMMs) (Breslow & Clayton 1993) can directly incorporate spatial correlation in count data, and have been used across multiple scientific disciplines, including ecological studies of pest populations (Barchia et al. 2003, Bennett et al. 2008,

Bianchi et al. 2008, Candy 2000, Elias et al. 2006, Elston et al. 2001, Paterson & Lello 2003, Takakura 2009).

Traditional pest assessment applications usually test hypotheses about a parameter θ that reflects the pest density within the whole orchard, such as the mean or median number of pests on each tree: $H_0 : \theta \leq \theta_c$ *vs.* $H_a : \theta > \theta_c$, where θ_c is a critical economic threshold for which the cost of treatment is equal to the cost of no treatment. Not rejecting H_0 would indicate no treatment intervention is required, whereas rejecting H_0 would call for treatment in an attempt to ward off serious crop loss (e.g., spraying pesticides or the release of natural enemies for pest control).

Often only specific areas of an orchard need treatment because many pest species exhibit clumped distributions, and it is within these "hotspots" where pest densities are high enough to warrant treatment. In this situation, under the present mode of operation pesticides may be applied to an entire orchard when treatment is required, including regions of the orchard that do not need treatment. Hence, within an IPM framework that is working to reduce unnecessary pesticide applications, a more sophisticated analytical procedure is desired to define localized regions with high pest infestations within an orchard for treatment.

There is very little literature about model-based co-clustering, and none of the literature has proposed a spatial co-clustering technique that co-clusters data such that

any co-cluster only contains a set of spatially consecutive rows and columns. In this chapter, we combine a spatial co-clustering technique with a statistical inference method to make pest assessments more reflective of their naturally occurring clumped distributions. In Section 4.2, we introduce a spatial GLMM to fit count data that exhibits spatial correlation within co-clusters. To avoid the high computational intensity associated with global optimization, we propose a heuristic optimization algorithm to search for a near optimal co-clustering. A sampling strategy is developed to maintain as much of the spatial information that is available from the data as possible, and the effect of sample size is studied. In Section 4.3, combining the heuristic optimization with the statistical inference, we develop a procedure to make assessment of pest density more accurate. We demonstrate the utility and power of our proposed procedure through simulation studies and apply the procedure to a study of assessing the density of persea mite (*Oligonychus perseae*) in Section 4.4.

4.2. METHODOLOGY

4.2.1. Spatial GLMM

4.2.1.1. Model Definition

Consider an $r \times c$ spatial grid with rows $(R_1, R_2, \cdots, R_r)$ and columns $(C_1, C_2, \cdots, C_c)$, in which each grid point is a potential sampling site. Co-clustering both

rows and columns, or simultaneously dividing both rows and columns into a number of contiguous and disjoint groups of rows and columns, we may obtain a “checkerboard” structure in the spatial grid, within which each block is referred to as a co-cluster. For a given co-clustering with n groups of rows and m groups of columns as shown in Figure 4.1, we use the term “design” to represent the specific row and column groupings that is denoted by

$$\{(R_1,\cdots,R_{i_1}),(R_{i_1+1},\cdots,R_{i_2}),\cdots,(R_{i_{n-1}+1},\cdots,R_{i_n})\}$$
$$\times\{(C_1,\cdots,C_{j_1}),(C_{j_1+1},\cdots,C_{j_2}),\cdots,(C_{j_{m-1}+1},\cdots,C_{j_m})\},$$

or more simply by the number of rows and columns within row groups and column groups respectively, $(i_1, i_2-i_1,\cdots,i_n-i_{n-1})\times(j_1, j_2-j_1,\cdots, j_m-j_{m-1})$. We also use the term “nomenclature” to represent the corresponding number of row groups and column groups and denote the nomenclature by $n\times m$. Notice that there is a one-to-one mapping between “co-clusterings” and “designs”, that is, each co-clustering corresponds to one and only one design, and vice versa. However, there exists a many-to-one mapping between “designs” and “nomenclatures”, that is, different designs may share the same nomenclature, but different nomenclatures must correspond to different designs.

The spatial GLMM for Figure 4.1 is defined to be:

$$Y_{j(i)} \mid \vec{\mathbf{s}} \overset{ind}{\sim} \textit{Negative Binomial}(\theta_i,\kappa),\quad i=1,2,\cdots,nm,\quad j=1,2,\cdots,n_i;$$
$$\log(\theta_i)=\mu+s_i; \tag{4.1}$$
$$\vec{\mathbf{s}}=(s_1,s_2,\cdots,s_{nm})'\sim MVN(0,\sigma^2\mathbf{I}_{nm});$$

where $Y_{j(i)}$ is the count number from the j^{th} sampling unit in the i^{th} co-cluster, n_i is the number of sampling units in the i^{th} co-cluster, θ_i is the conditional (on s_i) mean associated with the i^{th} co-cluster, κ quantifies the amount of overdispersion (relative to the Poisson distribution) for the Negative Binomial distribution (with $\kappa = \infty$ corresponding to no overdispersion), μ is the fixed intercept effect, s_i is a random effect associated with the i^{th} co-cluster, and $\mathbf{I}_{nm}$ is the identity matrix of size nm.

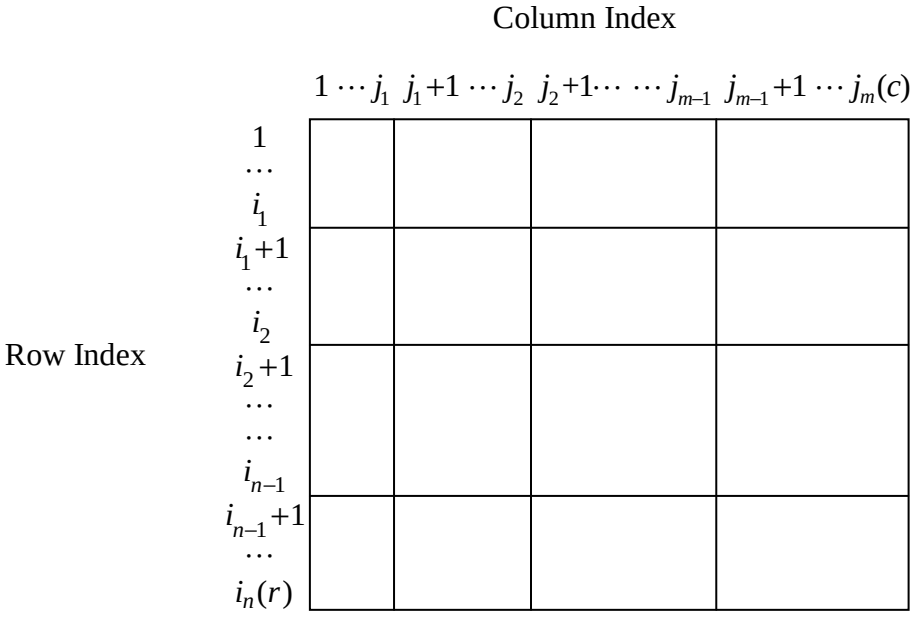


Figure 4.1. "Checkerboard" Co-cluster Structure

4.2.1.2. Likelihood and Parameter Estimation

The log-likelihood function corresponding to (4.1) can be derived as:

$$
\begin{aligned}
& l(\mu,\sigma^2,\kappa) \\
&= \log\prod_{i=1}^{nm}\left[\int_{-\infty}^{\infty} f(\vec{\mathbf{y}}_i, s_i)d(s_i)\right] \\
&= \sum_{i=1}^{nm}\log\left[\int_{-\infty}^{\infty} f(\vec{\mathbf{y}}_i \mid s_i) f(s_i)d(s_i)\right] \\
&= \sum_{i=1}^{nm}\log\left[\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f(y_{j(i)} \mid s_i) f(s_i)d(s_i)\right] \\
&= \sum_{i=1}^{nm}\log\left[\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\left(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\left(\frac{\kappa}{\exp(\mu+s_i)+\kappa}\right)^{\kappa}\left(\frac{\exp(\mu+s_i)}{\exp(\mu+s_i)+\kappa}\right)^{y_{j(i)}}\right)\right. \\
&\qquad \left.\cdot\frac{\exp(-s_i^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}}ds_i\right] \\
&= \sum_{i=1}^{nm}\sum_{j=1}^{n_i}\log\left(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\right)+ \\
&\quad \sum_{i=1}^{nm}\log\left[\int_{-\infty}^{\infty}\left(\frac{\kappa}{\exp(\mu+s_i)+\kappa}\right)^{n_i\kappa}\left(\frac{\exp(\mu+s_i)}{\exp(\mu+s_i)+\kappa}\right)^{\sum_{j=1}^{n_i}y_{j(i)}}\frac{\exp(-s_i^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}}ds_i\right], \qquad (4.2)
\end{aligned}
$$

where $\vec{\mathbf{y}}_i = (y_{1(i)}, y_{2(i)}, \cdots, y_{n_i(i)})'$.

Equation (4.2) involves nm one-dimensional integrals, each of which can be approximated as a weighted sum by the method of Gauss-Hermite quadrature:

$$
\begin{aligned}
l(\mu,\sigma^2,\kappa) &\approx \sum_{i=1}^{nm}\sum_{j=1}^{n_i}\log\left(\frac{\Gamma(y_{j(i)}+\kappa)}{\Gamma(y_{j(i)}+1)\Gamma(\kappa)}\right)+ \\
&\quad \sum_{i=1}^{nm}\log\left[\sum_{d=1}^{D}\left(\left(\frac{\kappa}{\exp(\mu+\sqrt{2\sigma^2}x_d)+\kappa}\right)^{n_i\kappa}\left(\frac{\exp(\mu+\sqrt{2\sigma^2}x_d)}{\exp(\mu+\sqrt{2\sigma^2}x_d)+\kappa}\right)^{\sum_{j=1}^{n_i}y_{j(i)}}\frac{w_d}{\sqrt{\pi}}\right)\right], \qquad (4.3)
\end{aligned}
$$

where x_d's and w_d's ($d = 1, 2, \cdots, D$) are the quadrature nodes and weights, respectively. Quadrature with $D = 30$ is usually enough for a good degree of approximation (McCulloch et al. 2008). Then (4.3) can be maximized numerically to obtain the MLEs of (μ, σ^2, κ), denoted as $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.

4.2.2. Model-based Co-clustering

4.2.2.1. Global Optimization

We define the optimal co-clustering to be the one with the maximum log-likelihood among all the possible co-clusterings. To avoid co-clusters that are too small, we specify the minimum co-cluster size to be $r_0 \times c_0$ ($r_0 > 1$ and $c_0 > 1$), in which r_0 is the minimum number of rows and c_0 is the minimum number of columns within the co-cluster. The global optimization algorithm is as follows:

1) Select a nomenclature and for each design associated with the nomenclature, fit the corresponding spatial GLMM and evaluate $l(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.
2) Repeat Step 1 for all the possible nomenclatures.
3) The global optimal co-clustering is the design with the maximum value of $l(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$.

In the global optimization algorithm, the number of possible co-clusterings that are evaluated can be shown to be:

$$\left[1+\sum_{n=2}^{\lfloor r/r_0 \rfloor}\left\{\binom{r-1}{n-1}+\sum_{i=1}^{n-1}\left[(-1)^i\binom{n}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}\right]\right\}\right]$$
$$\times\left[1+\sum_{m=2}^{\lfloor c/c_0 \rfloor}\left\{\binom{c-1}{m-1}+\sum_{i=1}^{m-1}\left[(-1)^i\binom{m}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{c_0-1}\binom{c-(j_1+j_2+\cdots+j_i)-1}{m-i-1}\right]\right\}\right]. \quad (4.4)$$

Proof of (4.4):

Consider an $r\times c$ spatial grid, and let the minimum co-cluster size be $r_0\times c_0$. Given a specific number of row groups n $(2\le n\le\lfloor r/r_0\rfloor)$, there are $\binom{r-1}{n-1}$ ways to obtain n row groups from r rows without considering the minimum co-cluster size.

In the situation that the minimum co-cluster size is not satisfied, there are $\sum_{j_1,j_2,\cdots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}$ ways to split n row groups into two parts, i $(1\le i\le n-1)$ specific row groups that do not satisfy the minimum co-cluster size, and the other $n-i$ row groups that may or may not satisfy the minimum co-cluster size. Also, there are $\binom{n}{i}$ ways to choose the above i specific row groups. However, summing up the groupings through all the possible i row groups duplicates the groupings that exact i' $(i'>i)$ row groups do not satisfy the minimum co-cluster size. Therefore, the total number of row groupings that do not satisfy the minimum co-cluster size is $\sum_{i=1}^{n-1}\left[(-1)^{i+1}\binom{n}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}\right]$.

It is trivial that there is only one row grouping for $n=1$. Hence the number of possible row groupings for all $1 \le n \le \lfloor r / r_0 \rfloor$ is

$$1+\sum_{n=2}^{\lfloor r/r_0 \rfloor}\left\{\binom{r-1}{n-1}+\sum_{i=1}^{n-1}\left[(-1)^i\binom{n}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}\right]\right\}.$$

The number of possible column groupings can be similarly derived as

$$1+\sum_{m=2}^{\lfloor c/c_0 \rfloor}\left\{\binom{c-1}{m-1}+\sum_{i=1}^{m-1}\left[(-1)^i\binom{m}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{c_0-1}\binom{c-(j_1+j_2+\cdots+j_i)-1}{m-i-1}\right]\right\}.$$

Therefore, the number of possible co-clusterings is

$$\left[1+\sum_{n=2}^{\lfloor r/r_0 \rfloor}\left\{\binom{r-1}{n-1}+\sum_{i=1}^{n-1}\left[(-1)^i\binom{n}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{r_0-1}\binom{r-(j_1+j_2+\cdots+j_i)-1}{n-i-1}\right]\right\}\right]$$
$$\times\left[1+\sum_{m=2}^{\lfloor c/c_0 \rfloor}\left\{\binom{c-1}{m-1}+\sum_{i=1}^{m-1}\left[(-1)^i\binom{m}{i}\sum_{j_1,j_2,\cdots,j_i=1}^{c_0-1}\binom{c-(j_1+j_2+\cdots+j_i)-1}{m-i-1}\right]\right\}\right]. \qquad \square$$

Table 4.1. Number of Co-clusterings for Global Optimization

	$r_0=c_0=6$	$r_0=c_0=8$	$r_0=c_0=10$	$r_0=c_0=12$
$r=c=20$	256	36	4	1
$r=c=25$	3,025	196	49	9
$r=c=30$	38,416	1,936	169	64
$r=c=35$	470,596	14,161	1,444	169
$r=c=40$	5,769,604	119,025	7,921	1,089
$r=c=45$	71,014,329	940,900	47,961	6,084

Some numerical examples that illustrate the formula in (4.4) are shown in Table 4.1, from which we notice the number of possible co-clusterings may be largely reduced by increasing r_0 and c_0, however, it still increases dramatically as r and c increase. For a relatively large spatial grid such as one of size 80×80, the number of possible co-clustering is 382,241,601 given that the minimum co-cluster size is 12×12. Therefore, exhaustively searching for the optimal co-clustering is usually not feasible in practice.

4.2.2.2. Heuristic Optimization

To avoid the extremely high computational intensity associated with global optimization, we propose the following heuristic optimization algorithm:

1) Starting with the original spatial grid, fit the corresponding spatial GLMMs for all the designs associated with the nomenclatures 1×2 and 2×1, and record the co-clustering with the maximum $l(\hat{\mu},\hat{\sigma}^2,\hat{\kappa})$ as the "Current Optimal Co-clustering" whose log-likelihood is denoted by $l^*(\hat{\mu},\hat{\sigma}^2,\hat{\kappa})$.
2) Starting with the "Current Optimal Co-clustering", fit the corresponding spatial GLMMs for all the designs with the nomenclature that has either one more row group or one more column group than the "Current Optimal Co-clustering", and record the co-clustering with the maximum $l(\hat{\mu},\hat{\sigma}^2,\hat{\kappa})$ as the "Potential Optimal Co-clustering" whose log-likelihood is denoted by $l^0(\hat{\mu},\hat{\sigma}^2,\hat{\kappa})$.

3) If $l^0(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa}) > l^*(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$, replace the "Current Optimal Co-clustering" with the "Potential Optimal Co-clustering" and repeat Step 2; otherwise, stop the procedure and report the "Current Optimal Co-clustering" as the heuristic optimal co-clustering.

With the minimum co-cluster size not considered, the number of co-clusterings that are evaluated in the heuristic optimization algorithm is

$$[r + c - 1 - (n^* + m^*)/2](n^* + m^* - 1),$$

where $n^* \times m^*$ is the nomenclature for the heuristic optimal co-clustering.

4.2.2.3. Efficiency of Heuristic Optimization Algorithm

To study the efficiency of our proposed heuristic optimization algorithm, we performed a simulation study to compare it to the global optimization algorithm. The simulation study is as follows:

1) Specify a nomenclature and a specific design.
2) Simulate count data for the spatial grid associated with the design selected in Step 1 using specified parameters.
3) Apply both the global optimization algorithm and the heuristic optimization algorithm to the spatial grid, and for each algorithm, check whether the original design and nomenclature can be retrieved or not. That is, we check whether the corresponding design and nomenclature of the reported optimal co-clustering from both the global

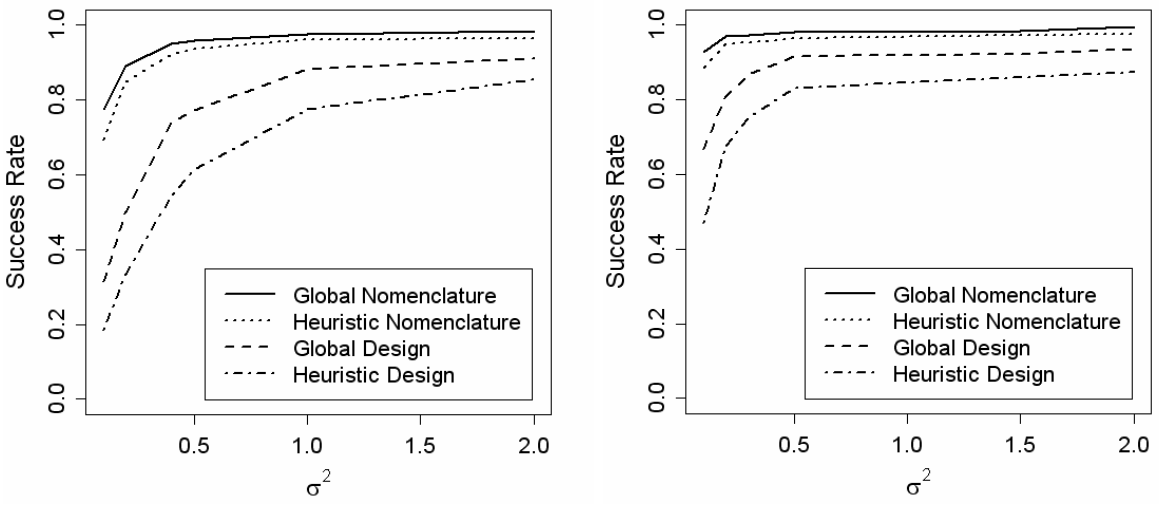
optimization algorithm and the heuristic optimization algorithm are same as the true design and nomenclature respectively.

4) Repeat Step 2–3 a number of times, and for each algorithm, record the success rates for the reported optimal design and nomenclature, i.e., the proportions of times that we succeed in retrieving the original design and nomenclature.

For all the simulation studies in this chapter, we considered a 40×40 spatial grid using the design $(10,17,13)\times(13,15,12)$ and hence the nomenclature 3×3, specified the minimum co-cluster size to be $r_0\times c_0 = 10\times 12$, and performed 1000 simulations for each setting. Based on the model fitting analyses discussed in Section 4.4, we chose $\mu = 6$ and different scenarios for (κ,σ^2). Throughout this chapter, the number of nodes used in the Gauss-Hermite quadrature was selected to be $D = 30$. The results are summarized in Figure 4.2, in which Figure 4.2(a) shows the success rates for the reported design and nomenclature for the different (κ,σ^2) scenarios when $\kappa = 1$, and Figure 4.2(b) similarly shows the success rates for the case $\kappa = 3$.

In this simulation study, the number of co-clusterings evaluated in the global optimization algorithm is 2937, whereas the number of co-clusterings evaluated in the heuristic optimization algorithm is around 85 on average. From either Figure 4.2(a) or Figure 4.2(b), we may notice the success rate of the design or nomenclature for the heuristic optimization algorithm is not that much lower than that for the global

optimization algorithm. Also, the success rate of the design or nomenclature increases as σ^2 increases given μ and κ, which indicates that greater difference among true co-clusters improves the chance of retrieving the true design or nomenclature. Comparing Figure 4.2(a) to Figure 4.2(b), we notice the success rate of the design or nomenclature increases as κ increases given μ and σ^2, meaning that less variability within true co-clusters also improves the chance of capturing the true design or nomenclature.



(a) $\kappa = 1$ (b) $\kappa = 3$

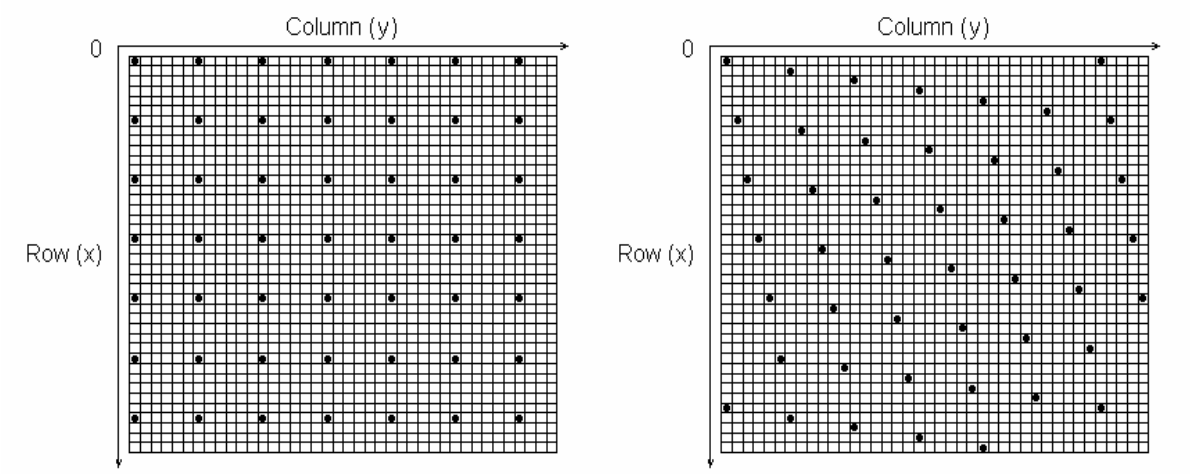
Figure 4.2. Heuristic Optimization vs. Global Optimization

4.2.2.4. Heuristic Optimization with Non-Exhaustive Samples

Concerning time and the cost of human resources, practitioners usually sample less than 100% of the grid points from the spatial grid. Next we develop a sampling strategy

for this case, and study how the sample size affects the success rates for the reported design and nomenclature.

First note that if we randomly sample a subset of grid points from the spatial grid, it is very likely that specific areas of the spatial grid will not be represented in the sample, especially when the sampling fraction is small. In this case, we can anticipate that some of the resulting co-clusters will not have been sampled and in some applications, such as the one we discuss in Section 4.3, this can lead to loss of precision in subsequent inference procedures.



(a) Even Sampling Scheme (b) Shifted Sampling Scheme

Figure 4.3. Sampling Strategy

Recall that the minimum co-cluster size is $r_0 \times c_0$. To ensure at least one grid point is taken from each co-cluster, we may start with the first grid point (the grid point located in the first row and the first column), and sample a grid point every r_0 rows along the row dimension and every c_0 columns along the column dimension. By doing so, all the sampled grid points are evenly distributed across the spatial grid such that any $r_0 \times c_0$ sub-grid contains at least one sampled grid point, as shown in Figure 4.3(a) in which the spatial grid is of size 40×40, the minimum co-cluster size is 6×6, and 49 sampled grid points are denoted by the black dots. Mathematically, by taking the row dimension as $x-$axis, the column dimension as $y-$axis, and row and column indices as $x-$coordinates and $y-$coordinates respectively, the positions of the sampled grid points in the Cartesian plane are the intersections of the lines $x = ir_0 + 1$ $(i = 0,1,\cdots,\lfloor (r-1)/r_0 \rfloor)$ and the lines $y = jc_0 + 1$ $(j = 0,1,\cdots,\lfloor (c-1)/c_0 \rfloor)$.

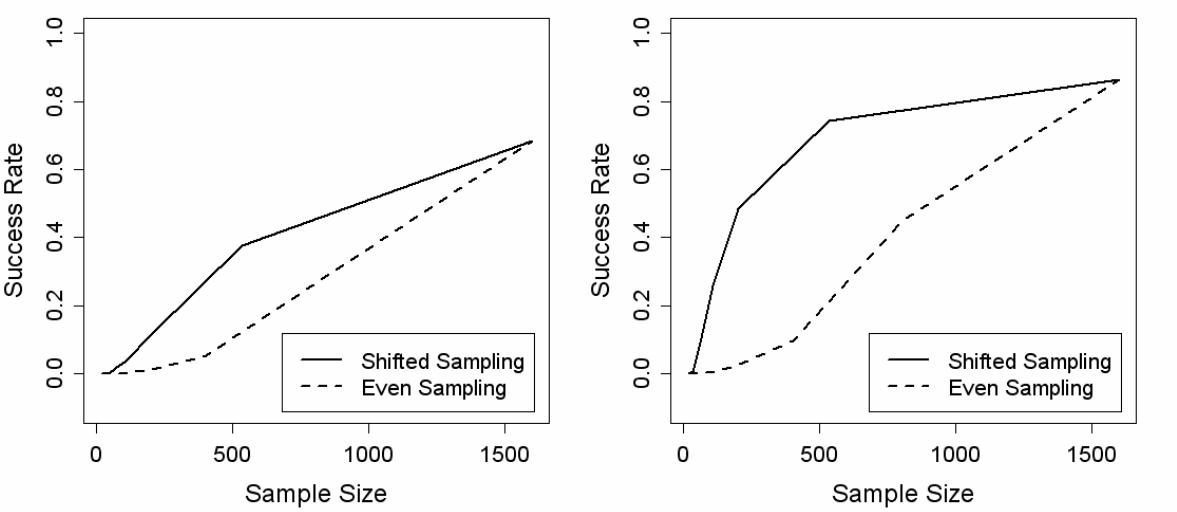
However, for a co-clustering with a row separation line or a column separation line located in the gap between the sampled grid points, moving around this separation line within the gap will not change the log-likelihood for the corresponding spatial GLMM. For example, consider two designs $(10,14,16)\times(12,13,15)$ and $(10,14,16)\times(12,14,14)$ for Figure 4.3(a), the only difference between which is a column separation line, one is between the 25^{th} and 26^{th} column, and the other is between the 26^{th} and 27^{th} column. Both designs lead to the same log-likelihood since there is no sampled grid point

coming from the 26^{th} column. Therefore, the reported optimal co-clustering is not unique.

To increase the chance of retrieving the true design, we propose an alternative sampling strategy as shown in Figure 4.3(b), in which each sampled grid point in Figure 4.3(a) is shifted one more row than the previously sampled grid point along the column dimension, and shifted one more column than the previously sampled grid point along the row dimension. Mathematically, the positions of the sampled grid points in the Cartesian plane are the intersections of the segments $x = i$ $(i = 1, 2, \cdots, r;\ 1 \le y \le c)$ and the lines $y = c_0(x - jr_0 - 1) + j + 1$ $(\ j = \lceil -(c-1)/(r_0 c_0 - 1) \rceil,\ \cdots, \lfloor c_0(r-1)/(r_0 c_0 - 1) \rfloor\ )$. By using this sampling strategy, not only does any $r_0 \times c_0$ sub-grid contain at least one sampled grid point, but also the sampled grid points overall reflect as much of the spatial information embedded in the spatial grid as possible. When the sampling fraction is small, it is still possible (though less likely) that moving row or column separation lines within a gap does not change the log-likelihood in the process of heuristic optimization algorithm. In this case, the heuristic optimization algorithm proceeds to further steps by randomly choosing from the alternatives that have the same log-likelihood.

Based on the minimum co-cluster size, the proposed sampling strategy would provide the minimum sample size required for co-clustering. For example, the minimum sample size in Figure 4.3(b) is 46. When practitioners can afford to sample more grid

points, we may increase the sample size by replacing r_0 with a smaller "row step" r^* ($1 \le r^* \le r_0$) and c_0 with a smaller "column step" c^* ($1 \le c^* \le c_0$) such that any $r^* \times c^*$ sub-grid contains at least one sampled grid point. For example, $r^* = c^* = 4$ leads to 108 sampled grid points in Figure 4.3(b) for the shifted sampling strategy.



(a) $\sigma^2 = 0.2$ (b) $\sigma^2 = 1$

Figure 4.4. Success Rate of Design vs. Sample Size

We performed a simulation study to evaluate how the sample size affects the success rate of the design. Here, we specified $\mu = 6$, $\kappa = 3$ and different values for σ^2. For each scenario of (μ, σ^2, κ), we sampled grid points based on both the even sampling strategy and the shifted sampling strategy, and chose different values of (r^*, c^*) to reach the corresponding sample sizes. The results are summarized in Figure 4.4, in which

Figure 4.4(a) shows the relationship between the success rate of the design and the sample size for $\sigma^2 = 0.2$ and Figure 4.4(b) for $\sigma^2 = 1$. From either Figure 4.4(a) or Figure 4.4(b), we notice the success rate of the design increases as the sample size increases given μ, σ^2 and κ, and the success rate of the design for the shifted sampling strategy is much higher than that for the even sampling strategy given μ, σ^2, κ and the sample size. Comparing Figure 4.4(a) to Figure 4.4(b), we may notice the success rate of the design increases as σ^2 increases given μ, κ and the sample size.

4.3. APPLICATION TO PEST DENSITY ASSESSMENT

4.3.1. Proposed Methodology

Here we consider an application to assess orchards of fruit-bearing trees for a potential pest problem. Our goal is to identify the infested regions within orchards that require treatment such as spraying pesticides or, alternatively, the release of natural enemies. Trees within orchards are frequently organized in a grid of rows and columns. Treating an orchard as a spatial grid, we first take a sample of trees (grid points) from the orchard based on the sampling strategy discussed in Section 4.2.2.4, and count pests of each sampled tree. Applying our proposed heuristic optimization algorithm to this spatial grid, we obtain the heuristic optimal co-clustering of the orchard. We then further analyze each co-cluster, as follows, to determine whether treatment is required or not.

For each co-cluster of the heuristic optimal co-clustering, we use the model in (4.1) to predict its conditional mean $\theta_i = \exp(\mu + s_i)$ $(i = 1, 2, \cdots, nm)$ using the Best Linear Predictor (BLP)

$$\tilde{\theta}_i = \mathrm{BLP}(\theta_i)$$

$$= \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1) \cdot \sum_{j=1}^{n_i} y_{j(i)} + \exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}. \quad (4.5)$$

The Mean Square Error (MSE) of $\tilde{\theta}_i$ is

$$\mathrm{MSE}(\tilde{\theta}_i) = \frac{\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)(\exp(\mu + 3\sigma^2/2)/\kappa + 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}. \quad (4.6)$$

Proof of (4.5) and (4.6):

It is trivial to show that

$$E(\theta_i) = E(\exp(\mu + s_i)) = \exp(\mu + \sigma^2/2).$$

$$Var(\theta_i) = Var(\exp(\mu + s_i)) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1).$$

For $j = 1, 2, \cdots, n_i$, we have

$$E(y_{j(i)}) = E\{E(y_{j(i)} \mid s_i)\} = E(\exp(\mu + s_i)) = \exp(\mu + \sigma^2/2).$$

$$\begin{aligned} & Cov(\exp(\mu + s_i), y_{j(i)}) \\ &= E\{Cov(\exp(\mu + s_i), y_{j(i)} \mid s_i)\} + Cov\{E(\exp(\mu + s_i) \mid s_i), E(y_{j(i)} \mid s_i)\} \\ &= 0 + Cov(\exp(\mu + s_i), \exp(\mu + s_i)) \\ &= \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1). \end{aligned}$$

Thus

$$E\left(\vec{\mathbf{y}}_i\right) = \exp(\mu + \sigma^2/2)\cdot\vec{\mathbf{1}}_{n_i},$$

$$Cov\left(\exp(\mu + s_i), \vec{\mathbf{y}}_i\right) = \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)\cdot\vec{\mathbf{1}}'_{n_i}, \tag{4.7}$$

where $\vec{\mathbf{y}}_i = (y_{1(i)}, y_{2(i)}, \cdots, y_{n_i(i)})'$, and $\vec{\mathbf{1}}_{n_i}$ is the n_i-tuple column vector of all 1's.

For $j,\ j' = 1,2,\cdots,n_i$ and $j \neq j'$, we have

$$\begin{aligned} Var\left(y_{j(i)}\right) &= E\left\{Var\left(y_{j(i)} \mid s_i\right)\right\} + Var\left\{E\left(y_{j(i)} \mid s_i\right)\right\} \\ &= E\left(\exp(2\mu+2s_i)/\kappa + \exp(\mu+s_i)\right) + Var\left(\exp(\mu+s_i)\right) \\ &= \exp(2\mu+2\sigma^2)/\kappa + \exp(\mu+\sigma^2/2) + \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1). \end{aligned}$$

$$\begin{aligned} Cov\left(y_{j(i)}, y_{j'(i)}\right) &= E\left\{Cov\left(y_{j(i)}, y_{j'(i)} \mid s_i\right)\right\} + Cov\left\{E\left(y_{j(i)} \mid s_i\right), E\left(y_{j'(i)} \mid s_i\right)\right\} \\ &= 0 + Cov\left(\exp(\mu+s_i), \exp(\mu+s_i)\right) \\ &= \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1). \end{aligned}$$

Thus

$$Var\left(\vec{\mathbf{y}}_i\right) = (\exp(2\mu+2\sigma^2)/\kappa + \exp(\mu+\sigma^2/2))\cdot\mathbf{I}_{n_i} + \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)\cdot\mathbf{J}_{n_i},$$

$$\text{and } Var^{-1}(\vec{\mathbf{y}}_i) = \frac{1}{\exp(2\mu+2\sigma^2)/\kappa + \exp(\mu+\sigma^2/2)} \cdot\left(\mathbf{I}_{n_i} - \frac{\exp(\mu+\sigma^2/2)(\exp(\sigma^2)-1)}{\exp(\mu+3\sigma^2/2)/\kappa + 1 + n_i\exp(\mu+\sigma^2/2)(\exp(\sigma^2)-1)}\mathbf{J}_{n_i}\right), \tag{4.8}$$

where $\mathbf{I}_{n_i}$ is the identity matrix of size n_i and $\mathbf{J}_{n_i}$ is the n_i-by-n_i matrix with all 1's .

Hence, from (4.7) and (4.8), we have

$$\begin{aligned}
\tilde{\theta}_i &= E(\theta_i) + Cov(\theta_i, \vec{\mathbf{y}}_i) \cdot Var^{-1}(\vec{\mathbf{y}}_i) \cdot (\vec{\mathbf{y}}_i - E(\vec{\mathbf{y}}_i)) \\
&= E(\exp(\mu + s_i)) + Cov(\exp(\mu + s_i), \vec{\mathbf{y}}_i) \cdot Var^{-1}(\vec{\mathbf{y}}_i) \cdot (\vec{\mathbf{y}}_i - E(\vec{\mathbf{y}}_i)) \\
&= \exp(\mu + \sigma^2/2) + \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \cdot \vec{\mathbf{1}}'_{n_i} \cdot \frac{1}{\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)} \\
&\quad \cdot \left(\mathbf{I}_{n_i} - \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)} \mathbf{J}_{n_i} \right) \\
&\quad \cdot \left(\vec{\mathbf{y}}_i - \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}_{n_i} \right) \\
&= \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1) \cdot \sum_{j=1}^{n_i} y_{j(i)} + \exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.
\end{aligned}$$

$$\begin{aligned}
MSE(\tilde{\theta}_i) &= Var\left(\tilde{\theta}_i - \theta_i\right) \\
&= Var(\theta_i) - Cov(\theta_i, \vec{\mathbf{y}}_i) \cdot Var^{-1}(\vec{\mathbf{y}}_i) \cdot Cov'(\theta_i, \vec{\mathbf{y}}_i) \\
&= \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) - \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \cdot \vec{\mathbf{1}}'_{n_i} \\
&\quad \cdot \frac{1}{\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)} \\
&\quad \cdot \left(\mathbf{I}_{n_i} - \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)} \mathbf{J}_{n_i} \right) \\
&\quad \cdot \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \cdot \vec{\mathbf{1}}_{n_i} \\
&= \frac{\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)(\exp(\mu + 3\sigma^2/2)/\kappa + 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.
\end{aligned}$$

□

With (μ, σ^2, κ) replaced with the MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ in (4.5) and (4.6), the empirical Best Linear Predictor (eBLP) of θ_i is

$$\begin{aligned}
\hat{\tilde{\theta}}_i &= \text{eBLP}(\theta_i) \\
&= \frac{\exp(\hat{\mu} + \hat{\sigma}^2/2)(\exp(\hat{\sigma}^2) - 1) \cdot \sum_{j=1}^{n_i} y_{j(i)} + \exp(2\hat{\mu} + 2\hat{\sigma}^2)/\hat{\kappa} + \exp(\hat{\mu} + \hat{\sigma}^2/2)}{\exp(\hat{\mu} + 3\hat{\sigma}^2/2)/\hat{\kappa} + 1 + n_i \exp(\hat{\mu} + \hat{\sigma}^2/2)(\exp(\hat{\sigma}^2) - 1)},
\end{aligned} \quad (4.9)$$

and the estimated MSE of $\tilde{\theta}_i$ is

$$\widehat{\text{MSE}}(\tilde{\theta}_i) = \frac{\exp(2\hat{\mu}+\hat{\sigma}^2)(\exp(\hat{\sigma}^2)-1)(\exp(\hat{\mu}+3\hat{\sigma}^2/2)/\hat{\kappa}+1)}{\exp(\hat{\mu}+3\hat{\sigma}^2/2)/\hat{\kappa}+1+n_i\exp(\hat{\mu}+\hat{\sigma}^2/2)(\exp(\hat{\sigma}^2)-1)}. \quad (4.10)$$

Define $U_i = \left(\hat{\tilde{\theta}}_i - \theta_i\right) \Big/ \sqrt{\widehat{\text{MSE}}(\tilde{\theta}_i)}$ and let $U_{i,\alpha}$ be the $100(1-\alpha)^{th}$ conditional percentile of U_i given $\vec{\mathbf{s}}$. Then a $100(1-\alpha)\%$ lower conditional prediction bound of θ_i given $\vec{\mathbf{s}}$ is

$$L_{i,\alpha} = \max\left(0, \hat{\tilde{\theta}}_i - U_{i,\alpha}\sqrt{\widehat{\text{MSE}}(\tilde{\theta}_i)}\right). \quad (4.11)$$

For a pre-specified threshold θ_c, the decision of "Treat" is made if $L_{i,\alpha} > \theta_c$; otherwise the decision of "Do Not Treat" is made. The value of $U_{i,\alpha}$ in (4.11) can be approximated from the following parametric bootstrap procedure:

1) Generate an $r \times c$ spatial grid based on the heuristic optimal co-clustering, with insect counts from trees in the co-clusters having independent distributions of *Negtive Binomial*$(\hat{\tilde{\theta}}_i, \hat{\kappa})$ $(i = 1, 2, \cdots, nm)$.
2) Fit the spatial GLMM based on the sampled tree locations to estimate (μ, σ^2, κ) as $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$. With $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ replaced with $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$ in (4.9) and (4.10), calculate $\hat{\tilde{\theta}}_i^* = \text{eBLP}(\theta_i)$ and $\widehat{\text{MSE}}^*(\tilde{\theta}_i)$, and then $U_i^* = \left(\hat{\tilde{\theta}}_i^* - \hat{\tilde{\theta}}_i\right) \Big/ \sqrt{\widehat{\text{MSE}}^*(\tilde{\theta}_i)}$;.
3) Repeat Step 1–2 a number of times, say B times, to obtain $(U_i^{*(1)}, U_i^{*(2)}, \cdots, U_i^{*(B)})$, and approximate $U_{i,\alpha}$ by the $100(1-\alpha)^{th}$ percentile of $(U_i^{*(1)}, U_i^{*(2)}, \cdots, U_i^{*(B)})$.

When the number of co-clusters of the heuristic optimal co-clustering is relatively large, we may adjust the significance level α to form the simultaneous lower conditional prediction bounds of the conditional means for co-clusters, such as by the method of Bonferroni correction or Sidak correction (Olejnik et al. 1997).

4.3.2. Coverage Probability

In Section 4.3.1, for each co-cluster of the heuristic optimal co-clustering, we may also use the model in (4.1) to predict the co-cluster effect s_i $(i = 1, 2, \cdots, nm)$ using the Best Linear Predictor (BLP)

$$\begin{aligned}\tilde{s}_i &= \mathrm{BLP}(s_i) \\ &= \frac{\sigma^2\left(\sum_{j=1}^{n_i} y_{j(i)} - n_i \exp(\mu + \sigma^2/2)\right)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.\end{aligned} \tag{4.12}$$

The Mean Square Error (MSE) of $\tilde{s}_i$ is

$$\mathrm{MSE}(\tilde{s}_i) = \sigma^2 - \frac{n_i \sigma^4 \exp(\mu + \sigma^2/2)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}. \tag{4.13}$$

Proof of (4.12) and (4.13):

For $j = 1, 2, \cdots, n_i$, we have

$$\begin{aligned}Cov\left(s_i, y_{j(i)}\right) &= E\left\{Cov\left(s_i, y_{j(i)} \mid s_i\right)\right\} + Cov\left\{E\left(s_i \mid s_i\right), E\left(y_{j(i)} \mid s_i\right)\right\} \\ &= 0 + Cov\left(s_i, \exp(\mu + s_i)\right) \\ &= \sigma^2 \exp(\mu + \sigma^2/2).\end{aligned}$$

Thus

$$Cov\left(s_i, \vec{\mathbf{y}}_i\right) = \sigma^2 \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}'_{n_i}, \tag{4.14}$$

where $\vec{\mathbf{y}}_i = (y_{1(i)}, y_{2(i)}, \cdots, y_{n_i(i)})'$, and $\vec{\mathbf{1}}_{n_i}$ is the n_i-tuple column vector of all 1's.

Hence, from (4.8) and (4.14), we have

$$\begin{aligned}
\tilde{s}_i &= E\left(s_i\right) + Cov\left(s_i, \vec{\mathbf{y}}_i\right) \cdot Var^{-1}\left(\vec{\mathbf{y}}_i\right) \cdot \left(\vec{\mathbf{y}}_i - E(\vec{\mathbf{y}}_i)\right) \\
&= 0 + \sigma^2 \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}'_{n_i} \cdot \frac{1}{\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)} \\
&\quad \cdot \left(\mathbf{I}_{n_i} - \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)} \mathbf{J}_{n_i} \right) \\
&\quad \cdot \left(\vec{\mathbf{y}}_i - \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}_{n_i} \right) \\
&= \frac{\sigma^2 \left(\sum_{j=1}^{n_i} y_{j(i)} - n_i \exp(\mu + \sigma^2/2) \right)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.
\end{aligned}$$

$$\begin{aligned}
MSE(\tilde{s}_i) &= Var\left(\tilde{s}_i - s_i\right) \\
&= Var\left(s_i\right) - Cov\left(s_i, \vec{\mathbf{y}}_i\right) \cdot Var^{-1}\left(\vec{\mathbf{y}}_i\right) \cdot Cov'\left(s_i, \vec{\mathbf{y}}_i\right) \\
&= \sigma^2 - \sigma^2 \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}'_{n_i} \cdot \frac{1}{\exp(2\mu + 2\sigma^2)/\kappa + \exp(\mu + \sigma^2/2)} \\
&\quad \cdot \left(\mathbf{I}_{n_i} - \frac{\exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)} \mathbf{J}_{n_i} \right) \\
&\quad \cdot \sigma^2 \exp(\mu + \sigma^2/2) \cdot \vec{\mathbf{1}}_{n_i} \\
&= \sigma^2 - \frac{n_i \sigma^4 \exp(\mu + \sigma^2/2)}{\exp(\mu + 3\sigma^2/2)/\kappa + 1 + n_i \exp(\mu + \sigma^2/2)(\exp(\sigma^2) - 1)}.
\end{aligned}$$

□

With (μ, σ^2, κ) replaced with the MLEs $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ in (4.12) and (4.13), the empirical Best Linear Predictor (eBLP) of s_i is

$$\begin{aligned}\hat{\tilde{s}}_i &= \text{eBLP}(s_i) \\ &= \frac{\hat{\sigma}^2\left(\sum_{j=1}^{n_i} y_{j(i)} - n_i \exp(\hat{\mu} + \hat{\sigma}^2/2)\right)}{\exp(\hat{\mu} + 3\hat{\sigma}^2/2)/\hat{\kappa} + 1 + n_i \exp(\hat{\mu} + \hat{\sigma}^2/2)(\exp(\hat{\sigma}^2) - 1)},\end{aligned} \tag{4.15}$$

and the estimated MSE of $\tilde{s}_i$ is

$$\widehat{\text{MSE}}(\tilde{s}_i) = \hat{\sigma}^2 - \frac{n_i \hat{\sigma}^4 \exp(\hat{\mu} + \hat{\sigma}^2/2)}{\exp(\hat{\mu} + 3\hat{\sigma}^2/2)/\hat{\kappa} + 1 + n_i \exp(\hat{\mu} + \hat{\sigma}^2/2)(\exp(\hat{\sigma}^2) - 1)}. \tag{4.16}$$

Define $V_i = \left(\hat{\tilde{s}}_i - s_i\right) \Big/ \sqrt{\widehat{\text{MSE}}(\tilde{s}_i)}$ and let $V_{i,\alpha}$ be the $100(1-\alpha)^{th}$ conditional percentile of V_i given $\vec{\mathbf{s}}$. Then a $100(1-\alpha)\%$ two-sided conditional prediction interval of s_i given $\vec{\mathbf{s}}$ is

$$\left(\hat{\tilde{s}}_i - V_{i,\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{s}_i)},\ \ \hat{\tilde{s}}_i - V_{i,1-\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{s}_i)}\right). \tag{4.17}$$

Intuitively, we may predict the conditional mean $\theta_i = \exp(\mu + s_i)$ using $\hat{\hat{\theta}}_i = \exp(\hat{\mu} + \hat{\tilde{s}}_i)$. And a $100(1-\alpha)\%$ two-sided conditional prediction interval of θ_i given $\vec{\mathbf{s}}$ can be constructed to be

$$PI^s_{i,\alpha} = \left(\exp\left(\hat{\mu} + \hat{\tilde{s}}_i - V_{i,\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{s}_i)}\right),\ \ \exp\left(\hat{\mu} + \hat{\tilde{s}}_i - V_{i,1-\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{s}_i)}\right)\right). \tag{4.18}$$

The value of $V_{i,\alpha/2}$ and $V_{i,1-\alpha/2}$ in (4.17) and (4.18) can be approximated from the following parametric bootstrap procedure:

1) Generate an $r \times c$ spatial grid based on the heuristic optimal co-clustering, with insect counts from trees in the co-clusters having independent distributions of $\textit{Negtive Binomial}(\exp(\hat{\mu} + \hat{\tilde{s}}_i), \hat{\kappa})$ $(i = 1, 2, \cdots, nm)$.

2) Fit the spatial GLMM based on the sampled tree locations to estimate (μ, σ^2, κ) as $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$. With $(\hat{\mu}, \hat{\sigma}^2, \hat{\kappa})$ replaced with $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\kappa}^*)$ in (4.15) and (4.16), calculate $\hat{\tilde{s}}_i^* = \text{eBLP}(s_i)$ and $\widehat{\text{MSE}}^*(\tilde{s}_i)$, and then $V_i^* = \left(\hat{\tilde{s}}_i^* - \hat{\tilde{s}}_i\right) \Big/ \sqrt{\widehat{\text{MSE}}^*(\tilde{s}_i)}$;.

3) Repeat Step 1–2 a number of times, say B times, to obtain $(V_i^{*(1)}, V_i^{*(2)}, \cdots, V_i^{*(B)})$, and approximate $V_{i,\alpha/2}$ and $V_{i,1-\alpha/2}$ by the $100(1-\alpha/2)^{th}$ and $100(\alpha/2)^{th}$ percentile of $(V_i^{*(1)}, V_i^{*(2)}, \cdots, V_i^{*(B)})$, respectively.

From Section 4.3.1, a $100(1-\alpha)\%$ two-sided conditional prediction interval of θ_i given $\vec{\mathbf{s}}$ can be constructed to be

$$PI_{i,\alpha}^{\theta} = \left(\hat{\tilde{\theta}}_i - U_{i,\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{\theta}_i)},\ \ \hat{\tilde{\theta}}_i - U_{i,1-\alpha/2}\sqrt{\widehat{\text{MSE}}(\tilde{\theta}_i)}\right). \tag{4.19}$$

To compare the coverage probability of (4.18) to that of (4.19), we performed a simulation study as follows:

1) Specify a design for the heuristic optimal co-clustering and set of model parameters. Generate the true co-cluster effects s_i's ($i = 1, 2, \cdots, M$, where M is the number of co-clusters of the heuristic optimal co-clustering), and record the true conditional means of co-clusters θ_i's.

2) Generate a spatial grid based on the heuristic optimal co-clustering, with insect counts from trees in the co-clusters having independent distributions of $Negtive\ Binomial(\theta_i, \kappa)$ $(i = 1, 2, \cdots, M)$.

3) Calculate both $PI^{\theta}_{i,\alpha}$ and $PI^{s}_{i,\alpha}$ for each co-cluster, and check whether the corresponding true conditional mean is captured or not. That is, we check whether θ_i falls in $PI^{\theta}_{i,\alpha}$ and/or $PI^{s}_{i,\alpha}$.

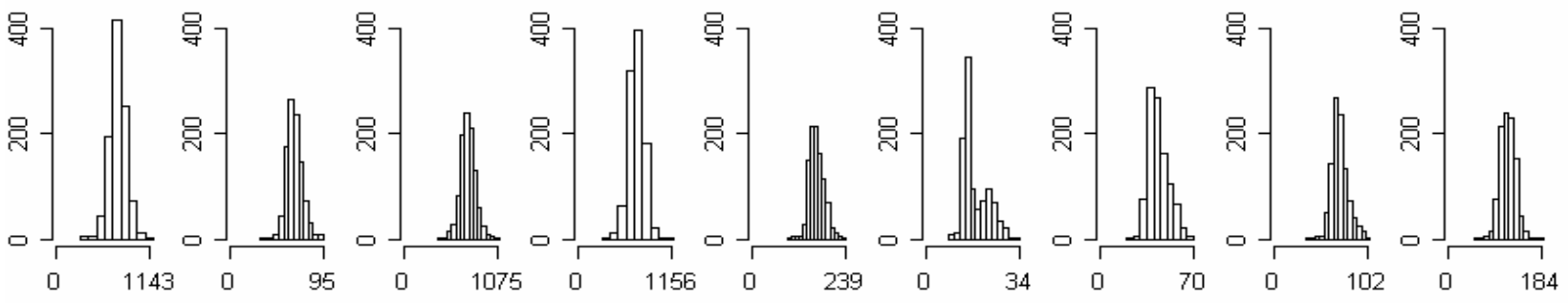
4) Repeat Step 2–3 a number of times, and the coverage probability of $PI^{\theta}_{i,\alpha}$ and $PI^{s}_{i,\alpha}$ for each co-cluster may be measured as the proportion of times that $PI^{\theta}_{i,\alpha}$ and $PI^{s}_{i,\alpha}$ capture the corresponding true conditional mean θ_i, respectively.

Table 4.2. Coverage Probability for $\kappa = 1$

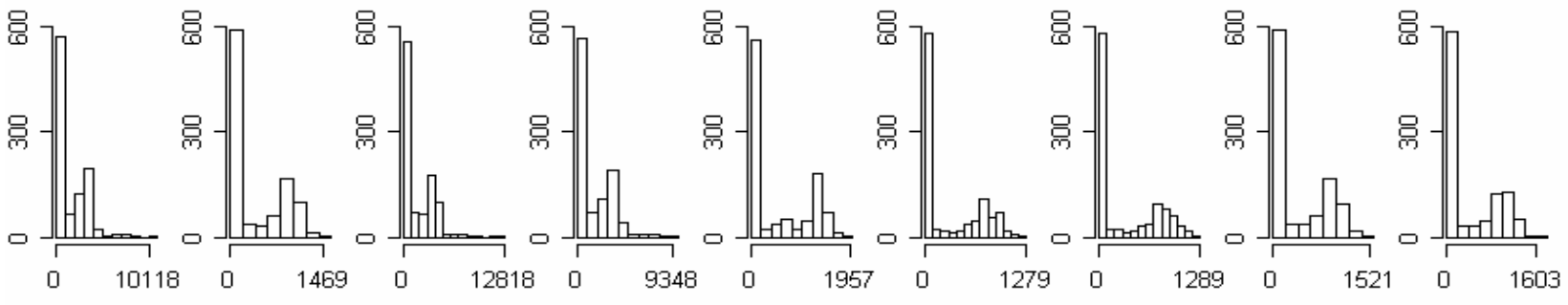
θ_i	2407	254	2646	2439	725	46	109	266	400
$PI^{\theta}_{i,\alpha}$	90.9%	92.5%	91.9%	91.9%	93.1%	95.6%	94.1%	92.0%	91.9%
$PI^{s}_{i,\alpha}$	35.9%	39.9%	36.3%	35.1%	42.7%	36.8%	37.4%	48.9%	44.6%

Table 4.3. Coverage Probability for $\kappa = 5$

θ_i	2407	254	2646	2439	725	46	109	266	400
$PI^{\theta}_{i,\alpha}$	96.1%	93.9%	94.6%	95.0%	93.1%	94.2%	93.1%	95.0%	92.8%
$PI^{s}_{i,\alpha}$	17.5%	23.8%	17.1%	17.5%	20.5%	16.8%	18.9%	70.3%	23.6%

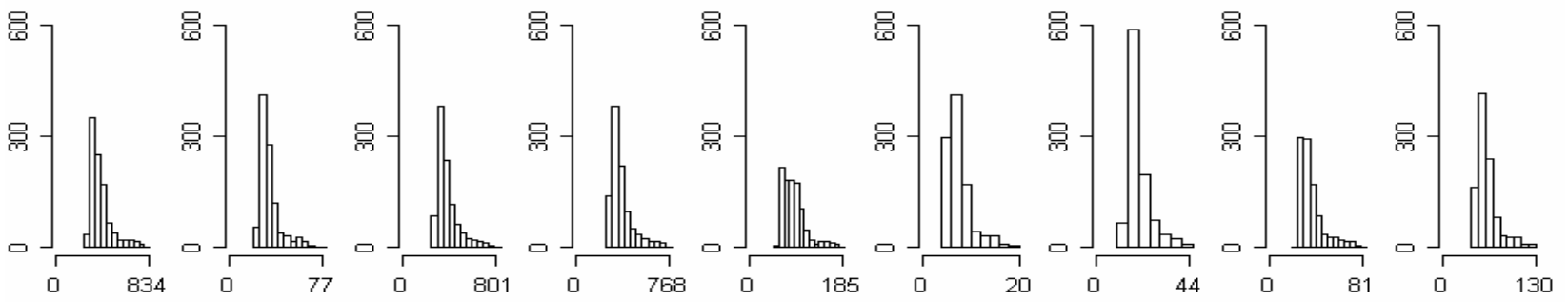


(a) $PI^{\theta}_{i,\alpha}$

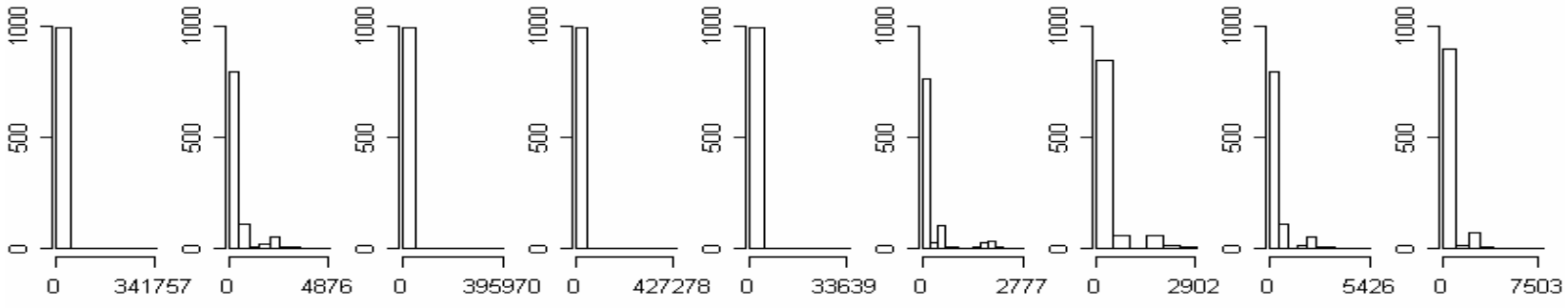


(b) $PI^{s}_{i,\alpha}$

Figure 4.5. Histogram of Conditional Prediction Interval Width for $\kappa = 1$



(a) $PI^{\theta}_{i,\alpha}$



(b) $PI^{s}_{i,\alpha}$

Figure 4.6. Histogram of Conditional Prediction Interval Width for $\kappa = 5$

Here, we considered a design $(10,17,13)\times(13,15,12)$ for the heuristic optimal co-clustering. We specified $\mu = 6$, $\sigma^2 = 2$ and different values for κ. $\alpha = 0.10$ was used for the conditional prediction interval, and $B = 1000$ resampling times were taken for the parametric bootstrap procedure. Table 4.2 shows the coverage probabilities of $PI^{\theta}_{i,\alpha}$ and $PI^{s}_{i,\alpha}$ for $\kappa = 1$, and Table 4.3 for $\kappa = 5$. We notice that the coverage probabilities of $PI^{\theta}_{i,\alpha}$ are satisfactorily close to nominal that is 90%, whereas most of the coverage probabilities of $PI^{s}_{i,\alpha}$ are pretty small. The histograms for the width of $PI^{\theta}_{i,\alpha}$ and $PI^{s}_{i,\alpha}$ are also shown in Figure 4.5 and Figure 4.6, from which we notice there is a very large variation for the width of $PI^{s}_{i,\alpha}$. Overall, $PI^{\theta}_{i,\alpha}$ performs better than $PI^{s}_{i,\alpha}$.

4.3.3. Simulation Study

To evaluate the proposed pest assessment procedure outlined in Section 4.3.1, we performed a simulation study as follows:

1) Simulate count data for an orchard based on a specified design and set of model parameters, and record the conditional means of true co-clusters θ_i's ($i = 1, 2, \cdots, M_0$), where M_0 is the number of true co-clusters.
2) Take a sample from the orchard based on the proposed sampling strategy. Apply the heuristic optimization algorithm to the orchard to search for the optimal co-clustering, and calculate $\hat{\tilde{\theta}}_j$ and $\widehat{\mathrm{MSE}}(\tilde{\theta}_j)$ for each co-cluster of the heuristic optimal

co-clustering ($j = 1, 2, \cdots, M$), where M is the number of co-clusters of the heuristic optimal co-clustering that may be different from M_0.

3) By comparing θ_i's ($i = 1, 2, \cdots, M_0$) with a pre-specified threshold θ_c, the true status of trees within the i^{th} true co-cluster is set to be "Treat" if $\theta_i > \theta_c$, and "Do Not Treat" otherwise. By comparing $L_{j,\alpha}$'s ($j = 1, 2, \cdots, M$) with θ_c, the decision status of trees within the j^{th} co-cluster of the heuristic optimal co-clustering is set to be "Treat" if $L_{j,\alpha} > \theta_c$, and "Do Not Treat" otherwise.

4) Investigate each tree of the orchard for consistency between its true status and decision status, and assign it into the corresponding combination of categories in a confusion matrix shown in Table 4.4. Count the number of trees ($d_{11}, d_{12}, d_{21}, d_{22}$) in the confusion matrix.

5) Repeat Step 1–4 a number of times, and update the confusion matrix by accumulating the values of ($d_{11}, d_{12}, d_{21}, d_{22}$).

Table 4.4. Confusion Matrix

		Decision	
		Do Not Treat	Treat
Truth	Do Not Treat	d_{11}	d_{12}
	Treat	d_{21}	d_{22}

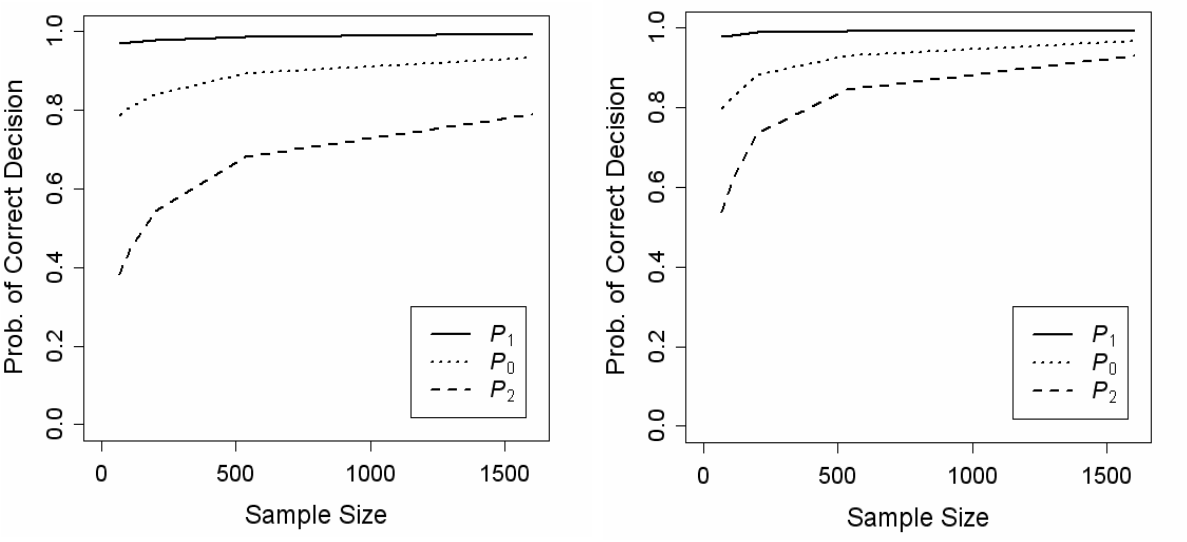
The probabilities of making correct decision may act as the measures to evaluate the pest assessment procedure:

$$P_1 = P(\text{Correct decision}|\text{Truth is "Do Not Treat"}) = d_{11} / (d_{11} + d_{12});$$

$$P_2 = P(\text{Correct decision}|\text{Truth is "Treat"}) = d_{22} / (d_{21} + d_{22});$$

$$P_0 = P(\text{Correct decision}) = (d_{11} + d_{22}) / (d_{11} + d_{12} + d_{21} + d_{22}).$$

The same simulation set-up as in Section 4.2.2.4 was used here. Based on the analyses discussed in Section 4.4, a critical economic threshold was set at $\theta_c = 500$. For this study, $\alpha = 0.05$ was considered for the prediction bound, and $B = 1000$ resampling times were taken for the parametric bootstrap procedure. The results are summarized in Figure 4.7, from which we notice all of P_0, P_1 and P_2 increase as the sample size increases given μ, σ^2 and κ, and increase as σ^2 increases given μ, κ and the sample size. Recall that previously we used the success rate of the design to evaluate the performance of heuristic optimization algorithm, which is a very conservative measure. For example, a co-clustering would be counted as a failure even if only one row or column is mistakenly assigned to a co-cluster that this row or column does not belong to. Figure 4.7 shows that the overall probability of making correct decision, P_0, is relatively high, even for a small sample size and small σ^2 such as $\sigma^2 = 0.2$, which further demonstrates the practical utility of our proposed pest assessment procedure.



(a) $\sigma^2 = 0.2$ (b) $\sigma^2 = 1$

Figure 4.7. Probabilities of Correct Decision

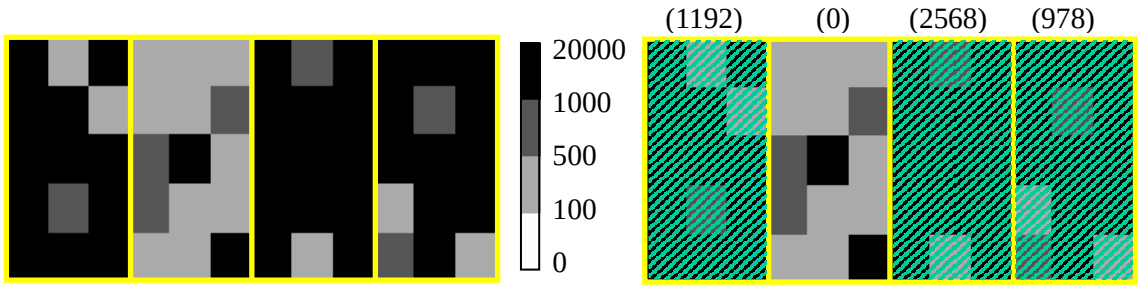
4.4. EXAMPLE

Persea mite (*Oligonychus perseae*) is an avocado leaf feeding pest that is native to Mexico and is a serious invasive pest in California (USA), Costa Rica, Israel, and Spain (Hoddle 2005). When pest populations build to sufficiently high densities leaves begin to drop from trees. To avoid premature leaf dropping some type of control procedure may be warranted (e.g., pesticide applications, or releases of commercially available natural enemies, like predatory mites that eat the pest).

Mite counts were determined during the Summer of 2009 from sampled trees in three commercial avocado orchards located in California, USA. Trees in the orchards were

planted on relatively flat terrain according to a grid system consisting of rows and columns. Sampled trees from orchard A were arranged on a 5×12 grid, from orchard B on a 5×6 grid, and from orchard C on a 5×6 grid. Eight leaves were collected from each tree, and summing up the number of mites provided a pest count for each sampled tree.

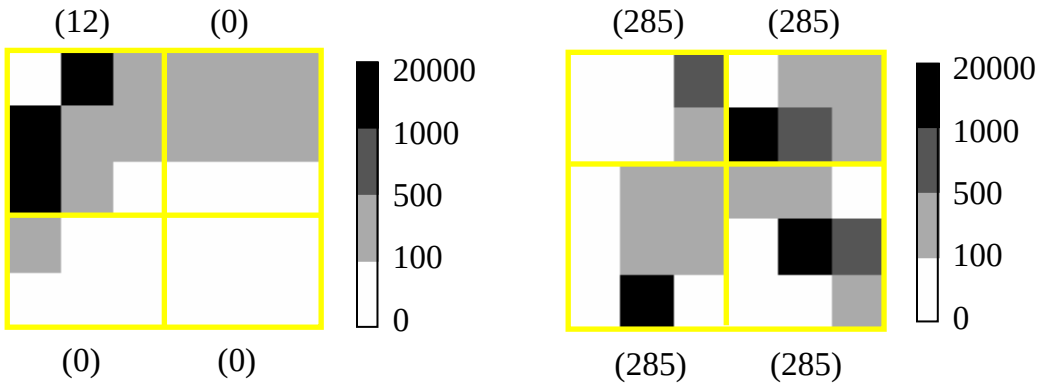
Applying the heuristic optimization algorithm to orchard A, we obtained the heuristic optimal co-clustering as shown in Figure 4.8(a), in which four co-clusters are separated by the solid lines. Here the minimum co-cluster size was set to be $r_0 \times c_0 = 2\times3$. We then compared the 95% lower conditional prediction bound of the conditional mean for each co-cluster to an established threshold of $\theta_c = 500$ (see Maoz et al. 2011). Figure 4.8(b) shows three (shaded) co-clusters that require treatment.



(a) Heuristic Optimal Co-clustering (b) Pest Treatment Decision

Figure 4.8. Pest Assessment for Orchard A (The value in parentheses next to each co-cluster is the corresponding lower conditional prediction bound of the conditional mean.)

Applying the same pest assessment procedure to orchard B, we obtained the heuristic optimal co-clustering as shown in Figure 4.9(a), in which four co-clusters are separated by the solid lines and none of them requires treatment. Similarly, the heuristic optimal co-clustering for orchard C is shown in Figure 4.9(b), in which four co-clusters are separated by the solid lines and none of them requires treatment.

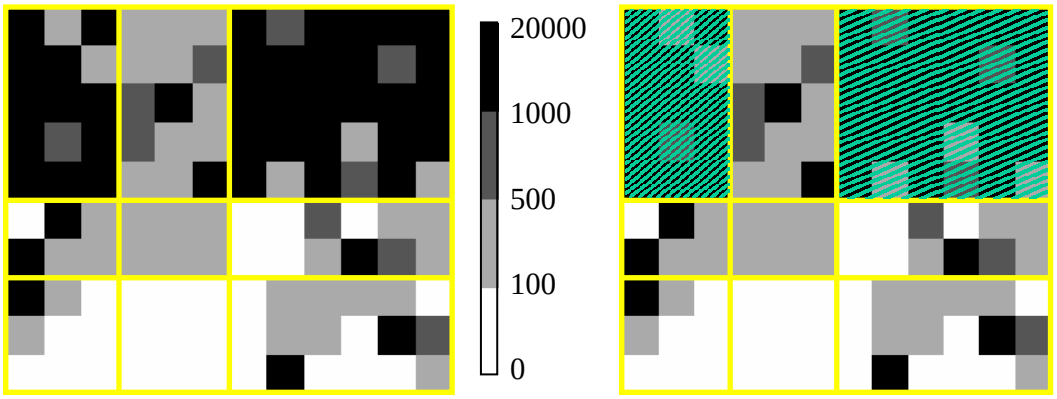


(a) Orchard B (b) Orchard C

Figure 4.9. Pest Assessment for Orchards B and C (The value in parentheses next to each co-cluster is the corresponding lower conditional prediction bound of the conditional mean.)

The analyses shown in Figure 4.8 and Figure 4.9 motivate us to anticipate which regions should be identified as infested if we were to merge orchards A, B and C and analyze them as one larger orchard. By combining orchards A, B and C with orchard A being on the top, orchard B on the bottom left and orchard C on the bottom right, we built

a synthetic integrated orchard, called orchard D, that contains 120 trees on a 10×12 grid. Applying the pest assessment procedure to orchard D, we obtained the heuristic optimal co-clustering as shown in Figure 4.10(a), in which nine co-clusters are separated by the solid lines. The pest treatment decision was then made that the two shaded co-clusters located within orchard A require treatment as shown in Figure 4.10(b), which is consistent with the results from analyzing orchards A, B and C one at a time. Again we specified the minimum co-cluster size to be $r_0\times c_0 = 2\times3$, and used $\theta_c = 500$.



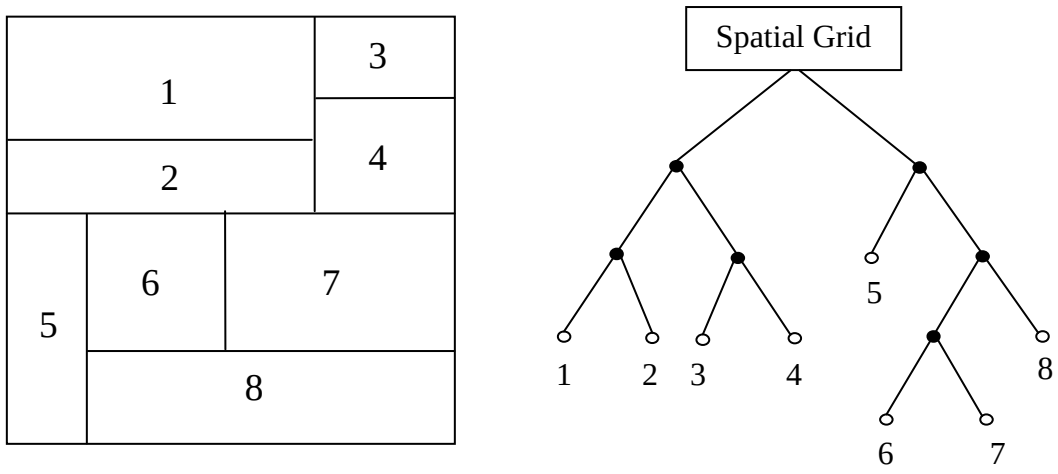
(a) Heuristic Optimal Co-clustering (b) Pest Treatment Decision

Figure 4.10. Pest Assessment for Orchard D (Integrated Orchard)

4.5. DISCUSSION

Our proposed model-based co-clustering method showed a significant utility and power in searching for the optimal co-clustering on a spatial grid. Combining the spatial

co-clustering technique with a statistical inference method, our proposed pest assessment procedure also showed an excellent performance in identifying the infested regions within orchards. Only treating the infested regions instead of the whole orchard can reduce pest management costs and minimize potential hazards to the environment. Although these methods were developed to analyze the pest data collected from perennial tree orchards (i.e., avocado orchards), we anticipate that this general approach will have utility for a wide range of investigations involving spatial information.



(a) Spatial Grid Representation (b) Tree Representation

Figure 4.11. “Tree” Co-cluster Structure

In this chapter, we considered the spatial GLMM with correlation within co-clusters, and all the co-clusters are independent to each other. Although this assumption makes

much practical sense with our application, we will further consider a spatial GLMM with both correlation within co-clusters and correlation between co-clusters as future work.

Furthermore, more flexible co-cluster structures will be investigated for the spatial grid in future work, such as the “tree” co-cluster structure. An example of the “tree” co-cluster structure is illustrated in Figure 4.11, in which the spatial grid contains eight co-clusters that can reproduce the original spatial grid through a sequence of leaf-to-root combinations as shown in Figure 4.11(b). The “Tree” co-cluster structure has been applied to the voting data in Hartigan (1972), which is considered the first co-clustering paper.

Chapter 5

Discussion

Clustering is rapidly becoming a powerful data mining technique. In Chapter 2 and 3, our proposed hierarchical clustering and co-clustering procedure showed a significant utility and power in handling a data matrix of scatter plots. In Chapter 4, we developed a model-based co-clustering method for spatial data. Specifically, the proposed pest assessment procedure that combines the spatial co-clustering with a statistical inference makes assessment of pest density more accurate.

Furthermore, extensive literature has shown a variety of clustering methods and their applications in many domains. Depending on how data are organized and how a cluster is defined, more clustering techniques may be potentially developed in the future to satisfy various needs in applications.

5.1. VECTOR-BASED CLUSTERING

In Chapter 2 and 3, the difference between a pair of scatter plots is directly measured by a quality index between the pair of corresponding bivariate distributions. Another possibility is to characterize scatter plots individually such that a scatter plot can be

represented by a vector of characteristics. Then the dissimilarity among scatter plots can be measured by comparing the corresponding vectors of characteristics.

Consider a scatter plot that is regarded as the sample from a bivariate distribution followed by $(X,Y)'$. Usually, it is insufficient to use a single characteristic such as the Pearson correlation coefficient, which only measures the strength and direction of a linear relationship between X and Y. Functional models offer an attractive tool that is flexible enough to capture a wide variety of non-linear association. The simplest function model is polynomial regression model and we use it for illustration purpose for the discussion below.

For each scatter plot, a low-order polynomial model may be fit:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i ,$$

where y_i is the value of Y for the i^{th} observation, x_i is the value of X for the i^{th} observation, and ε_i is the error term with $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ ($i = 1,2,\cdots,n$, where n is the number of observations).

To achieve the optimal fitting, we sequentially test three null hypotheses $H_{01}: \beta_3 = 0$, $H_{02}: \beta_2 = 0$, and $H_{03}: \beta_1 = 0$. That is, we test each null hypothesis only if we do not reject the preceding one. All the possible cases are illustrated in Figure 5.1, each of which would produce an estimate of the vector $\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ that act as characteristics of the relationship between X and Y. For example, in the case shown in

Figure 5.1(c), we do not reject $H_{01}: \beta_3 = 0$ but reject $H_{02}: \beta_2 = 0$. Then we fit a second-order polynomial regression model to estimate the parameters β_0, β_1 and β_2 by $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Therefore, the vector $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, 0)'$ can be used to characterize this scatter plot.

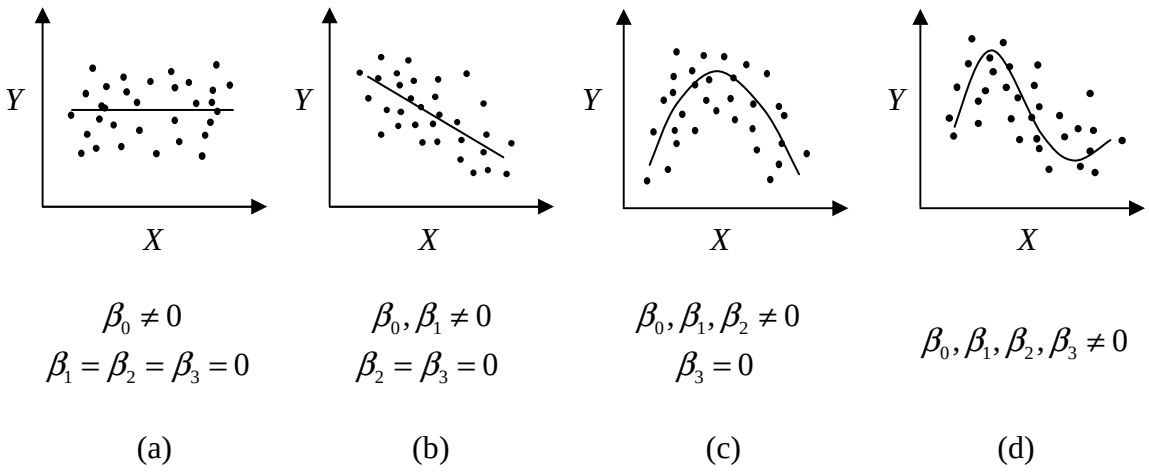


Figure 5.1. Polynomial Regression Model

If polynomial regression models do not fit the data well, a nonparametric approach may be used. For example, we divide the plotting region (such as the space covering all the possible values of X and Y, i.e., the space with borders at positions of $X_{\min}$, $X_{\max}$, $Y_{\min}$, and $Y_{\max}$ as shown in Figure 5.2) into a number of subplots and count the number of observations within each subplot. An example with 4 subplots Q_1, Q_2, Q_3 and Q_4 is presented in Figure 5.2, in which the center of the plotting region is used for

dividing the plotting region along both axes. By denoting the number of observations within the subplot Q_i by d_i ($i = 1, 2, 3, 4$), the relationship between X and Y is then characterized by the vector $\vec{d} = (d_1, d_2, d_3, d_4)'$.

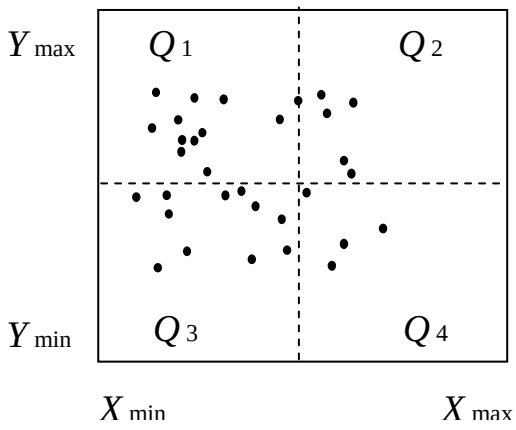


Figure 5.2. A Nonparametric Approach

To cluster the data matrix of scatter plots, we may define a one-dimensional or multi-dimensional objective function based on vectors of characteristics, either $\vec{\beta}$ or $\vec{d}$, to measure the quality of a specific clustering, and the optimal clustering would be the one that optimizes this objective function. The vector-based clustering method will be further investigated in future work. A straightforward approach could be to define a one-dimensional objective function that can be incorporated into the current clustering methods.

5.2. EXTENSION TO MULTI-DIMENSIONAL CLUSTERING

To obtain a greater understanding of gene expression regulation, host-microbe interactions, and to track and predict infectious disease outbreaks, it will be necessary to identify many of the associations among different variables. Consider the data that are arranged in a multi-way contingency table with each cell being a real number, for which we would like to simultaneously cluster all the dimensions. Some recent literature reflects efforts to extend co-clustering methods to multi-dimensional contexts. However, multi-dimensional clustering has not been well studied.

Information-theoretic clustering is a statistically based clustering technique with apparent flexibility to be applied in complicated cases such as multi-dimensional contexts. To extend the information-theoretic co-clustering to the general multi-dimensional case, we have to define mutual information for cases with more than two random variables. Consider three variables X, Y and Z. Several generalizations of two-way mutual information have been proposed, which are listed as follows:

1) Total Correlation (Watanabe 1960):

$$I_C(X;Y;Z) = \sum_x \sum_y \sum_z p(x,y,z) log \frac{p(x,y,z)}{p(x)p(y)p(z)}$$

2) Mutual Information (Yeung 1991):

$$I_M(X;Y;Z) = I(X;Y) - I(X;Y \mid Z) = I(X;Y) - E_Z[I(X;Y) \mid Z]$$
$$= \sum_x \sum_y \sum_z \left\{ p(x,y,z) log \frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)} \right\}$$

3) Interaction Information (McGill 1954):

$$I_I(X;Y;Z) = I(X;Y \mid Z) - I(X;Y) = E_Z[I(X;Y) \mid Z] - I(X;Y)$$
$$= \sum_x \sum_y \sum_z \left\{ p(x,y,z) log \frac{p(x)p(y)p(z)p(x,y,z)}{p(x,y)p(x,z)p(y,z)} \right\}$$

Notice that $I_I(X;Y;Z)$ is identical to $I_M(X;Y;Z)$ except for a change in sign. Therefore, we only focus on multi-dimensional clustering using either $I_C(X;Y;Z)$ or $I_M(X;Y;Z)$ in the following discussion.

It is trivial to prove $I_C(X;Y;Z) \geq 0$. By using I_C, the optimal three-dimensional clustering is the one that leads to the largest mutual information among the cluster random variables, $I_C(\hat{X};\hat{Y};\hat{Z})$, or equivalently, one that minimizes the difference (loss) between the mutual information among the original random variables and the mutual information among the cluster random variables, $I_C(X;Y;Z) - I_C(\hat{X};\hat{Y};\hat{Z})$.

$I_M(X;Y;Z)$ may be negative, whose interpretation is that the mutual information between any two of the random variables X, Y and Z increases when the other random variable is given, that is, any one of the random variables X, Y and Z affects the dependency between the other two random variables. By using I_M, the optimal three-dimensional clustering is the one that minimizes the difference (absolute value) between the mutual information among the original random variables and the mutual information among the cluster random variables, $\left| I_M(X;Y;Z) - I_M(\hat{X};\hat{Y};\hat{Z}) \right|$.

Our preliminary examples have shown that neither I_C nor I_M performs consistently better than the other one. In some examples we have seen, I_M is able to achieve the real optimal clustering whereas I_C has not ever been able to do that yet. An example for which both I_C and I_M fail to achieve the real optimal clustering is illustrated below. Figure 5.3 shows a joint probability distribution $p(x, y, z)$ among X, Y and Z each taking three levels of values, where X, Y and Z are pairwise independent but not mutually independent. Suppose two clusters are to be obtained for each dimension. By observing the data pattern, the real optimal clustering should be $\hat{x} = \{\{x_1\}, \{x_2, x_3\}\}$, $\hat{y} = \{\{y_1\}, \{y_2, y_3\}\}$ and $\hat{z} = \{\{z_1\}, \{z_2, z_3\}\}$. However, by using either I_C or I_M, the obtained optimal clustering is $\hat{x} = \{\{x_1\}, \{x_2, x_3\}\}$, $\hat{y} = \{\{y_1, y_2\}, \{y_3\}\}$ and $\hat{z} = \{\{z_1\}, \{z_2, z_3\}\}$.

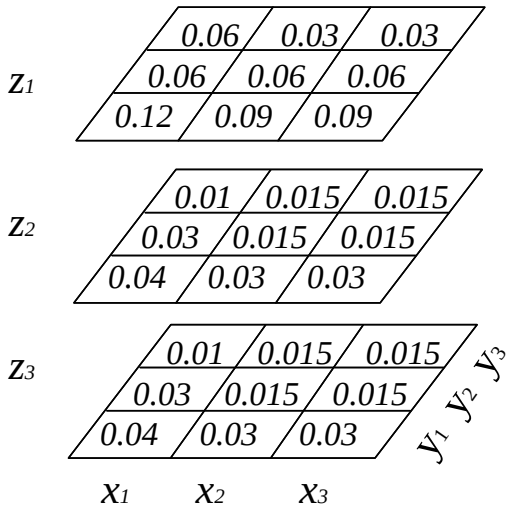


Figure 5.3. A Three-dimensional Example

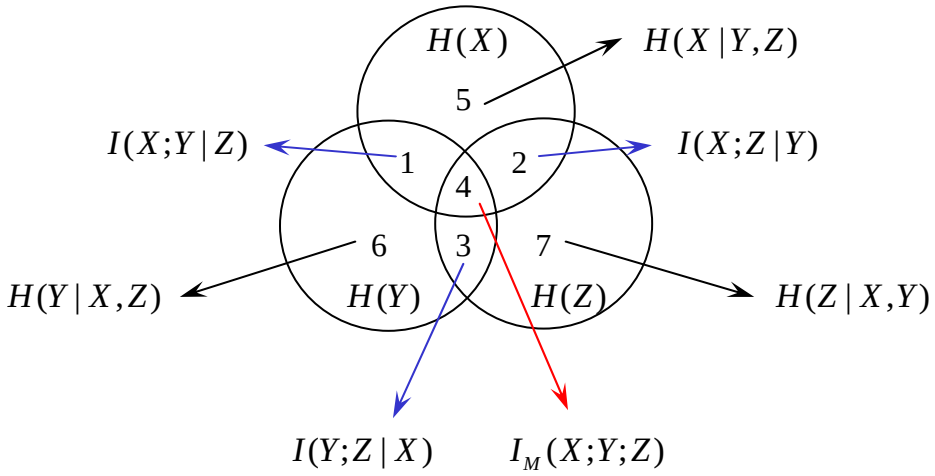


Figure 5.4. Information-theoretic Measures

Since the performance of existing definitions of multi-way mutual information is not consistent, we are led to seek other information measures to carry out information-theoretic multi-dimensional clustering. Analogous to Venn Diagrams in set theory, information-theoretic measures can be geometrically represented for the three-dimensional case as shown in Figure 5.4 (Yeung 1991), which may be used to motivate alternative information-theoretic measures, including for example, the following alternative measure of mutual information:

$$I_D(X;Y;Z) = I(X;Y|Z) + I(X;Z|Y) + I(Y;Z|X) + I_M(X;Y;Z)$$
$$= \sum_{x,y,z} \left\{ p(x,y,z) log \frac{[p(x,y,z)]^2}{p(x,y)\,p(x,z)\,p(y,z)} \right\}$$

It is easy to prove $I_D(X;Y;Z) \geq 0$. By adopting the associated criterion for defining the optimal clustering, a three-dimensional clustering method can be studied in future work.

Bibliography

Abdullah, A. and Hussain, A. (2006), “A New Biclustering Technique Based on Crossing Minimization,” *Neurocomputing*, 69, 1882-1896.

Agnholt, J., Kelsen, J., Schack, L., Hvas, C. L., Dahlerup, J. F., and Sørensen, E. S. (2007), “Osteopontin, a Protein with Cytokine-like Properties, is Associated with Inflammation in Crohn's Disease,” *Scandinavian Journal of Immunology*, 65(5), 453-460.

Aguilar-Ruiz, J. S. and Divina, F. (2005), “Evolutionary Biclustering of Microarray Data,” *EvoWorkshops ’05,* LNCS 3449, 1-10.

Ahrenstedt, O., Knutson, L., Nilsson, B., Nilsson-Ekdahl, K., Odlind, B., and Hällgren, R. (1990), “Enhanced Local Production of Complement Components in the Small Intestines of Patients with Crohn's Disease,” *The New England Journal of Medicine*, 322(19), 1345–1349.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997), “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs,” *Nucleic Acids Research*, 25(17), 3389-3402.

Andreopoulos, B., An, A., Wang, X., and Schroeder, M. (2009), “A Roadmap of Clustering Algorithms: Finding a Match for a Biomedical Application,” *Briefings in Bioinformatics*, 10(3), 297-314.

Baehrecke, E. H., Dang, N., Babaria, K., and Shneiderman, B. (2004), “Visualization and Analysis of Microarray and Gene Ontology Data with Treemaps,” *BMC Bioinformatics*, 5:84.

Bandyopadhyay, S. K., de la Motte, C. A., Kessler, S. P., Hascall, V. C., Hill, D. R., and Strong, S. A. (2008), “Hyaluronan-mediated Leukocyte Adhesion and Dextran Sulfate Sodium-induced Colitis are Attenuated in the Absence of Signal Transducer and Activator of Transcription 1,” *The American Journal of Pathology*, 173(5), 1361-1368.

Banerjee, A., Dhillon, I. S., Ghosh, J., Merugu, S., and Modha, D. S. (2007), “A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation,” *Journal of Machine Learning Research*, 8, 1919-1986.

Barchia, I. M., Herron, G. A., and Gilmour, A. R. (2003), "Use of a Generalized Linear Mixed Model to Reduce Excessive Heterogeneity in Petroleum Spray Oil Bioassay Data," *Journal of Economic Entomology*, 96(3), 983-989.

Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. (2006), "BicAT: A Biclustering Analysis Toolbox," *Bioinformatics*, 22(10), 1282-1283.

Baumgart, M., Dogan, B., Rishniw, M., Weitzman, G., Bosworth, B., Yantiss, R., Orsi, R. H., Wiedmann, M., McDonough, P., Kim, S. G., et al. (2007), "Culture Independent Analysis of Ileal Mucosa Reveals a Selective Increase in Invasive Escherichia Coli of Novel Phylogeny Relative to Depletion of Clostridiales in Crohn's Disease Involving the Ileum," *The ISME Journal*, 1(5), 403-418.

Bekkerman, R., El-Yaniv, R., and McCallum, A. (2005), "Multi-Way Distributional Clustering via Pairwise Interactions," *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, 41-48.

Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2002), "Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem," *Proceedings of the Sixth International Conference on Computational Biology (RECOMB '02)*, 49-57.

Bennett, K. E., Hopper, J. E., Stuart, M. A., West, M., and Drolet, B. S. (2008), "Blood-Feeding Behavior of Vesicular Stomatitis Virus Infected *Culicoides Sonorensis* (Diptera: Ceratopogonidae)," *Journal of Medical Entomology*, 45(5), 921-926.

Bianchi, F. J. J. A., Goedhart, P. W., and Baveco, J. M. (2008), "Enhanced Pest Control in Cabbage Crops Near Forest in The Netherlands," *Landscape Ecology*, 23(5), 595-602.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88(421), 9-25.

Broedl, U. C., Schachinger, V., Lingenhel, A., Lehrke, M., Stark, R., Seibold, F., Göke, B., Kronenberg, F., Parhofer, K. G., and Konrad-Zerna, A. (2007), "Apolipoprotein A-IV is an Independent Predictor of Disease Activity in Patients with Inflammatory Bowel Disease," *Inflammatory Bowel Diseases*, 13, 391-397.

Busygin, S., Jacobsen, G., and Kramer, E. (2002), "Double Conjugated Clustering Applied to Leukemia Microarray Data," *SIAM Data Mining Workshop on Clustering High Dimensional Data and its Applications*.

Busygin, S., Prokopyev, O., and Pardalos, P. M. (2008), "Biclustering in Data Mining," *Computers & Operations Research*, 35(9), 2964-2987.

Cai, R., Lu, L., and Hanjalic, A. (2008), "Co-clustering for Auditory Scene Categorization," *IEEE Transactions on Multimedia*, 10(4), 596-606.

Califano, A., Stolovitzky, G., and Tu, Y. (2000), "Analysis of Gene Expression Microarrays for Phenotype Classification," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, 8, 75-85.

Candy, S. G. (2000), "The Application of Generalized Linear Mixed Models to Multi-level Sampling for Insect Population Monitoring," *Environmental and Ecological Statistics*, 7(3), 217-238.

Cheng, Y. and Church, G. M. (2000), "Biclustering of Expression Data," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, 8, 93-103.

Chiaravalloti, A. D., Greco, G., Guzzo, A., and Pontieri, L. (2006), "An Information-Theoretic Framework for High-Order Co-clustering of Heterogeneous Objects," *ECML'06*, LNAI 4212, 598-605.

Cho, H. and Dhillon, I. S. (2008), "Coclustering of Human Cancer Microarrays Using Minimum Sum-Squared Residue," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3), 385-400.

Cho, H., Dhillon, I. S., Guan, Y., and Sra, S. (2004), "Minimum Sum-Squared Residue Co-Clustering of Gene Expression Data," *Proceedings of the Fourth SIAM International Conference on Data Mining*, 114-125.

Chopra, P., Kang, J., Yang, J., Cho, H., Kim, H. S., and Lee, M. G. (2008), "Microarray Data Mining Using Landmark Gene-guided Clustering," *BMC Bioinformatics*, 9:92.

Conyers, G., Milks, L., Conklyn, M., Showell, H., and Cramer, E. (1990), "A Factor in Serum Lowers Resistance and Opens Tight Junctions of MDCK Cells," *American Journal of Physiology*, 259(4), C577-585.

de la Motte, C. A., Hascall, V. C., Drazba, J., Bandyopadhyay, S. K., and Strong, S. A. (2003), "Mononuclear Leukocytes Bind to Specific Hyaluronan Structures on Colon Mucosal Smooth Muscle Cells Treated with Polyinosinic Acid:Polycytidylic Acid: Inter-alpha-trypsin Inhibitor is Crucial to Structure and Function," *The American Journal of Pathology*, 163(1), 121-133.

Deodhar, M. and Ghosh, J. (2007), "A Framework for Simultaneous Co-clustering and Learning from Complex Data," *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, V, 250-259.

Dhillon, I. S. (2001), "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269-274.

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003), "Information-Theoretic Co-clustering," *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 89-98.

Divina, F. and Aguilar-Ruiz, J. S. (2007), "A Multi-Objective Approach to Discover Biclusters in Microarray Data," *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO'07)*, 1, 385-392.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95(25), 14863-14868.

Elias, S. P., Lubelczyk, C. B., Rand, P. W., Lacombe, E. H., Holman, M. S., and Smith, R. P. (2006), "Deer Browse Resistant Exotic-Invasive Understory: An Indicator of Elevated Human Risk of Exposure to *Ixodes scapularis* (Acari: Ixodidae) in Southern Coastal Maine Woodlands," *Journal of Medical Entomology*, 43(6), 1142-1152.

Elston, D. A., Moss, R., Boulinier, T., Arrowsmith, C., and Lambin, X. (2001), "Analysis of Aggregation, a Worked Example: Numbers of Ticks on Red Grouse Chicks," *Parasitology*, 122(5), 563-569.

Fagerberg, U. L., Lööf, L., Lindholm, J., Hansson, L. O., and Finkel, Y. (2007), "Fecal Calprotectin: a Quantitative Marker of Colonic Inflammation in Children with Inflammatory Bowel Disease," *Journal of Pediatric Gastroenterology and Nutrition*, 45(4), 414–420.

Foell, D., Kucharzik, T., Kraft, M., Vogl, T., Sorg, C., Domschke, W., and Roth, J. (2003), "Neutrophil Derived Human S100A12 (EN-RAGE) is Strongly Expressed During Chronic Active Inflammatory Bowel Disease," *Gut*, 52(6), 847-853.

Foell, D., Wittkowski, H., and Roth, J. (2009), "Monitoring Disease Activity by Stool Analyses: From Occult Blood to Molecular Markers of Intestinal Inflammation and Damage," *Gut*, 58(6), 859-868.

Gao, B., Liu, T-Y., Zheng, X., Cheng, Q-S., and Ma, W-Y. (2005), "Consistent Bipartite Graph Co-Partitioning for Star-Structured High-Order Heterogeneous Data Co-Clustering," *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 41-50.

Gasch, A. P. and Eisen, M. B. (2002), "Exploring the Conditional Coregulation of Yeast Gene Expression Through Fuzzy k-Means Clustering," *Genome Biology*, 3(11), research0059.

Getz, G., Levine, E., and Domany, E. (2000), "Coupled Two-Way Clustering Analysis of Gene Microarray Data," *Proceedings of the National Academy of Sciences*, 97(22), 12079-12084.

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. (2008), "CCA: An R Package to Extend Canonical Correlation Analysis," *Journal of Statistical Software*, 23(12):1-14.

Gotway, C. A., and Stroup, W. W. (1997), "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2), 157-178.

Gozé, E., Nibouche, S., and Deguine, J-P. (2003), "Spatial and Probability Distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in Cotton: Systematic Sampling, Exact Confidence Intervals and Sequential Test," *Environmental Entomology*, 32(5), 1203-1210.

Greenstein, A. J., Sachar, D. B., Panday, A. K., Dikman, S. H., Meyers, S., Heimann, T., Gumaste, V., Werther, J. L., and Janowitz, H. D. (1992), "Amyloidosis and Inflammatory Bowel Disease. A 50-year Experience with 25 Patients," *Medicine* (Baltimore), 71, 261-270.

Halstensen, T. S., and Brandtzaeg, P. (1991), “Local Complement Activation in Inflammatory Bowel Disease,” *Immunologic Research*, 10, 485-492.

Halstensen, T. S., Mollnes, T. E., Garred, P., Fausa, O., and Brandtzaeg, P. (1992), “Surface Epithelium Related Activation of Complement Differs in Crohn's Disease and Ulcerative Colitis,” *Gut*, 33(7), 902-908.

Hansen, J. J., Holt, L., and Sartor, R. B. (2009), “Gene Expression Patterns in Experimental Colitis in IL-10-deficient Mice,” *Inflammatory Bowel Diseases*, 15, 890-899.

Hartigan, J. A. (1972), “Direct Clustering of a Data Matrix,” *Journal of the American Statistical Association*, 67(337), 123-129.

Hochbaum, D. S. and Shmoys, D. B. (1985), “A Best Possible Heuristic for the k-Center Problem,” *Mathematics of Operations Research*, 10(2), 180-184.

Hoddle, M. S. (2005), “Invasions of Leaf Feeding Arthropods: Why Are So Many New Pests Attacking California-Grown Avocados?” *California Avocado Society Yearbook* (2004-2005), 87, 65-81.

Holm, S. (1979), “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics*, 6(2), 65-70.

Huang, Z. (1998), “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery*, 2, 283-304.

Ifoulis, A. A., and Savopoulou-Soultani, M. (2006), “Use of Geostatistical Analysis to Characterize the Spatial Distribution of *Lobesia botrana* (Lepidoptera: Tortricidae) Larvae in Northern Greece,” *Environmental Entomology*, 35(2), 497-506.

Ihmels, J., Bergmann, S., and Barkai, N. (2004), “Defining Transcription Modules Using Large-scale Gene Expression Data,” *Bioinformatics*, 20(13), 1993-2003.

Jiang, D., Tang, C., and Zhang, A. (2004), “Cluster Analysis for Gene Expression Data: A Survey,” *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1370-1386.

Kaplan, N., Friedlich, M., Fromer, M., and Linial, M. (2004), “A Functional Hierarchical Organization of the Protein Sequence Space,” *BMC Bioinformatics*, 5:196.

Kaufman, L. and Rousseeuw, P. J. (1987), “Clustering by Means of Medoids,” *Statistical Data Analysis Based on the L1-Norm and Related Methods*, edited by Y. Dodge, North-Holland, 405-416.

Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008), “Techniques for Clustering Gene Expression Data,” *Computers in Biology and Medicine*, 38(3), 283-293.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003), “Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions,” *Genome Research*, 13(4), 703-716.

Kriegel, H-P., Kroger, P., and Zimek, A. (2009), “Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering,” *ACM Transactions on Knowledge Discovery from Data*, 3(1), 1-58.

Kruidenier, L., MacDonald, T. T., Collins, J. E., Pender, S. L., and Sanderson, I. R. (2006), “Myofibroblast Matrix Metalloproteinases Activate the Neutrophil Chemoattractant CXCL7 from Intestinal Epithelial Cells,” *Gastroenterology*, 130(1), 127-136.

Kung, S. Y., Mak, M-W., and Tagkopoulos, I. (2005), “Multi-Metric and Multi-Substructure Biclustering Analysis for Gene Expression Data,” *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, 123-134.

Lachmann, H. J., Goodman, H. J., Gilbertson, J. A., Gallimore, J. R., Sabin, C. A., Gillmore, J. D., and Hawkins, P. N. (2007), “Natural History and Outcome in Systemic AA Amyloidosis,” *The New England Journal of Medicine*, 356, 2361-2371.

Larsson, A. E., Melgar, S., Rehnström, E., Michaëlsson, E., Svensson, L., Hockings, P., and Olsson, L. E. (2006), “Magnetic Resonance Imaging of Experimental Mouse Colitis and Association with Inflammatory Activity,” *Inflammatory Bowel Diseases*, 12, 478–485.

Laufer, J., Oren, R., Goldberg, I., Horwitz, A., Kopolovic, J., Chowers, Y., and Passwell, J. H. (2000), “Cellular Localization of Complement C3 and C4 Transcripts in Intestinal Specimens from Patients with Crohn's Disease,” *Clinical and Experimental Immunology*, 120(1), 30-37.

Lazzeroni, L. and Owen, A. (2002), “Plaid Models for Gene Expression Data,” *Statistica Sinica*, 12(1), 61-86.

Li, J. X. (2004), “Visualization of High-dimensional Data with Relational Perspective Map, ” *Information Visualization*, 3(1), 49-59.

Li, X., LeBlanc, L., Elashoff, D., Borneman, J., Goodglick, L., and Braun, J. (2010), “Detecting Disease-related Biological Neighborhoods by Human Mucosal Interface Metaproteome Analysis,” *Gastroenterology*, 138, S-13.

Liu, R. Y. (1990), “On a Notion of Data Depth Based on Random Simplices,” *The Annals of Statistics*, 18, 405-414.

Liu, R. Y. and Singh, K. (1993), “A Quality Index Based on Data Depth and Multivariate Rank Tests,” *Journal of the American Statistical Association*, 88(421), 252-260.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), “Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference,” *The Annals of Statistics*, 27(3), 783-858.

Loewenstein, Y., Portugaly, E., Fromer, M., and Linial, M. (2008), “Efficient Algorithms for Accurate Hierarchical Clustering of Huge Datasets: Tackling the Entire Protein Space,” *Bioinformatics*, 24(13):i41-9.

Lonardi, S., Szpankowski, W., and Yang, Q. (2006), “Finding Biclusters by Random Projections,” *Theoretical Computer Science*, 368(3), 217-230.

Long, B., Zhang, Z. M., and Yu, P. S. (2005), “Co-clustering by Block Value Decomposition,” *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 635-640.

Lustgarten, J. L., Kimmel, C., Ryberg, H., and Hogan, W. (2008), “EPO-KB: A Searchable Knowledge Base of Biomarker to Protein Links,” *Bioinformatics*, 24(11), 1418-1419.

MacQueen, J. B. (1967), “ Some Methods for Classification and Analysis of Multivariate Observations,” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1, 281-297.

Madeira, S. C. and Oliveira, A. L. (2004), “Biclustering Algorithms for Biological Data Analysis: A Survey,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 24-45.

Madeira, S. C. and Oliveira, A. L. (2005), “A Linear Time Biclustering Algorithm for Time Series Gene Expression Data,” *Proceedings of the Fifth Workshop on Algorithms in Bioinformatics*, 39-52.

Mahalanobis, P. C. (1936), “On the Generalized Distance in Statistics,” *Proceedings of the National Academy of India*, 12, 49-55.

Maoz, Y., Gal, S., Zilberstein, M., Izhar, Y., Alchanatis, V., Coll, M., and Palevsky, E. (2011), “Determining an Economic Injury Level for the Persea Mite, *Oligonychus perseae*, a New Pest of Avocado in Israel,” *Entomologia Experimentalis et Applicata*, 138(2), 110-116.

Martinez-Medina, M., Aldeguer, X., Gonzalez-Huix, F., Acero, D., and Garcia-Gil, L. J. (2006), “Abnormal Microbiota Composition in the Ileocolonic Mucosa of Crohn's Disease Patients as Revealed by Polymerase Chain Reaction-denaturing Gradient Gel Electrophoresis,” *Inflammatory Bowel Diseases*, 12(12), 1136-1145.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.), John Wiley & Sons, Inc., Hoboken, New Jersey.

McGill, W. J. (1954), “Multivariate Information Transmission,” *Psychometrika*, 19(2), 97-116.

Mechelen, I. V., Bock, H.-H., and Boeck, P. D. (2004), “Two-mode Clustering Methods: A Structured Overview,” *Statistical Methods in Medical Research*, 13, 363-394.

Murali, T. M. and Kasif, S. (2003), “Extracting Conserved Gene Expression Motifs from Gene Expression Data,” *Pacific Symposium on Biocomputing*, 77-88.

Olejnik, S., Li, J., Supattathum, S., and Huberty, C. J. (1997), “Multiple Testing and Statistical Power with Modified Bonferroni Procedures,” *Journal of Educational and Behavioral Statistics*, 22(4), 389-406.

Papp, M., Lakatos, P. L.; Hungarian IBD Study Group, Palatka, K., Foldi, I., Udvardy, M., Harsfalvi, J., Tornai, I., Vitalis, Z., Dinya, T., Kovacs, A., Molnar, T., Demeter, P., Papp, J., Lakatos, L., and Altorjay, I. (2007), “Haptoglobin Polymorphisms are Associated with Crohn's Disease, Disease Behavior, and Extraintestinal Manifestations in Hungarian Patients,” *Digestive Diseases and Sciences*, 52(5):1279-1284.

Paterson, S., and Lello, J. (2003), "Mixed Models: Getting the Best Use of Parasitological Data," *Trends in Parasitology*, 19(8), 370-375.

Pensa, R. G. and Boulicaut, J-F. (2008), "Constrained Co-clustering of Gene Expression Data," *Proceedings SIAM SDM*, Atlanta, GA, 25-36.

Pensa, R. G. and Boulicaut, J-F., Cordero, F., and Atzori, M. (2010), "Co-clustering Numerical Data Under User-defined Constraints," *Statistical Analysis and Data Mining*, 3(1), 38-55.

Pensa, R. G., Robardet, C., and Boulicaut, J-F. (2005), "A Bi-clustering Framework for Categorical Data," *PKDD '05*, LNAI 3721, 643-650.

Pison, G., Struyf, A., and Rousseew, P. J. (1999), "Displaying a Clustering with CLUSPLOT," *Computational Statistics & Data Analysis*, 30(4), 381-392.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006), "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data," *Bioinformatics*, 22(9), 1122-1129.

Presley, L. L., Ye, J., Li, X., Leblanc, J., Zhang, Z., Ruegger, P. M., Allard, J., McGovern, D., Ippoliti, A., Roth, B., Cui, X., Jeske, D. R., Elashoff, D., Goodglick, L., Braun, J., and Borneman, J. (2011), "Host-microbe Relationships in Inflammatory Bowel Disease Detected by Bacterial and Metaproteomic Analysis of the Mucosal-luminal Interface," *Inflammatory Bowel Diseases*, doi: 10.1002/ibd.21793.

Puolamaki, K., Hanhijarvi, S., and Garriga, G. C. (2008), "An Approximation Ratio for Biclustering," *Information Processing Letters*, 108(2), 45-49.

Ramírez-Dávila, J. F., and Porcayo-Camargo, E. (2008), "Spatial Distribution of the Nymphs of *Jacobiasca Lybica* (Hemiptera: Cicadellidae) in a Vineyard in Andalucia, Spain," *Revista Colombiana de Entomologia*, 34(2), 169-175.

Rasmussen, M. and Karypis, G. (2004), "gCLUTO - An Interactive Clustering, Visualization, and Analysis System," *Technical Report, University of Minnesota*, TR#04-021.

Reimund, J. M., Arondel, Y., Escalin, G., Finck, G., Baumann, R., and Duclos, B. (2005), "Immune Activation and Nutritional Status in Adult Crohn's Disease Patients," *Digestive and Liver Disease*, 37(6), 424-431.

Reiss, D. J., Baliga, N. S., Bonneau, R. (2006), "Integrated Biclustering of Heterogeneous Genome-wide Datasets for the Inference of Global Regulatory Networks," *BMC Bioinformatics*, 7:280.

Ripollés Piquer, B., Nazih, H., Bourreille, A., Segain, J. P., Huvelin, J. M., Galmiche, J. P., and Bard, J.M. (2006), "Altered Lipid, Apolipoprotein, and Lipoprotein Profiles in Inflammatory Bowel Disease: Consequences on the Cholesterol Efflux Capacity of Serum Using Fu5AH Cell System," *Metabolism*, 55, 980-988.

Rocci, R. and Vichi, M. (2008), "Two-mode Multi-partitioning," *Computational Statistics & Data Analysis*, 52(4), 1984-2003.

Rogers, D. F. and Kulkarni, S. S. (2005), "Optimal Bivariate Clustering and a Genetic Algorithm with an Application in Cellular Manufacturing," *European Journal of Operational Research*, 160(2), 423-444.

Rousseeuw, P. J. and Ruts, I. (1996), "Algorithm AS 307: Bivariate Location Depth," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 45(4), 516-526.

Schotzko, D. J., and O'Keeffe, L. E. (1989), "Geostatistical Description of the Spatial Distribution of *Lygus hesperus* (Heteroptera: Miridae) in Lentils," *Journal of Economic Entomology*, 82(5), 1277-1288.

Sciola, V., Massironi, S., Conte, D., Caprioli, F., Ferrero, S., Ciafardini, C., Peracchi, M., Bardella, M. T., and Piodi, L. (2009), "Plasma Chromogranin a in Patients with Inflammatory Bowel Disease," *Inflammatory Bowel Diseases*, 15(6), 867-71.

Sheng, Q., Moreau, Y., and De Moor, B. (2003), "Biclustering Microarray Data by Gibbs Sampling," *Bioinformatics*, 19(Suppl. 2), ii196-205.

Sokol, H., Seksik, P., Furet, J. P., Firmesse, O., Nion-Larmurier, I., Beaugerie, L., Cosnes, J., Corthier, G., Marteau, P., and Doré, J. (2009), "Low Counts of Faecalibacterium Prausnitzii in Colitis Microbiota," *Inflammatory Bowel Diseases*, 15(8), 1183-1189.

Sun, J-T., Wang, X., Shen, D., Zeng, H-J., and Chen, Z. (2006), “Mining Clickthrough Data for Collaborative Web Search,” *WWW’06*, 947-948.

Swidsinski, A., Loening-Baucke, V., Vaneechoutte, M., and Doerffel, Y. (2008), “Active Crohn's Disease and Ulcerative Colitis can be Specifically Diagnosed and Monitored Based on the Biostructure of the Fecal Flora,” *Inflammatory Bowel Diseases*, 14(2), 147-161.

Takakura, K.-I. (2009), “Reconsiderations on Evaluating Methodology of Repellent Effects: Validation of Indices and Statistical Analyses,” *Journal of Economic Entomology*, 102(5), 1977-1984.

Tanay, A., Sharan, R., and Shamir, R. (2002), “Discovering Statistically Significant Biclusters in Gene Expression Data,” *Bioinformatics*, 18(Suppl. 1), S136-144.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), “Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression,” *Proceedings of the National Academy of Sciences of USA*, 99(10), 6567-6572.

Torrence, A. E., Brabb, T., Viney, J. L., Bielefeldt-Ohmann, H., Treuting, P., Seamons, A., Drivdahl, R., Zeng, W., and Maggio-Price, L. (2008), “Serum Biomarkers in a Mouse Model of Bacterial-induced Inflammatory Bowel Disease,” *Inflammatory Bowel Diseases*, 14(4), 480-490.

Tripathi, A., Lammers, K. M., Goldblum, S., Shea-Donohue, T., Netzel-Arnett, S., Buzza, M. S., Antalis, T. M., Vogel, S. N., Zhao, A., Yang, S., et al. (2009), “Identification of Human Zonulin, a Physiological Modulator of Tight Junctions, as Prehaptoglobin-2,” *Proceedings of the National Academy of Sciences of USA*, 106(39), 16799-16804.

Tukey, J. W. (1974), “Mathematics and Picturing Data,” *Proceedings of the 1974 International Congress of Mathematicians*, Vancouver, 2, 523–531.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007), “The Human Microbiome Project,” *Nature*, 449, 804–810.

Ueki, T., Mizuno, M., Uesu, T., Kiso, T., Nasu, J., Inaba, T., Kihara, Y., Matsuoka, Y., Okada, H., Fujita, T., and Tsuji, T. (1996), “Distribution of Activated Complement, C3b, and its Degraded Fragments, iC3b/C3dg, in the Colonic Mucosa of Ulcerative Colitis (UC),” *Clinical and Experimental Immunology*, 104(2), 286-292.

Ultsch, A. and Morchen, F. (2005), "ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM," *Technical Report, University of Marburg*, Germany, 46.

Watanabe, S. (1960), "Information Theoretical Analysis of Multivariate Correlation," *IBM Journal of Research and Development*, 4(1), 66-82.

Williams, L., Schotzko, D. J., and McCaffrey, J. P. (1992), "Geostatistical Description of the Spatial Distribution of *Limonius Californicus* (Coleoptera: Elateridae) Wireworms in the Northwestern United States, with Comments on Sampling," *Environmental Entomology*, 21(5), 983-995.

Willing, B., Halfvarson, J., Dicksved, J., Rosenquist, M., Järnerot, G., Engstrand, L., Tysk, C., and Jansson, J.K. (2009), "Twin Studies Reveal Specific Imbalances in the Mucosa-associated Microbiota of Patients with Ileal Crohn's Disease," *Inflammatory Bowel Diseases*, 15(5), 653-660.

Yamaguchi, N., Isomoto, H., Mukae, H., Ishimoto, H., Ohnita, K., Shikuwa, S., Mizuta, Y., Nakazato, M., and Kohno, S. (2009), "Concentrations of Alpha- and Beta-defensins in Plasma of Patients with Inflammatory Bowel Disease," Inflammation Research, 58(4), 192-197.

Yang, J., Wang, W., Wang, H., and Yu, P. (2002), "δ-Clusters: Capturing Subspace Correlation in a Large Data Set," *Proceedings of the 18th IEEE International Conference on Data Engineering*, 517-528.

Yang, J., Wang, H., Wang, W., and Yu, P. (2003), "Enhanced Biclustering on Expression Data," *Proceedings of the third IEEE Symposium on Bioinformatics and Bioengineering*, 321-327.

Yeung, R. W. (1991), "A New Outlook on Shannon's Information Measures," *IEEE Transactions on Information Theory*, 37(3), 466-474.

Yoon, S., Benini, L., and De Micheli, G. (2007), "Co-clustering: A Versatile Tool for Data Analysis in Biomedical Informatics," *IEEE Transactions on Information Technology in Biomedicine*, 11(4), 493-494.

Zhang, Z., Cui, X., Jeske, D. R., Li, X., Braun, J., and Borneman, J. (2010), "Clustering Scatter Plots Using Data Depth Measures," *The 6th International Conference on Data Mining (DMIN'10)*, Las Vegas, USA, 327-333.

Zhou, H., Yuan, X., Qu, H., Cui, W., and Chen, B. (2008), "Visual Clustering in Parallel Coordinates," *Computer Graphics Forum*, 27(3), 1047-1054.

Zissis, M., Afroudakis, A., Galanopoulos, G., Palermos, L., Boura, X., Michopoulos, S., and Archimandritis, A. (2001), "B2 Microglobulin: Is it a Reliable Marker of Activity in Inflammatory Bowel Disease?" *The American Journal of Gastroenterology*, 96(7), 2177-2183.

Zuo, Y. and He, X. (2006), "On the Limiting Distributions of Multivariate Depth-based Rank Sum Statistics and Related Test," *The Annals of Statistics*, 34(6), 2879-2896.