PhD Dissertation
Learning from SAGE Data

Tianjiao Chu
Department of Philosophy
Carnegie Mellon University

Jan, 2003

Abstract

Serial Analysis of Gene Expression (SAGE) is a technology for measuring quantitatively the expression levels of all genes in a cell population simultaneously. In this thesis, I study what we could not learn, what we could learn, and how to learn what we could learn, from SAGE data.

The first chapter of this thesis is a short introduction of the SAGE technology and another popular technology for measuring gene expression levels — the microarray technology. In the second chapter I show that, because all the current technologies can only measure the summed expression levels of genes from an aggregate of cells, in principle, we cannot learn the conditional independence relations among the expression levels of the genes in a single cell. Furthermore, the number of experiments required to estimate the correlation matrix of the gene expression levels is also too large to be feasible. The only feature of the expression levels of the genes we can learn reliably in practice is their expectations. Hence no algorithm of learning the gene regulatory network based on the conditional independence relations or the partial correlations among the expression levels of the genes could work with the data generated by the current technologies.

Chapter 3 is a study of a new sampling scheme — the sampling, amplification, and resampling (SAR) scheme — which is an abstraction of several key steps of the SAGE protocol. I present, in chapter 3, the asymptotic distribution of the data generated through the SAR scheme. In chapter 4, I explore several ways of learning from SAGE data. I shall first derive a statistical model for SAGE data based on the results of the chapter 3, then show how we can, based on SAGE data, test for differentially expressed genes, search for housekeeping genes, cluster genes according to their expression level patterns, and identify marker genes for a group of cell populations. Computer programs for the analysis of SAGE data are also developed based on the results of chapter 4, and used to analyze some real data. Chapter 5 briefly summarizes this thesis.

Acknowledgments

First I would like to give special thanks to my advisors Peter Spirtes and Clark Glymour, who have given me invaluable helps during my graduate study at CMU. They pointed to me the topic of this thesis, and provided me insightful guidance and a lot of constructive suggestions about this thesis.

Also I would like to thank the other members of the dissertation committee: David Peters, Larry Wasserman, and Greg Cooper. I have benefited a lot from rewarding conversations with them. David also kindly provided several SAGE libraries generated by his lab. The final version of this thesis has incorporated many comments and suggestions from Larry and Greg.

Finally, I am very grateful to my parents and my wife for their encouragement and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The past few years have seen the development of many new technologies for the measurement of gene expression levels, such as microarray and SAGE (Serial Analysis of Gene Expression). These technologies make it possible, for the first time, to measure the expression levels of the thousands of genes in a whole genome simultaneously. In this thesis, I explore what we can, what we cannot, and how to, learn from the gene expression level data generated by the SAGE technology. Some of the results presented in this thesis can be applied not only to the study of the SAGE data, but also to the analysis of the data generated by other technologies.

In this chapter, I shall give a short description of the microarray and SAGE technologies.

Microarray technology

Microarray technology is also called gene chip technology, because the key instrument in a microarray experiment is the microarray chip. Here I describe one type of microarray chip. To fabricate a Microarray chip, thousands of types of DNAs with known sequences are spotted to a specially treated glass or nylon surface by high speed robots, and then immobilized. Each spot contains only one type of DNA, usually a 500-5000 base long fragment of a known gene or a promoter region of some known gene. To compare the expression levels of the genes in two cell populations, we first need a microarray chip containing the DNA fragments of the genes of interest. The mRNA transcripts are then extracted from the two cell populations, reverse-transcribed into cDNA, and labeled with red and green fluorescent dyes (Cy3 and Cy5). The two pools of labeled cDNA are mixed together. Ideally, the mixture should contain equal amount of cDNA from each cell population. In practice, though, this is not always the case. The mixture is then deposited to each spot of the microarray. The cDNA that match the DNA in a spot, i.e., with complementary sequences, will hybridize to that spot, and the unhybridized cDNA will be washed away. The amount of cDNA generated from the two cell populations hybridized to each spot can be measured by scanning the spot and reading off the intensities of the red and green fluorescent signals. The relative abundance of the cDNA for a certain gene will tell the relative abundance of the mRNA transcripts of that gene in the two cell populations. (For more information about microarray technology and its applications, see `http://www.gene-chips.com/`.)

The microarray technology is relatively cheap and very fast — a microarray experiment can be done within a few hours. There have been a lot of studies on the analysis of microarray data in the past few years. Some have focused on the estimation of the gene expression levels and the identification of differentially expressed genes (Chen et al 1997, Duoit et al , 2000, Yang et al 2000, Tusher et al 2001), some on the clustering of the genes based on the gene expression level patterns (Eisen et al 1998, D'haeseleer et al 2000, Bar-Joseph et al 2001), some on the inference of gene regulatory networks based on the data generated from disturbed or undisturbed cell populations (Liang et al 1998, D'haeseleer 2000, Friedman et al 2000, Ideker et al 2001, Spirtes et al 2001, Yoo et al 2002).

SAGE technology

The SAGE technology (Velculescu et al 1995) is based on a common practice in the study of molecular genetics — identifying a gene by an expression sequence tag (EST), i.e., a short fragment of that gene. By sequencing a segment of the gene, and comparing it with the library of the known ESTs, we could figure out from which gene the EST is read off. However, this practice was used only for the study of the expression levels of a few genes, because the length of an EST usually is in the range of 100-300, and sequencing many long ESTs could be very expensive. The innovation of the SAGE technology is that, instead of sequencing the relatively long ESTs, it uses short 10-12 base long tags to identify the genes.

In the beginning of a SAGE experiment, a collection of mRNA transcripts is extracted from a sample of cells, e.g., a tissue sample, using some standard methods. Then cDNA clones are synthesized from the mRNA transcripts. Each cDNA clone is cut at specific sites by some anchoring enzyme, such as NlaIII. The 3' ends of the resulting cDNA fragments are bound to magnetic beads, divided into two parts, and then ligated with two types of linkers respectively. The tags are released from the magnetic beads by tagging enzyme, blunt-ended, and ligated to form ditags. These ditags are amplified by the PCR procedure, purified, and then linked together to form concatemers. The concatemer of certain lengths are isolated, and then cloned and sequenced. The counts of various tags can be read off from the sequencing data. (Velculescu et al 2000).

The result of a SAGE experiment, called a SAGE library, is a list of counts of the sequenced tags. Thus, strictly speaking, a SAGE experiment only measures, in some sense, the expression levels of the tags. To get the gene expression levels from a SAGE library, we need a map from the tags to the genes. Ideally, we would like to have a bijection from the tags to the genes. However, in reality, a tag may correspond to more than one gene, and a gene may generate more than one tag. This means, in some cases, that the difference in the relative frequencies of a tag in two tissues may only imply that at least one of the genes mapped to this tag is differentially expressed. Fortunately, there are a large number of tags each of which is mapped to only one gene, and a large number of genes each of which is mapped to only one tag. (For more information about the SAGE technology, see the web-site of National Center for Biotechnology Information (NCBI):

http://www.ncbi.nlm.nih.gov/SAGE/).

The SAGE technology is more expensive and slower than the microarray technology, and probably because of this, it is not as well-studied as the microarray technology. However, it does have two distinct advantages. First, the results of the SAGE experiments are portable, in the sense that we can compare directly the results from two SAGE experiments. For example, given any two SAGE libraries, regardless of whether they are generated by the same or different labs at the same or different times, we can always compare whether a gene is differentially expressed in the two tissues where the two SAGE libraries are generated from by comparing the relative frequencies of the tag representing that gene in the two libraries. This is not true, however, for microarray technology. Indeed, to account for the chip to chip variation, dye to dye variation, and other variations introduced in the microarray experiments, often sophisticated experiment design is needed in microarray experiments (Kerr et al 2001a, Kerr et al 2001b, Yang & Speed 2002).

Another distinct advantage of the SAGE technology is that we do have a better understanding of the process that generates a SAGE library. As shown in chapter 4, we can find, for each step of the SAGE protocol, an accurate statistical model. Moreover, because the SAGE results are essentially the counting of the genes in the sample cells, the general statistical model for the SAGE results are much simpler than that for the microarray results. Therefore, a careful study of the SAGE data could shed light on what we can, and cannot, learn from gene expression level data.

In this thesis, I focus my study on the gene expression level data generated by the SAGE technology. In the second chapter of this thesis, I discuss the implication of the fact the current technologies, including SAGE and microarray, can only measure the summed expression levels of the genes from a large aggregate of cells. I show that, first, in general, the conditional independence relations among the expression levels of the genes in a single cell are not the same as the conditional independence relations among the summed expression levels of genes from an aggregate of cells. Indeed, in most cases, where the summed gene expression levels are measured from a large number of cells, it can be shown that the conditional independence relations among the summed gene expression levels are essentially determined by the correlation matrix among the gene expression levels, regardless of the conditional independence relations among the expression levels of the genes in an individual cell. Therefore, in principal, we cannot learn the conditional independence relations among the expression levels of the genes in a single cell from the gene expression level data generated by the SAGE or the microarray technology, unless we make the biologically implausible assumption that the gene expression levels can be approximated by a linear Gaussian model. Furthermore, even if we assume that the gene expression levels follow roughly the linear Gaussian Model, at least for the SAGE data, the number of experiments required to estimate the correlation matrix of the gene expression levels is too large to be feasible. This suggests that any algorithm for learning the gene regulatory network based on the conditional independence relations among the expression levels of the genes would not work with the data generated by the current technologies.

In the third chapter of this thesis I discuss a new sampling scheme — the sampling, amplification, and resampling (SAR) scheme. To generate a sample according to the SAR scheme, we need first to get an original sample, then each element of the original sample is amplified independently and identically by a random folder. The final sample is drawn with or without replacement from the amplified sample. This scheme is an abstraction of three keys steps of a typical SAGE experiment. The main result of Chapter 3 is the derivation of the asymptotic distribution of the data generated by the SAR sampling scheme. During the derivation of the asymptotic distribution of the SAR sample, I also prove some other theorems, including the theorem about the asymptotic distribution of the ratio of two sums of iid sequences, and the theorem about the convergence of marginal distributions given the convergence of conditional distributions. The result of Chapter 3 provides the theoretical foundation for the analysis of SAGE data in Chapter 4. Readers who are mainly interested in the practical analysis of the SAGE data can skip this chapter and go directly to Chapter 4.

Based on the results of Chapter 3, and a close examination of the SAGE protocol, in Chapter 4, I propose a new statistical model for the SAGE gene expression level data. While two critical parameters for the new model are missing, it can be shown that, in many cases, the multinomial model can be used to approximate the new model of the SAGE data. Therefore, many technologies designed for the analysis of multinomial data can be used to analyze the SAGE data. For example, the test of association in a two way table can be used to test whether a set of genes are differentially expressed over a set of cell populations. I also examine the concept of housekeeping gene, propose several alternative rigorous definitions of housekeeping gene based on different levels of statistical constraints, and present the corresponding algorithms for the search of housekeeping genes for the SAGE gene expression level data. Another application of the statistical model for the SAGE data is the clustering of genes based on the SAGE gene expression level data. I consider several different approached of gene clustering algorithm, and suggest two new algorithms that take into account the fact that the SAGE data follow approximately the multinomial distributions. Finally, I study marker genes, i.e., genes that express roughly one way in one family of cell populations, and a different way in another family of cell populations. I propose a rigorous statistical definition of marker genes, and give an algorithm for searching for the marker genes from the SAGE data generated from two families of cell populations. The algorithms are implemented in C, Java, and Splus, and are used to analyze some SAGE data.

Chapter 5 concludes the thesis with a short discussion about what we could do possibly in the future to discover gene regulatory networks.

Chapter 2

The problem of aggregation

The inference of causal information from purely observed data has been an active field of study in the past decade. Combining statistics, graph theory, and computer science, various algorithms for making causal inferences from observational data have been proposed, analyzed, and applied to solve real world problems (Spirtes et al 2001, Pearl 2000). This technique seems promising for the task of deriving the gene regulatory networks from the large collection of gene expression data set generated using microarray, SAGE, and other technologies. Indeed, there have already been some publications about using causal inference techniques to infer gene regulatory network from gene expression data (Akutsu, 1998; D'hasseleer, 2000; D'hasseleer, et al., 2000; Friedman, 2000; Hartemink, 2001; Liang, et al., 1998; Shrager, et al., 2002). The basic idea is to get the expression levels in repeated samples from the same cell population, or similar cell populations, possibly in the form of time series data, and to infer the regulatory structure from the statistical dependencies and independencies among the measured expression levels.

The apparent advantage of this approach is that it offers the possibility of determining multiple relationships without separate experimental interventions. That is, in principle, we could figure out the gene regulatory network just by observing the expression levels of the genes, without conducting elaborate experiments to interfere with the regulatory network in various ways and checking how the gene expression levels react to the experimental interference. If, for example, gene A regulates gene C only by regulating gene B which in turn regulates C, the expression level of A should be independent, or nearly independent, of the expression level of gene C conditional on the expression level of gene B. In principle, if adequate sample sizes were available, the method could also be used as a supplement to gain additional information from experiments in which the expression of particular genes are experimentally suppressed or enhanced. The requisite statistical procedures for this strategy are more elaborate, and require direct or indirect (e.g., implicit in the posterior probabilities) estimates of conditional probability relationships among expression levels.

However, there are many statistical obstacles to the causal inference approach of deriving gene regulatory network from the gene expression level data. They include the joint influence of unmeasured factors (e.g., unmeasured gene expressions or extra-cellular factors), a variety of sources of measurement error, an unknown family of probability distributions governing the errors, the presence of cycles in the regulatory network, etc. Some of these difficulties—in particular the presence of latent common causes and cycles—have, in principle, been overcome. (Spirtes, et al, 2001). However, as we are going to see in this chapter, there is a more elementary statistical difficulty with the causal inference approach that calls its value into question and raises a set of important research problems. This difficulty arises from the fact that the gene expression level data obtainable by the current technologies are all measurement of the aggregate of the mRNA transcripts from a large number of cells.

The next section is a short introduction of the basic ideas of causal inference. In section 2, I show that, except for some special cases, the conditional independence relations among the aggregates of genes from a number of cell in general are not the same as the conditional independence relations among the genes in a single cell. I then study the conditional independence relations among the aggregates of genes from a large number of cells, as well as the difficulty in detecting these relations. The conclusion of this chapter is that, in principle, we can only learn from the gene expression data both the variances and the means of the gene expression levels. In practice, however, we probably can only learn the means reliably. The proofs of the theorems presented in this chapter are given in the appendix to this chapter.

2.1 Causal graph and gene regulatory network

A directed graph consists of a set of vertices, and a set of directed edges connecting pairs of vertices. If there is an edge coming out of vertex X and ending at vertex Y, X is called a parent of Y, and Y a child of X. If in a directed graph, the edges cannot form any directed cycle, then the graph is called directed acyclic graph (DAG). DAG provides an intuitive representation of the causal relations among a set of random variables: Each vertex in the graph represents a random variable, and a variable X is a direct cause of variable Y if and only if X is a parent of Y, i.e., there is a direct edge from X to Y in the graph. A DAG with causal interpretation is called a causal graph.

A causal model consists of a causal graph, and, for each variable in the graph, the conditional distribution of this variable given all of its parents. Usually, the conditional distribution of a variable is expressed as a function of all of its parents and an independent error terms. For example, for the variables in Figure 2.1, we could specify the following functional relations:

$$\begin{aligned} Z &= f(Y, W) + \epsilon_z \\ Y &= g(X) + \epsilon_y \\ W &= h(X) + \epsilon_w \end{aligned} \tag{2.1}$$

Where f, g, h are any functions and ϵ_z , ϵ_y, ϵ_w are independently dis-

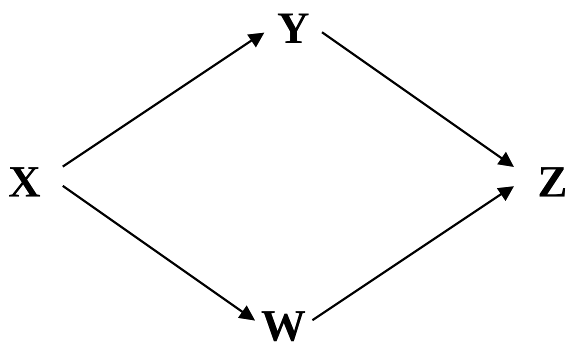


Figure 2.1: A simple causal graph

tributed noises. It follows that the joint probability density of Z, Y, W, X admits a Markov factorization

$$d(X, Y, Z, W) = d(Z|Y, W)d(Y|X)d(W|X)d(X) \tag{2.2}$$

The Markov factorization implies that Y, W are independent conditional on X, and that X, Z are independent conditional on Y, W, and is in fact equivalent to specifying that these two relationships hold. More generally, in a causal model, the following condition will be satisfied:

Markov Condition: *Consider a causal model G. Let X be a variable in G, $\boldsymbol{Y}$ be the set of parents of X in G, and $\boldsymbol{Z}$ a set of variables that are neither parents nor descendents of X. Then conditional on $\boldsymbol{Y}$, X and $\boldsymbol{Z}$ are independent.*

The Markov condition is a sufficient condition for conditional independence relation in the sense that a conditional independence relation pre-

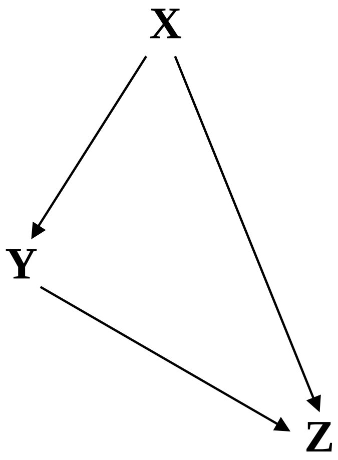


Figure 2.2: Is X and Z independent?

dicted by the Markov condition must be observed. However, it is possible that in a causal model some observed conditional independence relations are not predicted by the Markov condition. For example, suppose we have three random variables X, Y, and Z, as shown in figure 2.2. Their functional relationships are given as:

$$X, \epsilon_Y, \epsilon_Z \sim N(0,1)$$

$$Y = aX + \epsilon_Y$$

$$Z = bX + cY + \epsilon_Z$$

The Markov condition does not predict that X and Z are independent, and indeed X and Z will be dependent as long as $b + ac \neq 0$. However, in the case where $b + ac = 0$, X and Z becomes independent. This would be a problem if we want to infer the causal relations from the conditional independence relations. Fortunately, it can be shown that, at least for the most familiar types of causal models, i.e., the structural equation model and the Bayes network model, the set of values of the parameters that lead to conditional independence relations not predicted by the causal graph has Lebesgue measure 0. This seems to justify the following condition:

Faithfulness Condition: *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ be three disjunct sets of variables in a causal model G. Then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given $\boldsymbol{Z}$ only if this is implied by the Markov condition.* [1]

The Markov and the faithfulness conditions establish a close relation between causation and conditional independence. Under the Markov and the faithfulness conditions, each causal graph uniquely specifies a set of conditional independence relations. It is possible that different causal graphs may specify the same set of conditional independence relations. In this case, we call these causal graphs Markov equivalent, and the set of all these graphs constitute a Markov equivalent class. In the example of figure 2.1, the Markov equivalence class consists of the graph shown and the graphs obtained by reorienting exactly one of the edges from X to Y or X to W.

Absent extra knowledge from other sources, the Markov equivalence class represents the most information that could be obtained from conditional independencies among the variables. Several search algorithms have been de-

[1] For more discussion about the faithfulness condition and its implication, see Robins et al (2000).

veloped to output a graphical representation of the Markov equivalent class based on the conditional independence relations observed in a population.

Further studies have been focused on the causal inference with the presence of unobserved variables that are parents of pairs of observed variables. Algorithms, such as FCI, have been developed to infer common causal patterns from populations sharing same set of conditional independence relations among observed variables. Moreover, causal inference from the population where there is feedback are also studied (Richardson, 1996)

To represent gene regulatory networks with causal graphs, each variable in the causal graph will be the level of expression of a particular gene. A directed edge from one variable X to another variable Y in such a graph indicates that gene X produces a protein that regulates gene Y. It is well known that the gene regulatory networks contain self-loops and cycles, i.e., some gene may regulate itself either directly, or through some other genes. In principle, this type of regulatory networks could be represented by cyclic causal graphs. However, most proposed search methods have been confined to acyclic graphs, hence, for simplicity, one usually assumes the regulatory network could be represented by an acyclic graph with noises and random measurement errors for each measurement of each gene that are independent of those for any other gene. This simplification becomes unnecessary when data are obtained in a time series, because here the regulatory relationships can be represented by a directed acyclic causal graph, but with vertices appropriately labeled by gene and time.

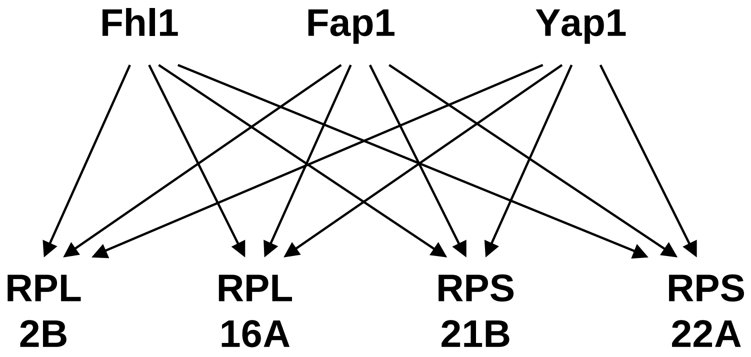


Figure 2.3: A yeast gene regulatory network

Figure 2.3 shows a yeast gene regulatory network represented by a directed acyclic graph (Lee et al 2002).

2.2 Conditional independence under aggregation[2]

Our goal is to discover the regulatory structure in individual cells from the gene expression level data. To achieve this goal, we need the measurements of the gene expression levels for many single cells. For example, suppose figure 2.1 represents a true regulatory network. To infer this network, we need to collect a number of cells, and for each cell, measure the expression levels of genes X, Y, Z, and W. Let the number of cells be n, and the expression levels of X, Y, Z, and W in the ith cell be X_i, Y_i, Z_i, and W_i. We should get a sample of size n, where each data point is a 4-dimensional

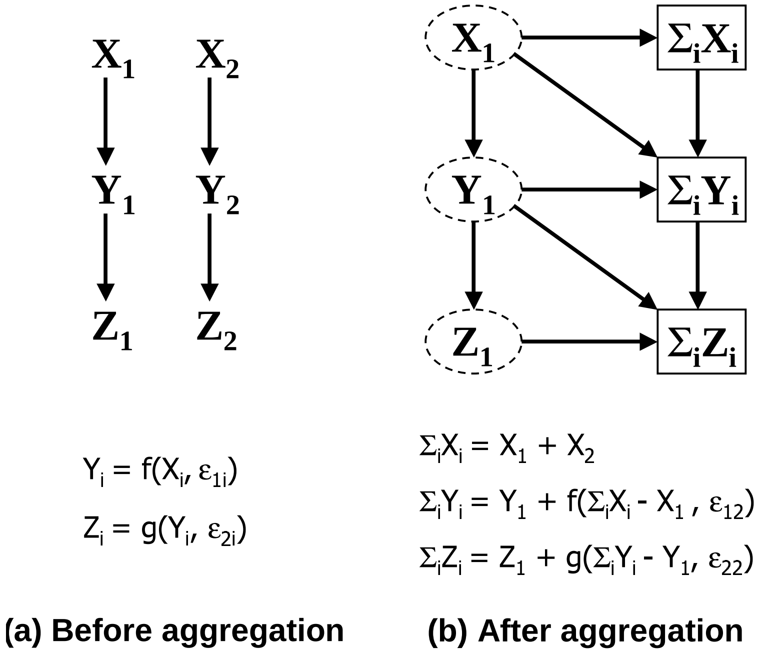
[2] This section is based on a joint work with Glymour, Scheines, and Spirtes.

vector representing the expression levels of the 4 genes in a single cell. Then we could apply various algorithms to this data set to infer the regulatory network.

However, the gene expression data we could get using today's technology are measurements of mRNA transcripts obtained from thousands, or even millions, of cells. Such measurements are not of variables such as X, Y, Z, and W in figure 2.1, but are instead, ideally, of the sums of the values of X, Y, Z, and W over many cells. That is, for each measurement, we get a single data point, which is not a 4-dimensional vector (X_i, Y_i, Z_i, W_i) for some i, but the vector $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Z_i, \sum_{i=1}^{n} W_i)$.

This proves to be a problem for the causal inference approach for discovering regulatory structures, which relies on the statistical dependencies among the gene expression levels, because the conditional dependencies/independencies among the gene expression levels of a single cell in general are not the same as those among the summed gene expression levels over a number of cells. In other words, the conditional independence relations do not preserve under aggregation. For example, if the variables in figure 2.1 are binary, and each measurement is of the aggregate of transcript concentrations from two or more cells, $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} Z_i$ are not independent conditional on $\sum_{i=1}^{n} Y_i$, $\sum_{i=1}^{n} W_i$, and the associations obtained from repeated samples will not therefore satisfy the Markov factorization implied by the graph in figure 2.1 (Danks and Glymour, 2001).

A graphical heuristic explanation of the problem of aggregation is shown



(a) Before aggregation **(b) After aggregation**

Figure 2.4: Problem of aggregation

in figure 2.4. Here we consider a simple regulatory network: Gene X regulates gene Y, and Y regulates gene Z. Figure 2.4(a) shows the ideal case where we could make two measurements from two separate cells. Using the Markov condition, it is easy to see, from the graph, that X_i and Z_i are independent given Y_i for $i = 1, 2$. Note that here X_1, X_2, ϵ_{11}, ϵ_{12}, ϵ_{21}, and ϵ_{22} are independent. Figure 2.4(b) shows what happens when we can only measure the gene express levels of the aggregate of the two cells. X_1, Y_1, and Z_1 now are latent variables, represented by dashed ovals. The three observed variables, $\sum_{i=1}^{2} X_i$, $\sum_{i=1}^{2} Y_i$, and $\sum_{i=1}^{2} Z_i$, are represented by solid rectangles. Each of them is expressed as a function of its parent(s) and an independent error term, where the error term for $\sum_{i=1}^{2} X_i$ is X_2, the error term for $\sum_{i=1}^{2} Y_i$ is ϵ_{12}, and the error term for $\sum_{i=1}^{2} Z_i$ is ϵ_{22}. It is not difficult to see that the causal graph shown in figure 2.4(b) does not imply that $\sum_{i=1}^{2} X_i$ and $\sum_{i=1}^{2} Z_i$ are independent conditional on $\sum_{i=1}^{2} Y_i$.

Interestingly, there are some special cases where the conditional independencies are invariant under aggregation. For example, although the graph in figure 2.4(b) does not entail that $\sum_{i=1}^{2} X_i$ and $\sum_{i=1}^{2} Z_i$ are independent conditional on $\sum_{i=1}^{2} Y_i$, if X, Y, and Z are all binary variables, the implied conditional independence of X, Z given Y will hold as well for $\sum_i X_i$, $\sum_i Y_i$ and $\sum_i Z_i$ (Danks and Glymour, 2001).

Linear, normal distributions have special virtues for invariance. Whatever the directed acyclic graph of cellular regulation may be, if the noise terms, as in equations 2.1, are normally distributed and each variable is a linear function of its parents and an independent Gaussian noise, then the Markov factorization holds for the summed variables. For in that case, conditional independence is equivalent to vanishing partial correlation, and the partial correlation of the two variables, each respectively composed of the sum of n like variables, will be the same as the partial correlation of the unsummed variables.

Two less restrictive sufficient conditions for conditional independence of variables to be the same as the conditional independence of their sums, are given in the following two theorems. The general setting is an acyclic graph such that each node is a function—not necessarily additive—of its parents and an independent noise term.

Theorem 1 (Local Markov Theorem). *Given an acyclic graph G representing the causal relations among a set $\boldsymbol{V}$ of causal sufficient random variables.*[3] *Let $Y, X^1, \cdots, X^k \in \boldsymbol{V}$, and $\boldsymbol{X} = \{X^1, \cdots, X^k\}$ be the set of parents of Y in G. If $Y = \boldsymbol{c}^T\boldsymbol{X} + \epsilon$,* [4] *where $\boldsymbol{c}^T = (c^1, \cdots, c^k)$, and ϵ is a noise term independent of all non-descendents of Y, then Y is independent of all its non-parents and non-descendents conditional on its parents $\boldsymbol{X}$, and this relation holds under aggregation.*

The above theorem states that, under the local linearity condition, the conditional independence relation between a random variable and its non-descendent and non-parent is invariant under aggregation. In the next theorem, we give another sufficient condition for the conditional independence relation to be invariant under aggregation.

Theorem 2 (Markov Wall Theorem). *Given an acyclic graph G representing the causal relations among a set $\boldsymbol{V}$ of random variables. Let $\boldsymbol{X} = \{X^1, \cdots, X^h\}$, $\boldsymbol{Y} = \{Y^1, \cdots, Y^k\}$, $\boldsymbol{W} = \{W^1, \cdots, W^m\}$, and $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{W} = \boldsymbol{V}$. Suppose that the following three conditions hold:*

1. *The joint distribution of $X^1, \cdots, X^h$, Y^1, $\cdots$, Y^k is multivariate normal with nonsingular covariance matrix.*

[3] A set $\boldsymbol{V}$ of random variables are causal sufficient if, for any $X, Y \in \boldsymbol{V}$, if Z is a common cause of X and Y, then $Z \in \boldsymbol{V}$.

[4] In this and the next theorems, we shall use the same bold face symbol to represent both a set of variables, and a vector of that set of variables.

2. *For $i = 1, \cdots, k$, Y^i is neither a parent, nor a child, of any variable $W^j \in \boldsymbol{W}$. That is, there is no direct edge between a variable in $\boldsymbol{Y}$ and a variable in $\boldsymbol{W}$.*

3. *For $i = 1, \cdots, h$, X^i is not a child of any variable $W^j \in \boldsymbol{W}$. That is, if there is an edge between a variable in $\boldsymbol{X}$ and a variable in $\boldsymbol{W}$, the direction of the edge must be from the variable in $\boldsymbol{X}$ to the variable in $\boldsymbol{W}$.*

Then conditional on $\boldsymbol{X}$, $\boldsymbol{Y}$ is independent of $\boldsymbol{W}$, and this relation holds under aggregation.

Although there are established regulatory mechanisms in which some regulators of a gene act linearly in the presence of a suitable combination of other regulators of the same gene (Yuh, 1998), there does not appear to be any known regulatory system that is simply linear. One of the best-established regulatory functional relations seems to be the expression of the Endo16 gene of the sea urchin (Yuh, et al., 1998). The expression level of the gene is controlled by a Boolean regulatory switch between two functions, each of which is a product of a Boolean function of regulator inputs multiplied by a linear function of other regulator inputs. Even much simplified versions of such transmission functions do not preserve conditional independence over sums of variables.

For example, consider the causal structure shown in figure 2.5, where $Y = UX$ and $Z = VY$. (This causal structure is a simplification of the

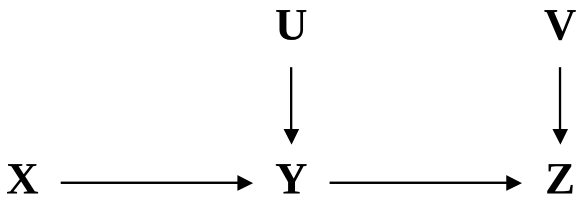


Figure 2.5: A Sea Urchin type regulatory network

proposed Sea Urchin endo16 gene regulatory network.) Suppose X has a Poisson distribution with parameter λ, U and V are Bernoulli random variables with parameters p_1 and p_2 respectively.

It is obvious that X and Z are independent conditional on Y. However, it can be shown that this relation does not hold under aggregation. For example, let X_1, U_1, Y_1, V_1, Z_1 and X_2, U_2, Y_2, V_2, Z_2 be two independent samples generated from the same causal structure. Assuming that U and V are not degenerate, that is, $p_1, p_2 \neq 1$ and $p_1, p_2 \neq 0$, through straightforward calculation, we can show that:

$$\begin{aligned} &P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2) \\ &\quad = p_2 - p_2(1 - p_2)\frac{p_1 \lambda e^{-\lambda}}{1 - p_1 + p_1 e^{-\lambda} + p_1 \lambda e^{-\lambda}} \end{aligned}$$

$$P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2, X_1 + X_2 = 4) = p_2$$

Clearly, conditional independence relation is not preserved under aggregation for the causal structure shown in figure 2.5, because as long as $p_1, p_2 \neq 1$ and $p_1, p_2 \neq 0$,

$$\begin{aligned} & P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2) \neq \\ & \qquad P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2, X_1 + X_2 = 4) \end{aligned}$$

In the above examples, we treat the number n of cells in an aggregated sample as a constant. In practice, however, when several samples are obtained, the number of cells in each sample is a random variable. This could make the inference of conditional association even more problematic. When n is held constant, we know that there is a fixed set of conditional associations among the aggregated genes, though they are not the same as the genes within each individual cell. If n is a random variable, we are not sure if the aggregated genes in different samples share the same set of conditional associations.

2.3 Conditional independence among large sample means

The discussion in section 2 suggests that, other than by chance, inference of genetic regulatory networks from associations among measured expression levels is possible only if the graphical structure and transmission functions from regulator concentrations to expression concentrations of regulated genes preserve conditional independence relations over sums of i.i.d. units.

The few sufficient conditions we have provided are not biologically relevant, but, unfortunately, the negative example based on a simplification of Endo 16 regulation (figure 2.5) is relevant.

Of course, there are certainly many real gene regulatory networks that are not similar to this simplified Endo 16 regulatory network. While the Endo 16 regulatory network fails to preserve conditional independence under aggregation, we cannot conclude that the other types of networks will also fail. Thus, it would be very nice if we could find some interesting general sufficient conditions for conditional independence *not* to be invariant. However, a general theory that works for the aggregation of arbitrary number of cells seems very complicated, if not impossible. Instead, in this section, we are going to explore some general conditions under which we can predict whether the conditional independence relations will not hold as the number of cells aggregated goes to infinity. The main result of this section is that, if the joint distribution of the measurements of the genes in a cell falls into either of two general classes of distributions, the conditional independence relations among the measurements of the genes from an aggregate of large number of cells will be essentially determined by the covariance matrix of the original joint distribution.

Recall that for a set of variable whose joint distribution is a multivariate normal, the conditional independence relations are entirely determined by the covariance/correlation matrix. More precisely, if the random vector $(X, Y, \boldsymbol{Z})$ has a multivariate normal distribution, then X and Y are inde-

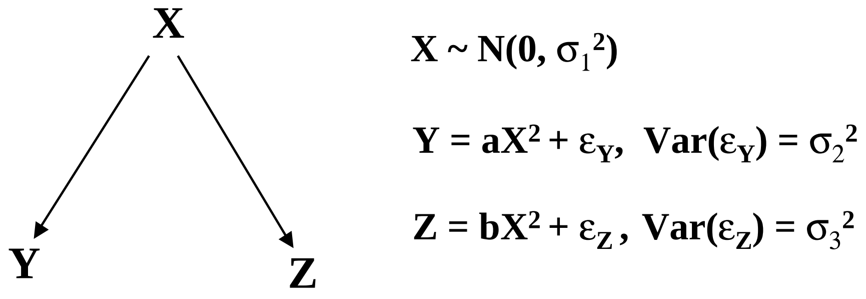


Figure 2.6: Covariance matrix and conditional independence

pendent given $\boldsymbol{Z}$ if and only if the partial covariance/correlation of X and Y with respect to $\boldsymbol{Z}$ is 0. However, this special relation between conditional independence and covariance matrix does not hold in general. For example, consider the causal model shown in figure 2.6. The covariance matrix for (X, Y, Z) is:

$$\mathrm{Cov}(X,Y,Z) = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & 2a^2\sigma_1^2 + \sigma_2^2 & 2ab\sigma_1^2 \\ 0 & 2ab\sigma_1^2 & 2b^2\sigma_1^2 + \sigma_3^2 \end{bmatrix}$$

The partial covariance of Y and Z with respect to X and the partial covariance of X and Z with respect to Y are, respectively:

$$\begin{aligned} \mathrm{Cov}(Y,Z;X) &= \mathrm{Cov}(Y,Z) - \mathrm{Cov}(Y,X)\mathrm{Var}(X)^{-1}\mathrm{Cov}(X,Z) = 2ab\sigma_1^2 \\ \mathrm{Cov}(X,Z;Y) &= \mathrm{Cov}(X,Z) - \mathrm{Cov}(X,Y)\mathrm{Var}(Y)^{-1}\mathrm{Cov}(Y,Z) = 0 \end{aligned}$$

However, it is easy to see, from figure 2.6, that Y and Z are *independent* given X, and that X and Z are *dependent* given Y.

However, we are going to show, under certain conditions, that the conditional independence relations among the sums of a large sample of a set of random variables will be, in some sense, more and more determined, as the sample size increases, by the covariance matrix of this set of variables. The basic idea is that, by the central limit theorems, the (properly normalized) sums of a large sample of a set of random variables will converge weakly to a multivariate normal distribution, which, as we mentioned before, has the unique property that a one-to-one relation exists between the set of conditional independencies and the covariance matrix. Of course, some conditions are required to ensure that the conditional distribution of the sums of a set of variables given the sums of another set of variables will also converge in the right way.

First we look at the class of distributions with non-singular covariance matrices and bounded densities (with respect to the Lebesgue measure). We will show that, for a random vector $(X, Y, \boldsymbol{Z})$ belonging to this class of distributions, the density of conditional distribution of the large sample sums $(\sum_i X_i, \sum_i Y_i)$ given large sample sums $\sum_i \boldsymbol{Z}_i$ converges in total variation distance to the product of the densities of $\sum_i X_i$ given $\sum_i \boldsymbol{Z}_i$ and $\sum_i Y_i$ given $\sum_i \boldsymbol{Z}_i$ if and only if the partial correlation of X and Y with respect to $\boldsymbol{Z}$ is 0. To prove this, we need a few lemmas about the characteristic functions of multivariate distributions.

Lemma 1. *Let* $\boldsymbol{X} = (X_1, \cdots, X_k)$ *be a random vector with characteristic* *function $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$. Then if $\boldsymbol{X}$ has a density with respect to the Lebesgue measure, $|\phi(\boldsymbol{t})|$ equals 1 only when $\boldsymbol{t} = \mathbf{0}$*

Lemma 2. *Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ be a random vector with a bounded density with respect to the Lebesgue measure, and $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$ be the characteristic function of $\boldsymbol{X}$. Then $\phi(\boldsymbol{t})$ is integrable if $\phi(\boldsymbol{t}) \geq 0$.*

Lemma 3. *Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ be a random vector with a bounded density $f(\boldsymbol{x})$ with respect to the Lebesgue measure, and $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$ be the characteristic function of $\boldsymbol{X}$. Then $|\phi(\boldsymbol{t})|^n$ is integrable for all $n \geq 2$.*

With the previous lemmas, we can prove the first main theorem of this section, which is a generalization of the well known theorem of convergence in density for univariate random variables (Feller 1971, van der Vaart 1998):

Theorem 3. *Let $\boldsymbol{X}_n$ be i.i.d. random vectors with 0 mean and non-singular covariance matrix $\boldsymbol{\Sigma}_X$, and $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i/\sqrt{n}$. Suppose the characteristic function $\phi(\boldsymbol{t}) = E[\exp(\boldsymbol{t}^T\boldsymbol{X})]$ is integrable, then $\overline{\boldsymbol{X}}_n$ have bounded continuous densities that converge uniformly to the density of a multivariate normal distribution with 0 mean and covariance matrix $\boldsymbol{\Sigma}_X$.*

The following corollary is a direct consequence of Theorem 3.

Corollary 1. *Let $\{(X_n, Y_n, \boldsymbol{Z}_n)\}$ be a sequence of i.i.d. $k+2$ dimensional random vectors with mean $\mathbf{0}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}$. Suppose $(X_n, Y_n, \boldsymbol{Z}_n)$ and $\boldsymbol{Z}_n$ both have bounded densities (with respect to the Lebesgue measure). Let $\overline{X}_n = (\sum_{i=1}^n X_i)/\sqrt{n}$, $\overline{Y}_n = (\sum_{i=1}^n Y_i)/\sqrt{n}$, and*

$\overline{\boldsymbol{Z}}_n = (\sum_{i=1}^n \boldsymbol{Z}_i)/\sqrt{n}$, and $(U, V, \boldsymbol{W})$ be a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ converges to the total variation distance between the conditional distribution of (U, V) given $\boldsymbol{W}$ and the product of the conditional distributions of U given $\boldsymbol{W}$ and V given $\boldsymbol{W}$ almost surely with respect to the measure induced by $\boldsymbol{W}$.

Note that the conditions for Theorem 3 and Corollary 1 could be made even more general. However, the current conditions for Theorem 3 and Corollary 1 are more intuitive.

The main implication of Corollary 1 is that, under the conditions for Corollary 1, the conditional independence relations among the summed expression levels of the genes from large number of cells will eventually be determined by the covariance matrix of the expression levels of the genes within a single cell. Recall that unlike conditional independence relations, the covariance matrix, with appropriate normalization, is invariant under aggregation. That is, for the variables in Corollary 1, we have:

$$n\text{Var}(X, Y, \boldsymbol{Z}) = \text{Var}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i, \sum_{i=1}^n \boldsymbol{Z}_i\right) = n\text{Var}(\overline{X}_n, \overline{Y}_n, \overline{\boldsymbol{Z}}_n)$$

Therefore, $\text{Var}(\overline{X}_n, \overline{Y}_n, \overline{\boldsymbol{Z}}_n) = \text{Var}(U, V, \boldsymbol{W})$. Now consider the case where X and Y are independent given $\boldsymbol{Z}$, but the partial correlation of X and Y with respect to $\boldsymbol{Z}$ is not 0. Clearly, the partial correlation of

U and V with respect to $\boldsymbol{W}$ cannot be 0 either, hence U and V must be dependent given $\boldsymbol{W}$. As we aggregate more and more cells, the total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ converges to a positive value. [5] Therefore, there must be a number N such that for all $n \geq N$, $\overline{X}_n$ and $\overline{Y}_n$ are dependent given $\overline{\boldsymbol{Z}}_n$. On the other hand, if the partial correlation of X and Y with respect to $\boldsymbol{Z}$ is 0, then total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ will converge to 0. As the total variation distance goes to 0, it becomes harder and harder for any general statistical procedure to distinguish the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ from the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$. Hence the power of any general test of conditional independence for the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ will be too poor to be useful.

As a special case of Corollary 1, when $\boldsymbol{Z}$ is empty, we can show that whether $\overline{X}_n$ and $\overline{Y}_n$ are independent is also determined by the covariance matrix, or more precisely, by the value of $\mathrm{Cov}(X, Y)$. The relation between independence and covariance is less complicated, thanks to the fact that if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. Basically, if $\mathrm{Cov}(X, Y) \neq 0$,

[5] This value is the total variation distance between the conditional distribution of (U, V) given $\boldsymbol{W}$ and the product of the conditional distributions of U given $\boldsymbol{W}$ and V given $\boldsymbol{W}$.

then $\overline{X}_n$ and $\overline{Y}_n$ are dependent for all n, because of the invariance of the covariance matrix. Moreover, the total variation distance between the joint distribution of $(\overline{X}_n, \overline{Y}_n)$ and the product of the marginal distributions of $\overline{X}_n$ and $\overline{Y}_n$ will converge to a non-zero value, which is the total variation distance between the joint distribution of (U, V) and the product of the marginal distributions of U and V. Therefore, we do not need to worry about the power of the test of independence. On the other hand, if $\text{Cov}(X, Y) = 0$, we need to consider two cases: If X and Y are also independent, then because the independent relation is invariant under aggregation, [6] the total variation distance between the joint distribution of $(\overline{X}_n, \overline{Y}_n)$ and the product of the marginal distributions of $\overline{X}_n$ and $\overline{Y}_n$ will remain 0 for all n, which is just fine. If X and Y are dependent, then the total variation distance between the joint distribution of $(\overline{X}_n, \overline{Y}_n)$ and the product of the marginal distributions of $\overline{X}_n$ and $\overline{Y}_n$ will converge to the total variation distance between the joint distribution of (U, V) and the product of the marginal distributions of U and V, which is 0. This means that regardless of whether X and Y are independent, insofar as $\text{Cov}(X, Y) = 0$, for large n, any genearl independence test will likely return that $\overline{X}_n$ and $\overline{Y}_n$ are independent.

An interesting implication of Corollary 1 and the local Markov theorem is that, if a variable X is a linear function of its parents $\boldsymbol{Z}$ and an independent error term, then the partial correlation of this variable and any non-parental

[6] This statement is universally true, regardless of the distribution of X and Y. To show this, we note that if X and Y are independent, then $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ are also independent, hence $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} Y_i$ are independent.

non-descendent variable Y with respect to $\boldsymbol{Z}$ must be 0. This is because we know that the conditional independence relation between X and Y given $\boldsymbol{Z}$ is preserved under aggregation, hence the total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ is always 0 for all n. Suppose $(\overline{X}_n, \overline{Y}_n, \overline{\boldsymbol{Z}}_n)$ converges weakly to a multivariate random variable $(U, V, \boldsymbol{W})$, then U and V must be independent conditional on $\boldsymbol{W}$. The partial correlation of U and V with respect to $\boldsymbol{W}$ then must be 0, hence the partial correlation of X and Y with respect to $\boldsymbol{Z}$ then must be 0 too. [7]

While the conditions for Theorem 3 and Corollary 1 seem to be quite general, they do not cover the class of discrete distributions. After all, the expression level of any type of gene in a cell, which is the number of mRNA transcripts for that gene at a moment, is an integer valued random variable. The continuous distributions could approximate a discrete distribution arbitrarily well, but only in term of the distribution function. (The total variation distance between a continuous distribution and a discrete distribution is always 1, regardless of how close the distribution functions of these

[7] Of course, we can also prove this directly. Let $X = \boldsymbol{c}^T\boldsymbol{Z} + \epsilon$, where ϵ is independent of Y. Then we have:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{Cov}(\boldsymbol{c}^T\boldsymbol{Z}, Y) \ = \ \boldsymbol{c}^T\mathrm{Cov}(\boldsymbol{Z}, Y) \\ &= \boldsymbol{c}^T\mathrm{Var}(\boldsymbol{Z})\mathrm{Var}(\boldsymbol{Z})^{-1}\mathrm{Cov}(\boldsymbol{Z}, Y) \\ &= \mathrm{Cov}(X, \boldsymbol{Z})\mathrm{Var}(\boldsymbol{Z})^{-1}\mathrm{Cov}(\boldsymbol{Z}, Y) \end{aligned}$$

Hence the partial covariance of X and Y with respect to $\boldsymbol{Z}$ is:

$$\mathrm{Cov}(X,Y) - \mathrm{Cov}(X, \boldsymbol{Z})\mathrm{Var}(\boldsymbol{Z})^{-1}\mathrm{Cov}(\boldsymbol{Z}, Y) = 0$$

two distribution are.) However, as we are going to show in the remaining part of this section, Theorem 3 and Corollary 1 could be extended to an important class of discrete distributions — the regular lattice distributions — which covers the possible distributions of the numbers of mRNA transcripts of any set of genes in a cell.

A lattice distribution for a random vector $\boldsymbol{X}$ is a discrete distribution that only assigns non-zero probabilities to points $\boldsymbol{x} = (x_1, \cdots, x_k)$ such that $x_i = mh_i + b_i$, where m is an integer, h_i a positive real value, and b_i a constant. If h_i is the largest positive real number such that X_i can only take values of the form $mh_i + b_i$, h_i is called the span of X_i. The regular lattice distribution is defined as:

Definition 1. *Suppose a random vector $\boldsymbol{X} = (X_1, \cdots, X_k)$ has a lattice distribution, and h_i is the span of the ith coordinate X_i. Then $\boldsymbol{X}$ has a regular lattice distribution if, for any $1 \leq i \leq k$, there are at least two vectors $\boldsymbol{x}^i = (x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_k)$ and $\boldsymbol{y}^i = (x_1, \cdots, x_{i-1}, y_i, x_{i+1}, \cdots, x_k)$, such that $|y_i - x_i| = h_i$, $P(\boldsymbol{X} = \boldsymbol{x}^i) > 0$, and $P(\boldsymbol{X} = \boldsymbol{y}^i) > 0$.*

Let $\phi(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}$, define $T = \{\boldsymbol{t} : |\phi(\boldsymbol{t})| = 1, \boldsymbol{t} \neq \boldsymbol{0}\}$. By Lemma 1, $|\phi(\boldsymbol{t})| = 1$ implies that $\boldsymbol{t}^T\boldsymbol{X} = b + 2m\pi$ a.s. for $m = 0, \pm 1, \cdots$. In particular, $\boldsymbol{t}^T(\boldsymbol{x}^i - \boldsymbol{y}^i) = t_i(y_i - x_i) = 2m_1\pi$ for some integer m_1. That is, either $t_i = 0$, or $|t_i| \geq 2\pi/|y_i - x_i| = 2\pi/h_i$. Thus, we have shown that, if $\boldsymbol{X}$ has a regular lattice distribution, $|\phi(\boldsymbol{t})| < 1$ if $0 < |t_i| < h_i$ for all $1 \leq i \leq k$.

Now we can extend Theorem 3 and Corollary 1 to the regular lattice distributions.

Theorem 4. *Let $\boldsymbol{X}_n$ be i.i.d. discrete random vectors with a lattice distribution that satisfies the regularity condition given above. Suppose $\boldsymbol{X}_n$ has mean $\mathbf{0}$ and a non-singular covariance matrix $\boldsymbol{\Sigma}_X$. Let h_i be the span of the marginal distribution of the ith coordinate of $\boldsymbol{X}_n$, $\phi(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}_n$, and $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i/\sqrt{n}$. Then the probability mass functions $p_n(\boldsymbol{x})$ of $\overline{\boldsymbol{X}}_n$ converge uniformly to the density g of a multivariate normal distribution with 0 mean and covariance matrix $\boldsymbol{\Sigma}_X$ in the following way:*

$$\sup_x \left[\frac{n^{k/2}}{\prod_{i=1}^k h_i} p_n(\boldsymbol{x}) - g(\boldsymbol{x}) \right] \to 0 \tag{2.3}$$

Corollary 2. *Let $\{(X_n, Y_n, \boldsymbol{Z}_n)\}$ be a sequence of i.i.d. $k+2$ dimensional random vector with mean $\mathbf{0}$ and nonsingular covariance matrix $\boldsymbol{\Sigma}$. Suppose that $(X_n, Y_n, \boldsymbol{Z}_n)$ has a regular lattice distribution with a nonsingular covariance matrix $\boldsymbol{\Sigma}$. Let $\overline{X}_n = (\sum_{i=1}^n X_i)/\sqrt{n}$, $\overline{Y}_n = (\sum_{i=1}^n Y_i)/\sqrt{n}$, and $\overline{\boldsymbol{Z}}_n = (\sum_{i=1}^n \boldsymbol{Z}_i)/\sqrt{n}$, and $(U, V, \boldsymbol{W})$ be a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ converges to the total variation distance between the conditional distribution of (U, V)*

given $\boldsymbol{W}$ *and the product of the conditional distributions of* U *given* $\boldsymbol{W}$ *and* V *given* $\boldsymbol{W}$ *almost surely with respect to the measure induced by* $\boldsymbol{W}$*.*

The implication of Corollary 2 is similar to that of Corollary 1, except that it is applied to the regular lattice distributions.

Combining Corollaries 1 and 2, we have shown that, if given only the data about the summed gene expression levels from a large number of cells, we could learn virtually nothing about the exact causal models for the gene expression levels in a single cell, which has a lattice distribution, or the approximated continuous model, except the mean vector and the covariance matrix.

2.4 Estimate correlation matrix from noisy aggregation data

In section 2 of this chapter, it has been shown that, except for some special cases, we should not expect that the conditional independence relations among the expression levels of the genes in a single cell would be the same as the relations among the summed expression levels from an aggregate of multiple cells. In section 3, it was shown that, for two general classes of distributions, when a large number of cells are aggregated, the independence and conditional independence relations among the summed genes expression levels from the aggregated cells are essentially determined by the covariance matrix, or more precisely, by the covariance and partial covariances, of the expression levels of genes in a single cell. Given that it is typically the case, for the current technologies such as microarray or SAGE, that often hundreds of thousands of cells are used in a single measurement, it seems that in principle we are not going to learn the conditional independence information among the expression levels of the genes in a single cell, unless we were to make the biologically implausible assumption that in the true models for the gene expression levels in a single cell, the partial correlation between the expression levels of two genes with respect to other genes is 0 if and only if the two genes are independent conditional on other genes.

Nevertheless, we do know that two important features of the joint distribution of the gene expression levels—the mean vector and the covariance matrix—are invariant under aggregation up to a simple linear transformation, and we know that non-zero covariance does imply dependent relation. Therefore, theoretically, we can always claim that the expression levels of two genes are dependent if we find that the covariance of the summed expression levels of these two genes from a large number of cell is non-zero.

Unfortunately, even such a weak statement is problematic in practice, at least if we are going to use one of the two popular technologies, i.e., microarray or SAGE. The main reason is that, compared to the measurement error of the current technologies, the covariance between the summed expression levels of any pair of genes from an aggregate of a large number of cells is too small to be reliably estimated.

Let us first look at the SAGE data. A typical SAGE experiment needs 10^8 cells (Velculescu et al., 1997), and a yeast cell contains roughly 15000 mRNA transcripts (Hereford & Rosbash, 1977). Using some modified protocols, such as microSAGE, the number of cells can be reduced to 10^5 (Datson et al, 1999). Typically the result of a SAGE experiment is a library consisting of 30000 tags. Consider the following experiment: 10^5 cell, each with 15000 mRNA transcripts, are used as input, and the output is a SAGE library containing 30000 tags. Let X_i and Y_i represent respectively the numbers of mRNA transcripts of two genes A and B in the ith cell, and S and T the counts of tags for A and B in the resulting SAGE library. Suppose that $\mathrm{E}[X_i] = \mathrm{E}[Y_i] = 15$, $\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 225$, and $\mathrm{Cov}(X_i, Y_i) = 112.5$, (hence $\mathrm{Corr}(X_i, Y_i) = 0.5$). Let $\hat{p} = \sum_{i=1}^{100000} X_i/(1.5 \times 10^9)$, and $\hat{q} = \sum_{i=1}^{100000} Y_i/(1.5 \times 10^9)$. Assuming the PCR is unbiased, ignoring the sequencing error, conditional on $(\hat{p}, \hat{q})$, it can be shown that: [8]

$$\begin{aligned}
\mathrm{Var}(S|\hat{p}) &\approx 30000\,\hat{p}(1-\hat{p}) \\
\mathrm{Var}(T|\hat{q}) &\approx 30000\,\hat{q}(1-\hat{q}) \\
\mathrm{Cov}(S,T|\hat{p},\hat{q}) &\approx 30000\,\hat{p}\hat{q}
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\mathrm{Var}(S) &= \mathrm{E}[\mathrm{Var}(S|\hat{p})] + \mathrm{Var}(\mathrm{E}[S|\hat{p}]) \approx 30 \\
\mathrm{Var}(T) &= \mathrm{E}[\mathrm{Var}(T|\hat{q})] + \mathrm{Var}(\mathrm{E}[T|\hat{q}]) \approx 30 \\
\mathrm{Cov}(S,T) &= \mathrm{E}[\mathrm{Cov}(S,T|\hat{p},\hat{q})] + \mathrm{Cov}(\mathrm{E}[S|\hat{p}], \mathrm{E}[T|\hat{q}]) \approx -2.55 \times 10^{-2}
\end{aligned}$$

[8] For the details of the proof, see Chapter 3.

which implies that $\mathrm{Corr}(S,T) \approx -8.5 \times 10^{-4}$. On the other hand, if we assume that $\mathrm{Cov}(X_i, Y_i) = 0$, and everything else remains the same, the correlation between S and T would be -1×10^{-3}. Thus to test the null hypothesis that $\mathrm{Corr}(X_i, Y_i) = 0$ versus the alternative that $\mathrm{Corr}(X_i, Y_i) = 0.5$, we have to test $\mathrm{Corr}(S,T) = -1\times 10^{-3}$ versus $\mathrm{Corr}(S,T) = -8.5\times 10^{-4}$. Using Fisher's z transformation, the sample size must be greater than 1.7×10^8 so that the rates of both type I and II errors are approximately 15%. [9] That is, we need to perform at least 1.7×10^8 SAGE experiments so that we can detect a rather strong correlation of 0.5 between the expression levels of two genes. (Note that if there were no measurement errors, to test whether $\mathrm{Corr}(X_i, Y_i) = 0$ or $\mathrm{Corr}(X_i, Y_i) = 0.5$, using Fisher's z transformation, we would only need a sample of size 19 to control the rates of the two types of error at the level of approximately 15%.) In practice, the problem is even more difficult, because the correlation between two dependent genes could be smaller, and the alternative hypothesis should be $\mathrm{Corr}(X_i, Y_i) \neq 0$.

It is difficult to estimate how many microarray measurements are required so that we can test reliably whether $\mathrm{Corr}(X_i, Y_i) = 0$, because so far all the statistical models for the microarray data treat the expression levels of the genes as constants. However, it is generally believed that, while relatively cheap and fast, the microarray experiments usually provide qualitative measurements of the gene expression levels, in contrast to the quantitative

[9] Let z be the test statistic. Under the null, $\mathrm{E}_0[z] \approx -0.001$, under the alternative, $\mathrm{E}_1[z] \approx -0.00085$. The variance of z is approximately $1/(n-3)$, where n is the sample size. The level 15% test will reject the null if $z > -0.000925$.

nature of the SAGE technology. Our own experience with the two technologies also suggests that the quality of the data from microarray experiments usually is not as good as the SAGE data. Therefore, we may expect that we would need even more experiments to test whether $\mathrm{Corr}(X_i, Y_i) = 0$.

Thus we have reached the conclusion of this section and the whole chapter: In theory, the data obtained using current technologies such as microarray and SAGE cannot be used to identify the conditional independence relations among the expression levels of the genes in a single cell, though they could be used to estimate the covariance matrix of the gene expression levels. In practice, these data cannot even be used to estimate the covariance matrix of the gene expression levels in a single cell, unless we have an astronomical number of measurements. Thus, the only thing we can learn reliably from these data is the mean of the gene expression levels in a single cell.

This conclusion appears to conflict with many reports of successful machine learning searches for regulatory structure. In many cases, however, the successes are with simulated data in which the simulated values for individual cell representatives are not summed in forming the simulated measured values, and are therefore unfaithful to the actual measurement processes. In several other cases results with real data are not independently confirmed, but merely judged plausible. Rarely, results are obtained that agree with independent biological knowledge; in these cases the actual regulatory structure among the genes considered may approximately satisfy invariance of conditional independence for summed variables, or the procedures may simply have been lucky. Feasible, economical techniques for measuring concentrations of transcripts in single cells could make machine learning techniques based on associations of expressions valuable in identifying regulatory structure, but such techniques are not yet available. In the meanwhile, absent biological evidence that regulatory dependencies have the requisite invariance over sums of variables, there seems little warrant for thinking accurate methods are possible for inferring regulatory structures that depend on conditional associations.

Of course, there are other ways to determine the networks of regulatory relationships among genes. One approach, the intervention approach (Yuh, et al., 1998; Ideker, et al., 2001; Davidson, et al., 2002, and Yoo et al., 2002), experimentally suppresses (or enhances) the expression of one or more genes, and measures the resulting increased or decreased expression of other genes. A single knockout of gene A resulting in changed expression of genes B and C, for example, implies that either A regulates both B and C directly, or A regulates B which in turn regulates C, etc. The method, while laborious, has proved fruitful in unraveling small pieces of the regulatory networks of several species. Its chief disadvantage is that each experiment provides information only about the effects of the manipulated gene or genes, and it is often impossible to distinguish the direct effect from indirect effect with a single experiment. To identify a regulatory network, the number of experiments required will be super exponential in the number of distinct genes in the network.

Another promising approach is the genome-wide location analysis (Ren et al, 2000). The basic idea is to use formaldehyde to cross-link proteins and nuclei acids in living cells. The cells then are lysed and sonicated. The DNA fragments bound by certain proteins, which represent the promoter regions of the genes regulated by these proteins, are then enriched by immunoprecipitation with corresponding antibodies. The cross-links are then reversed, the DNA fragments are purified, amplified, and identified, and their concentration levels are measured (Orlando, 2000). This technology allows direct monitor of the protein-DNA interactions, and has been used to construct the regulatory network of yeast (Lee et al, 2002).

Using the above two experimental approaches, we can make inference of regulatory network without the knowledge of the statistical associations among the expression levels of the genes in a single cell. Of course we still need to know the mean expression levels of each gene (under certain conditions), which, fortunately, could be estimated from the gene expression data obtained by the microarray and the SAGE technologies. In the next two chapters, we are going to focus on the SAGE technology, construct a statistical model for the SAGE data, and explore the various applications of this statistical model.

2.5 Appendix: Proofs

Proofs for section 2

Theorem 1 (Local Markov Theorem). *Given an acyclic graph G representing the causal relations among a set $\boldsymbol{V}$ of random variables. Let $Y, X^1, \cdots, X^k \in \boldsymbol{V}$, and $\boldsymbol{X} = \{X^1, \cdots, X^k\}$ be the set of parents of Y in G. If $Y = \boldsymbol{c}^T\boldsymbol{X} + \epsilon$,* [10] *where $\boldsymbol{c}^T = (c^1, \cdots, c^k)$, and ϵ is a noise term independent of all non-descendents of Y, then Y is independent of all its non-parents, non-descendents conditional on its parents $\boldsymbol{X}$, and this relation holds under aggregation.*

Proof:

Let $\boldsymbol{U}$ be the set of the variables in $\boldsymbol{V}$ that are neither parents nor descendents of Y. That Y is independent of $\boldsymbol{U}$ conditional on its parents $\boldsymbol{X}$ is a direct consequence of the local Markov condition for acyclic graphs (Spirtes, et al, 2001).

Let Y_i, ϵ_i, $\boldsymbol{X}_i$, and $\boldsymbol{U}_i$ be the i^{th} i.i.d. copy of Y, ϵ, $\boldsymbol{X}$, and $\boldsymbol{U}$ respectively, we have,

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n (\boldsymbol{c}^T\boldsymbol{X}_i + \epsilon_i) = \boldsymbol{c}^T \sum_{i=1}^n \boldsymbol{X}_i + \sum_{i=1}^n \epsilon_i$$

Clearly, $(\epsilon_1, \cdots, \epsilon_n)$ is independent of $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{U}_1, \cdots, \boldsymbol{U}_n)$. This means that $\sum_{i=1}^n \epsilon_i$ is independent of $(\sum_{i=1}^n \boldsymbol{U}_i, \sum_{i=1}^n \boldsymbol{X}_i)$, which again implies that $\sum_{i=1}^n \epsilon_i$ is independent of $\sum_{i=1}^n \boldsymbol{U}_i$ conditional on $\sum_{i=1}^n \boldsymbol{X}_i$. Con-

[10] In this and the next theorems, we shall use the same bold face symbol to represent both a set of variables, and a vector of that set of variables.

sequently, $\boldsymbol{c}^T \sum_{i=1}^n \boldsymbol{X}_i + \sum_{i=1}^n \epsilon_i$ is independent of $\sum_{i=1}^n \boldsymbol{U}_i$ given $\sum_{i=1}^n \boldsymbol{X}_i$. (Note that $\boldsymbol{c}^T \sum_{i=1}^n \boldsymbol{X}_i$ is a constant conditional on $\sum_{i=1}^n \boldsymbol{X}_i = \boldsymbol{x}$, where $\boldsymbol{x}$ is an arbitrary constant vector.)

□

Theorem 2 (Markov Wall Theorem). *Given an acyclic graph G representing the causal relations among a set $\boldsymbol{V}$ of random variables. Let $\boldsymbol{X} = \{X^1, \cdots, X^h\}$, $\boldsymbol{Y} = \{Y^1, \cdots, Y^k\}$, $\boldsymbol{W} = \{W^1, \cdots, W^m\}$, and $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{W} = \boldsymbol{V}$. Suppose that the following three conditions hold:*

1. *The joint distribution of $X^1, \cdots, X^h$, Y^1, $\cdots$, Y^k is multivariate normal with nonsingular covariance matrix.*

2. *For $i = 1, \cdots, k$, Y^i is neither a parent, nor a child, of any variable $W^j \in \boldsymbol{W}$. That is, there is no direct edge between a variable in $\boldsymbol{Y}$ and a variable in $\boldsymbol{W}$.*

3. *For $i = 1, \cdots, h$, X^i is not a child of any variable $W^j \in \boldsymbol{W}$. That is, if there is an edge between a variable in $\boldsymbol{X}$ and a variable in $\boldsymbol{W}$, the direction of the edge must be from the variable in $\boldsymbol{X}$ to the variable in $\boldsymbol{W}$.*

Then conditional on $\boldsymbol{X}$, $\boldsymbol{Y}$ is independent of $\boldsymbol{W}$, and this relation holds under aggregation.

Proof:

The conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{W}$ given $\boldsymbol{X}$ is obvious, because

$\boldsymbol{W}$ can be represented as a function of $\boldsymbol{X}$ and some other random variables independent of $(\boldsymbol{X} \cup \boldsymbol{Y})$. [11]

Now let $\boldsymbol{Z} = (X^2, \cdots, X^h, Y^1, \cdots, Y^k)^T$, suppose the joint distribution of X^1 and $\boldsymbol{Z}$ is:

$$\begin{bmatrix} X^1 \\ \boldsymbol{Z} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \vec{\nu} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^T \\ \vec{\alpha} & \boldsymbol{\Sigma}_Z \end{bmatrix} \right)$$

Let $\boldsymbol{Z}_i = (X_i^2, \cdots, X_i^h, Y_i^1, \cdots, Y_i^k)^T$, which is the i^{th} i.i.d. copy of $\boldsymbol{Z}$, we are going to show that X_1^1 is independent of $\sum_{i=1}^n \boldsymbol{Z}_i$ given $\sum_{i=1}^n X_i^1$. First, let us see the joint distribution of $X_1^1, \sum_{i=1}^n X_i^1$, and $\sum_{i=1}^n \boldsymbol{Z}_i$:

$$\begin{bmatrix} X_1^1 \\ \sum_{i=1}^n X_i^1 \\ \sum_{i=1}^n \boldsymbol{Z}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ n\mu \\ n\vec{\nu} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1^2 & \vec{\alpha}^T \\ \sigma_1^2 & n\sigma_1^2 & n\vec{\alpha}^T \\ \vec{\alpha} & n\vec{\alpha} & n\boldsymbol{\Sigma}_Z \end{bmatrix} \right)$$

We claim that conditional on $\sum_{i=1}^n X_i^1 = nx$ and $\sum_{i=1}^n \boldsymbol{Z}_i = n\vec{z}$, the mean of X_1^1 is x.

Note that:

$$\begin{aligned} &\mathrm{E}[X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \boldsymbol{Z}_i = n\vec{z}] = \\ &\quad \mu + \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^T \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^T \\ n\vec{\alpha} & n\boldsymbol{\Sigma}_Z \end{bmatrix}^{-1} \begin{bmatrix} nx - n\mu \\ n\vec{z} - n\vec{\nu} \end{bmatrix} \end{aligned}$$

Let $\vec{\beta}^T = n\vec{\alpha}^T (n\boldsymbol{\Sigma}_Z)^{-1}$, $\gamma = 1/(n\sigma_1^2 - \vec{\beta}^T n\vec{\alpha})$, inverting by partition, we have:

[11] More precisely, these variables are the exogenous variables in $\boldsymbol{W}$ and the independent noise terms associated with the endogenous variables in $\boldsymbol{W}$.

$$\begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^T \\ n\vec{\alpha} & n\mathbf{\Sigma}_Z \end{bmatrix}^{-1} =$$
$$\begin{bmatrix} \gamma & -\gamma\vec{\beta}^T \\ -\gamma\vec{\beta} & (n\mathbf{\Sigma}_Z)^{-1}[I + (n\vec{\alpha})\gamma\vec{\beta}^T] \end{bmatrix}$$

It then can be shown that:

$$\begin{bmatrix} \sigma_1^2 & \vec{\alpha}^T \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^T \\ n\vec{\alpha} & n\mathbf{\Sigma}_Z \end{bmatrix}^{-1} = \begin{bmatrix} 1/n & \vec{0}^T \end{bmatrix}$$

It then follows:

$$\begin{aligned} &\mathrm{E}[X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \boldsymbol{Z}_i = n\vec{z}] \\ &= \mu + \begin{bmatrix} 1/n & \vec{0}^T \end{bmatrix} \begin{bmatrix} nx - n\mu \\ n\vec{z} - n\vec{\nu} \end{bmatrix} = x \end{aligned}$$

The conditional variance of X_1^1 given $\sum_{i=1}^n X_i^1 = nx$ and $\sum_{i=1}^n \boldsymbol{Z}_i = n\vec{z}$ is:

$$\begin{aligned} &\mathrm{Var}\left(X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \boldsymbol{Z}_i = n\vec{z} \right) \\ &= \sigma_1^2 - \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^T \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^T \\ n\vec{\alpha} & n\mathbf{\Sigma}_Z \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 \\ \vec{\alpha} \end{bmatrix} \\ &= \frac{n-1}{n}\sigma_1^2 \end{aligned}$$

Thus, we have shown that both the conditional mean and the conditional variance of X_1^1 is constant in $n\vec{z}$. Given that the conditional distribution of X_1^1 is normal, this implies that X_1^1 is independent of $\sum_{i=1}^n \boldsymbol{Z}_i$ given

$\sum_{i=1}^n X_i^1$. Note that by the same argument, we could show that, conditional on $\sum_{i=1}^n X_i^1$, X_1^1 is independent of $\sum_{i=1}^n X_i^2, \cdots, \sum_{i=1}^n X_i^h$. Let $\boldsymbol{X}_i$ be the i^{th} copy of $\boldsymbol{X}$, it follows that, conditional on $\sum_{i=1}^n \boldsymbol{X}_i$, X_1^1 is independent of $\sum_{i=1}^n \boldsymbol{Y}_i$. Because the choice of X_1^1 is arbitrary, we actually have shown that, conditional on $\sum_{i=1}^n \boldsymbol{X}_i$, X_i^j is independent of $\sum_{i=1}^n \boldsymbol{Y}_i$ for any $1 \leq i \leq n$ and $1 \leq j \leq h$. Moreover, the joint distribution of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ and $\sum_{i=1}^n \boldsymbol{Y}_i$ conditional on $\sum_{i=1}^n \boldsymbol{X}_i$ is multivariate normal, and for multivariate normal, marginal independence relations imply the joint independence relation. [12] It then follows that $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ is independent of $\sum_{i=1}^n \boldsymbol{Y}_i$ given $\sum_{i=1}^n \boldsymbol{X}_i$.

We note that $\boldsymbol{W}_i$, the i^{th} copy of $\boldsymbol{W}$, can be represented as a function of $\boldsymbol{X}_i$ and some other random variables independent of ($\boldsymbol{X}_1$, $\cdots$, $\boldsymbol{X}_n$, $\boldsymbol{Y}_1$, $\cdots$, $\boldsymbol{Y}_n$). Thus, as a function of $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ and other random variables independent of $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)$, $\sum_{i=1}^n \boldsymbol{W}_i$ is independent of $\sum_{i=1}^n \boldsymbol{Y}_i$ given $\sum_{i=1}^n \boldsymbol{X}_i$.

□

Proofs for section 2

Most of the lemmas in this section are multivariate versions of some well known facts about the univariate characteristic functions.

Lemma 1. *Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ be a random vector with characteristic function $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$. Then if $\boldsymbol{X}$ has a density with respect to the Lebesgue measure, $|\phi(\boldsymbol{t})|$ equals to 1 only when $\boldsymbol{t} = \boldsymbol{0}$*

[12] Suppose X, Y, Z are multivariate normal. If X is independent of Y, and X is also independent of Z, then X is independent of (Y, Z).

Proof:

Suppose for some $\boldsymbol{a} = (a_1, \cdots, a_k) \neq \mathbf{0}$, $|\phi(\boldsymbol{a})| = 1$. Let $\phi(\boldsymbol{a}) = \exp(ic)$ for some real number c. Without loss of generality, suppose $a_1 \neq 0$. Then the characteristic function of $\boldsymbol{X}_c = (X_1 - c/a_1, X_2, \cdots, X_k)$ is:

$$\mathrm{E}[\exp(i\boldsymbol{X}_c^T\boldsymbol{t})] = \exp\Big(-i\frac{ct_1}{a_1}\Big)\phi(t_1, \cdots, t_k) \tag{2.4}$$

Therefore, when $\boldsymbol{t} = \boldsymbol{a}$:

$$\begin{aligned} 1 &= \mathrm{E}[\exp(i\boldsymbol{X}_c^T\boldsymbol{a})] = \mathrm{E}[\cos(\boldsymbol{X}_c^T\boldsymbol{a}) + i\sin(\boldsymbol{X}_c^T\boldsymbol{a})] \\ &= \mathrm{E}[\cos(\boldsymbol{X}^T\boldsymbol{a} - c) + i\sin(\boldsymbol{X}^T\boldsymbol{a} - c)] \end{aligned}$$

This is possible only when $\mathrm{E}[\cos(\boldsymbol{X}^T\boldsymbol{a} - c)] = 1$, which implies that $\boldsymbol{X}^T\boldsymbol{a} = 2m\pi + c$ a.s. for $m = 0, \pm 1, \cdots$. Therefore, the distribution of $\boldsymbol{X}$ is not absolutely continuous with respect to the Lebesgue measure. □

Lemma 2. *Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ be a random vector with a bounded density with respect to the Lebesgue measure, and $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$ be the characteristic function of $\boldsymbol{X}$. Then $\phi(\boldsymbol{t})$ is integrable if $\phi(\boldsymbol{t}) \geq 0$.*

Proof:

Let $f(\boldsymbol{x})$ be the density of $\boldsymbol{X}$. Suppose $f(\boldsymbol{x}) \leq M$, where M is a constant. Let $g_\sigma(\boldsymbol{t})$ be the density of a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{I}_k$, then by the Parseval's identity:

$$\int \exp(-i\boldsymbol{t}^T\boldsymbol{u})\phi(\boldsymbol{t})g_\sigma(\boldsymbol{t})\,d\boldsymbol{t} = \int f(\boldsymbol{x})\exp(-(\sigma^2/2)(\boldsymbol{x}-\boldsymbol{u})^T(\boldsymbol{x}-\boldsymbol{u}))\,d\boldsymbol{x} \quad (2.5)$$

This implies that:

$$\begin{aligned} & \int \exp(-i\boldsymbol{t}^T\boldsymbol{u})\phi(\boldsymbol{t})\exp\left(-\frac{\boldsymbol{t}^T\boldsymbol{t}}{2\sigma^2}\right)d\boldsymbol{t} \\ = \;& (2\pi)^k \int f(\boldsymbol{x})\frac{1}{(\sqrt{2\pi})^k\sigma^{-1}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{u})^T(\boldsymbol{x}-\boldsymbol{u})}{2\sigma^{-2}}\right)d\boldsymbol{x} \\ \le \;& (2\pi)^k M \end{aligned}$$

Set $\boldsymbol{u} = \boldsymbol{0}$, we have:

$$\int \phi(\boldsymbol{t})\exp\left(-\frac{\boldsymbol{t}^T\boldsymbol{t}}{2\sigma^2}\right)d\boldsymbol{t} \le (2\pi)^k M \quad (2.6)$$

Given that $\phi(\boldsymbol{t}) \ge 0$, and $\phi(\boldsymbol{t})\exp[-(\boldsymbol{t}^T\boldsymbol{t})/(2\sigma^2)] \uparrow \phi(\boldsymbol{t})$ as $\sigma \to \infty$,

$$\int \phi(\boldsymbol{t})\exp\left(-\frac{\boldsymbol{t}^T\boldsymbol{t}}{2\sigma^2}\right)d\boldsymbol{t} \to \int \phi(\boldsymbol{t})\,d\boldsymbol{t} \quad (2.7)$$

Hence $\phi(\boldsymbol{t})$ is integrable and $\int \phi(\boldsymbol{t})\,d\boldsymbol{t} \le (2\pi)^k M$. □

Lemma 3. *Let $\boldsymbol{X} = (X_1, \cdots, X_k)$ be a random vector with a bounded density $f(\boldsymbol{x})$ with respect to the Lebesgue measure, and $\phi(\boldsymbol{t}) = \phi(t_1, \cdots, t_k)$ be the characteristic function of $\boldsymbol{X}$. Then $|\phi(\boldsymbol{t})|^n$ is integrable for all $n \ge 2$.*

Proof:

It sufficies to show that $|\phi(\boldsymbol{t})|^2$ is integragable.

First, we note that $|\phi(\boldsymbol{t})|^2 \geq 0$, and $|\phi(\boldsymbol{t})|^2$ is the characteristic function of $\boldsymbol{X}_1 - \boldsymbol{X}_2$, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are i.i.d. with density $f(\boldsymbol{x})$. The density of $\boldsymbol{X}_1 - \boldsymbol{X}_2$ is:

$$g(\boldsymbol{y}) = \int f(\boldsymbol{x}+\boldsymbol{y}) f(\boldsymbol{x})\, d\boldsymbol{x} \leq \sup_{\boldsymbol{x}} f(\boldsymbol{x}) < \infty \tag{2.8}$$

By Lemma 2, $|\phi(\boldsymbol{t})|^2$ is integragable. □

Theorem 3. *Let $\boldsymbol{X}_n$ be i.i.d. random vectors with 0 mean and non-singular covariance matrix $\boldsymbol{\Sigma}_X$, and $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i/\sqrt{n}$. Suppose the characteristic function $\phi(\boldsymbol{t}) = E[\exp(\boldsymbol{t}^T\boldsymbol{X})]$ is integratable, then $\overline{\boldsymbol{X}}_n$ have bounded continuous densities that converge uniformly to the density of a multivariate normal distribution with 0 mean and covariance matrix $\boldsymbol{\Sigma}_X$.*

Proof:

This theorem is a generalization of the well known fact about the convergence of density in the univariate case. The following proof is similar to the one (for the univariate case) given in Feller (1971).

Because $\phi(\boldsymbol{t})$ is integrable, the density of $\overline{\boldsymbol{X}}_n$ can be obtained by the inversion formula:

$$f_n(\boldsymbol{x}) = \left(\frac{1}{2\pi}\right)^k \int \exp(-i\boldsymbol{t}^T\boldsymbol{x})\phi(\boldsymbol{t}/\sqrt{n})^n\, d\boldsymbol{t} \tag{2.9}$$

We need to show that, uniformly over $\boldsymbol{x}$:

$$\int \left| \exp(-i\boldsymbol{t}^T\boldsymbol{x})\left(\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\, \boldsymbol{t})\right)\right| d\boldsymbol{t} \to 0 \tag{2.10}$$

where $\exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})$ is the characteristic function of the multivariate normal distribution with 0 mean and covariance matrix $\boldsymbol{\Sigma}_X$.

Compare $\phi(\boldsymbol{t})$ with $\exp(-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})$. They are both equal to 1 when evalauted at $\boldsymbol{t} = \boldsymbol{0}$, their first derivatives are both equal to $\boldsymbol{0}$ when evaluated at $\boldsymbol{t} = \boldsymbol{0}$, and their second derivatives are $-\boldsymbol{\Sigma}_X$ and $-\frac{1}{2}\boldsymbol{\Sigma}_X$ when evaluated at $\boldsymbol{t} = \boldsymbol{0}$. Given that $\boldsymbol{\Sigma}_X$ is positive definite, there must be a postive δ such that $|\phi(\boldsymbol{t})| \leq \exp(-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})$ for $|\boldsymbol{t}| \leq \delta$. Let $h = \sup_{|\boldsymbol{t}|=\delta}\{\exp(-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\}$. It is easy to see that $h < 1$. On the other hand, by the Riemann-Lebesgue Theorem, (see Stein & Weiss 1971, p. 2), $\phi(\boldsymbol{t}) \to 0$ as $|\boldsymbol{t}| \to \infty$. Thus, given that $|\phi(\boldsymbol{t})| < 1$ for all $\boldsymbol{t} \neq \boldsymbol{0}$, $|\phi(\boldsymbol{t})|$ must achieve a maximum m on $|\boldsymbol{t}| \geq \delta$, where $m < 1$.

Let $\epsilon > 0$. First we choose a c such that $\int_{|\boldsymbol{t}|\geq c}\exp(-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\,d\boldsymbol{t} < \epsilon$. Given the fact that $[\phi(\boldsymbol{t}/\sqrt{n})]^n \to \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})$ uniformly on any compact set, there is an N_1 such that, for all $n \geq N_1$,

$$\int_{|\boldsymbol{t}|\leq c}\left|\exp(-i\boldsymbol{t}^T\boldsymbol{x})\Big(\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\Big)\right|\,d\boldsymbol{t} < \epsilon \qquad (2.11)$$

Now choose N_2 such that for all $n \geq N_2$, $\sqrt{n}m^{n-1}\int|\phi(\boldsymbol{t})|\,d\boldsymbol{t} < \epsilon$, and $\sqrt{n}\delta > c$. We have, for $n \geq \max(N_1, N_2)$:

$$\begin{aligned}&\int\left|\exp(-i\boldsymbol{t}^T\boldsymbol{x})\Big(\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\Big)\right|\,d\boldsymbol{t}\\ &\leq \int\left|\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\right|\,d\boldsymbol{t}\end{aligned}$$

$$
\begin{aligned}
\leq\ & \int_{|\boldsymbol{t}|\leq c} \left|\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\right| d\boldsymbol{t} \\
& + \int_{|\boldsymbol{t}|>c} \left[\exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t}) + \left|\phi(\boldsymbol{t}/\sqrt{n})^n\right|\right] d\boldsymbol{t} \\
\leq\ & \epsilon + \epsilon + \int_{c<|\boldsymbol{t}|\leq\sqrt{n}\delta} \exp(-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\,d\boldsymbol{t} + m^{n-1}\int_{|\boldsymbol{t}|>\sqrt{n}\delta} \left|\phi(\boldsymbol{t}/\sqrt{n})\right| d\boldsymbol{t} \\
\leq\ & 3\epsilon + \sqrt{n}m^{n-1}\int|\phi(\boldsymbol{t})|\,d\boldsymbol{t} \leq 4\epsilon
\end{aligned}
$$

□

Theorem 4. *Let $\boldsymbol{X}_n$ be i.i.d. discrete random vectors with a lattice distribution that satisfies the regularity condition given above. Suppose $\boldsymbol{X}_n$ has mean $\mathbf{0}$ and a non-singular covariance matrix $\boldsymbol{\Sigma}_X$. Let h_i be the span of the marginal distribution of the ith coordinate of $\boldsymbol{X}_n$, $\phi(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}_n$, and $\overline{\boldsymbol{X}}_n = \sum_{i=1}^n \boldsymbol{X}_i/\sqrt{n}$. then the probability mass functions $p_n(\boldsymbol{x})$ of $\overline{\boldsymbol{X}}_n$ converge uniformly to the density g of a multivariate normal distribution with 0 mean and covariance matrix $\boldsymbol{\Sigma}_X$ in the following way:*

$$\sup_x \left[\frac{n^{k/2}}{\prod_{i=1}^k h_i} p_n(\boldsymbol{x}) - g(\boldsymbol{x})\right] \to 0 \tag{2.12}$$

Proof: Given that the span of the marginal distribution of X_i is h_i, the marginal distribution of the ith coordinate of $\boldsymbol{X}_n$ must have a lattice distribution with span $h_i/\sqrt{n}$. Let $\phi(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}_n$, the probability mass function for $\boldsymbol{X}_n$ is:

$$p_n(\boldsymbol{x}) = \frac{\prod_{i=1}^k h_i}{(2\sqrt{n}\pi)^k} \int_{\frac{-\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{\frac{-\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \exp(-i\boldsymbol{t}^T\boldsymbol{x})\phi(\boldsymbol{t}/\sqrt{n})^n \, d\boldsymbol{t} \qquad (2.13)$$

Let $\phi(\boldsymbol{t})$ be the characteristic function of $\boldsymbol{X}_n$, we need to show that, uniformly on $\boldsymbol{x}$,

$$\int_{\frac{-\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{\frac{-\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \exp(-i\boldsymbol{t}^T\boldsymbol{x})[\phi(\boldsymbol{t}/\sqrt{n})^n \, d\boldsymbol{t} - \int \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})] \, d\boldsymbol{t} \to 0 \qquad (2.14)$$

It is easy to see that it suffices to prove that:

$$\int_{\frac{-\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{\frac{-\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \left|\phi(\boldsymbol{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_X\,\boldsymbol{t})\right| \, d\boldsymbol{t} \to 0 \qquad (2.15)$$

The proof will be essentially the same as the proof for Theorem 3, except that here the Riemann-Lebesgue Theorem does not hold. Instead, we use the fact that under the regularity condition, $|\phi(\boldsymbol{t})| < 1$ if $0 < |t_i| < h_i$ for all $1 \leq i \leq k$. □

Corollary 2. *Let $\{(X_n, Y_n, \boldsymbol{Z}_n)\}$ be a sequence of i.i.d. $k+2$ dimensional random vector, where $\boldsymbol{Z}_n$ is a k dimensional random vector. Suppose that $(X_n, Y_n, \boldsymbol{Z}_n)$ has a regular lattice distribution with a nonsingular covariance matrix $\boldsymbol{\Sigma}$. Let $\overline{X}_n = (\sum_{i=1}^n X_i)/\sqrt{n}$, $\overline{Y}_n = (\sum_{i=1}^n Y_i)/\sqrt{n}$, and $\overline{\boldsymbol{Z}}_n = (\sum_{i=1}^n \boldsymbol{Z}_i)/\sqrt{n}$, then the total variation distance between the conditional distribution of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$ and the product of the conditional distributions of $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$ and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ converges to 0.*

Proof: Without loss of generality, suppose the spans for $(X_n, Y_n, \boldsymbol{Z}_n)$ are $h_1, h_2, \cdots, h_{k+2}$ respectively, and that the lattice points of the distribution, i.e., the values that $(X_n, Y_n, \boldsymbol{Z}_n)$ could possible take, are of the form $(m_1 h_1 + c_1, m_2 h_2 + c_2, \cdots, m_{k+2} h_{k+2} + c_{k+2})$, where $m_1, \cdots, m_{k+2}$ are arbitrary intergers, and $0 \leq c_i < h_i$ for $i = 1, \cdots, k+2$.

Let $F^n_{X,Y|\boldsymbol{Z}}$, $F^n_{X|\boldsymbol{Z}}$, and $F^n_{Y|\boldsymbol{Z}}$ be the conditional distributions of $(\overline{X}_n, \overline{Y}_n)$ given $\overline{\boldsymbol{Z}}_n$, $\overline{X}_n$ given $\overline{\boldsymbol{Z}}_n$, and $\overline{Y}_n$ given $\overline{\boldsymbol{Z}}_n$ respectively. Clearly they are also lattice distributions. We are going to approximate these three conditional distributions with three continuous distributions $G^n_{X,Y|\boldsymbol{Z}}$, $G^n_{X|\boldsymbol{Z}}$, and $G^n_{Y|\boldsymbol{Z}}$. The basic idea is to transform the probability mass functions of the latttice distributions into the probability density functions (w.r.t. the Lebesgue measure) of the continuous distributions. Generally speaking, to approximate a m dimensional lattice distribution with a continuous distribution, we shall first divide the m dimensional Eucidean space into identical m dimensional rectangles such that the lengths of the "edges" of each rectangle are equal to the spans of the lattice distribution, and that at the geometric center of each rectangle is a lattice point. The probability density function then will be uniform within each of the rectangles, and the total mass for each rectangle will be the same as the mass assigned to the lattice point in the center of that rectangle by the corresponding lattice distribution. The three densities are given as:

$$\begin{aligned}
g^n_{X,Y|\boldsymbol{z}}(x,y|\boldsymbol{z}) &= \frac{n}{h_1h_2}\sum_{(m_1,m_2)\in\mathbb{Z}^2} I_{C^{x,y,n}_{m_1,m_2}}(x,y) \\
&\qquad P((\overline{X}_n,\overline{Y}_n)\in C^{x,y,n}_{m_1,m_2} \mid \overline{\boldsymbol{Z}}_n = d_n(\boldsymbol{z})) \\
g^n_{X|\boldsymbol{z}}(x|\boldsymbol{z}) &= \frac{\sqrt{n}}{h_1}\sum_{m_1\in\mathbb{Z}} I_{C^{x,n}_{m_1}}(x)P(\overline{X}_n\in C^{x,n}_{m_1} \mid \overline{\boldsymbol{Z}}_n = d_n(\boldsymbol{z})) \\
g^n_{Y|\boldsymbol{z}}(y|\boldsymbol{z}) &= \frac{\sqrt{n}}{h_2}\sum_{m_2\in\mathbb{Z}} I_{C^{y,n}_{m_2}}(y)P(\overline{Y}_n\in C^{y,n}_{m_2} \mid \overline{\boldsymbol{Z}}_n = d_n(\boldsymbol{z}))
\end{aligned}$$

where

$$\begin{aligned}
C^{x,y,n}_{m_1,m_2} &= \Bigg\{(w_1,w_2): \\
&\qquad \frac{(m_i-0.5)h_i+nc_i}{\sqrt{n}} < w_i \le \frac{(m_i+0.5)h_i+nc_i}{\sqrt{n}}, i=1,2\Bigg\} \\
C^{x,n}_{m_1} &= \left(\frac{(m_1-0.5)h_1+nc_1}{\sqrt{n}}, \quad \frac{(m_1+0.5)h_1+nc_1}{\sqrt{n}}\right] \\
C^{y,n}_{m_2} &= \left(\frac{(m_2-0.5)h_2+nc_2}{\sqrt{n}}, \quad \frac{(m_2+0.5)h_2+nc_2}{\sqrt{n}}\right]
\end{aligned}$$

and $d_n(w_3,\cdots,w_{k+2}) = (v_3,\cdots,v_{k+2})$, with $v_i = [\text{ceil}((\sqrt{n}w_i - nc_i)/h_i - 0.5)h_i + nc_i]/\sqrt{n}$ for $3 \le i \le k+2$. [13]

Let $p^n_{X,Y|Z}$, $p^n_{X|Z}$, and $p^n_{Y|Z}$ be the probability mass functions for $F^n_{X,Y|Z}$, $F^n_{X|Z}$, and $F^n_{Y|Z}$ respectively. Let $q^n_{X,Y|Z} = p^n_{X|Z}p^n_{Y|Z}$, and $Q^n_{X,Y|Z}$ be the correpsonding distribution function. Let $h^n_{X,Y|Z} = g^n_{X|Z}g^n_{Y|Z}$, and $H^n_{X,Y|Z}$ be the correpsonding distribution function. It is easy to see that the total

[13] ceil(x) returns the smallest integer that is greater than or equal to x. Thus $d_n(\boldsymbol{z})$ returns the lattice point of $\overline{\boldsymbol{Z}}_n$ closest to $\boldsymbol{z}$. In cases of tie, it will return the smallest. This way we have defined the conditional distribution for the cases where $P(\overline{\boldsymbol{Z}}_n = \boldsymbol{z}) = 0$.

variation of the signed measure $Q^n_{X,Y|Z} - F^n_{X,Y|Z}$ is the same as the total variation of the signed measure $H^n_{X,Y|Z} - G^n_{X,Y|Z}$, i.e., $|Q^n_{X,Y|Z} - F^n_{X,Y|Z}| = |H^n_{X,Y|Z} - G^n_{X,Y|Z}|$. As a direct consequence of Theorem 4, we have $h^n_{X,Y|Z} - g^n_{X,Y|Z} \to 0$ as $n \to \infty$, although the convergence may be not uniform. By the bounded convergence theorem, we have $|H^n_{X,Y|Z} - G^n_{X,Y|Z}| \to 0$, hence $|Q^n_{X,Y|Z} - F^n_{X,Y|Z}| \to 0$, as $n \to \infty$. □

Chapter 3

Sampling, amplification, and resampling

Compared to other technologies for gene expression level measurement, such as microarray, SAGE has a distinct advantage. That is, we have a better understanding of most of the critical steps of the SAGE protocol. This makes it possible to construct a statistical model for the SAGE data that based on our knowledge about how the data are generated, rather than by simply trying to fit the data with some convenient models. In this chapter, I shall discuss a new sampling method for generating discrete data — sampling, amplification, and resampling (SAR) — that is a generalization of the three critical steps of the SAGE protocol. This new sampling scheme can be used to model not only the SAGE gene expression data, but also many other biological experiments involving Polymerase Chain Reaction (PCR). I shall derive the asymptotic distribution for the data generated by the SAR scheme. The results of this chapter provide a theoretical foundation for the detailed study of the SAGE data in the next chapter. Readers who are only interested in the practical analysis of the SAGE data can skip this chapter and go directly to the next chapter: SAGE data analysis.

The first section of this chapter gives the definition of the the sampling, amplification, and resampling scheme (SAR). In section 2, I study the asymptotic behavior of the amplification step, and prove some theorems about the asymptotic behavior of the ratio of the large sample means. Then in section 3, I derive the main result of this chapter, the asymptotic distribution of the discrete data generated by the SAR procedure. In the last section, I present several test statistics for some frequently used tests, and give the asymptotic distributions for these statistics.

3.1 Introduction

In their classic work on the multivariate discrete analysis, Bishop, Fienberg, and Holland (1975) discuss several popular sampling methods that generate multivariate discrete data (contingency tables). Basically, there are two types of sampling methods: the multinomial type, and the hypergeometric type. The multinomial type methods again include sampling methods that could generate the following three families of distributions: the multinomial sampling, which generates data of multinomial distributions; the Poisson sampling, which generates data of Poisson distributions; and the negative multinomial sampling, which generate data of negative multinomial distributions. These sampling methods are closely related to each other. For example, among the three multinomial type methods, the joint distribution of k independent Poisson random variables, conditional on their sum, is a k dimensional multinomial distribution. The multinomial distribution, on the other hand, could be seen as the limit of the multivariate hypergeometric distribution, and is often a good approximation for the latter when the population size is large compared to the sample size.

Because of the popularity of the above models, people may tend to treat any contingency table as being generated by one of these methods. However, there are many other types of data that do not fall in one of the above models. One such type of data are the gene expression level data generated by the SAGE experiments.

In a SAGE experiment, a sample of mRNA transcripts is extracted from a cell population, transcribed into cDNA clones. Then, from a specific site of each cDNA clone, a short 10 base long tag is cut. A certain number of cycles of PCR then are performed to amplify the tags. Finally, the tags are linked together to form longer sequences. Among these longer sequences, those of certain length that are suitable for sequencing are chosen and get sequenced. The counts of the tags contained in the sequenced sequences are reported as the experimental result, called the SAGE library.

While the generation of a SAGE data involves many steps, there are three steps critical to the statistical modeling of the SAGE data. These three steps are the collection of the sample cells from a certain cell population, PCR, and the collection of long sequences for sequencing. As an abstraction of these three steps, we can derive a new sampling scheme for generating multivariate discrete data: the *sampling, amplification, and resampling* (SAR) procedure. A typical sampling, amplification, and resampling procedure include the following steps:

1. Draw the original sample, which has either the multinomial, or the multivariate hypergeometric distribution.

2. Amplify the original sample. Each element in the original sample is amplified independently such that the integer valued amplification factors for each element are nonnegative and identically distributed with positive mean and finite variance. (Starting with a single element, let X be the total number of elements obtained after the amplification, then X is the amplification factor.) The amplified sample is called the intermediate sample.

3. Generate the final sample from the intermediate sample by drawing randomly *with* or *without* replacement. The final sample is also called the SAR sample.

Note that the generation of the final sample by sampling without replacement from the intermediate sample is complex. The problem is that the size of the intermediate sample is a random variable, hence the size of the final sample in general will also be a random variable. For example, suppose the initial plan is to draw a sample of size n, but the size of the intermediate sample is $n' < n$, then the final sample size will be n', instead of n. However, this is less an issue in asymptotic study if n is so selected that it is less than the size of the intermediate sample with probability one as the size of the original sample goes to infinity.

In the remaining part of this chapter, we shall analyze the asymptotic behavior of the data generated by the SAR procedure. The theorems proved for the SAR procedure could be used not only to model the SAGE data, but also to study the results of other biological experiments that employ PCR. [1]

3.2 Asymptotic distribution of the ratio in amplification

If the intermediate sample in the SAR scheme were obtained by multiplying the original sample by a factor k, then the relative frequencies of each category in the intermediate sample will be the same as the relative frequencies in the original sample. However, if the original sample is amplified

[1] For example, competitive RT-PCR (reverse transcription-polymerase chain reaction) is believed to be one of the most accurate methods of quantifying the mRNA expression. Our result about the asymptotic distribution of the ratio of sums of two sequences of iid random variables (see section 2) could be used to analyze the results obtained from the RT-PCR experiments. The basic idea of RT-PCR is that the target mRNA sample is mixed with a known mount of synthesized mRNA, which share the same primer pair for amplification with the target mRNA. The PCR procedure then is applied to the mixture. The amounts of the amplified target mRNA and the amplified synthesized mRNA then get measured. Let X_t and X_s be the amounts of the target mRNA and synthesized mRNA before PCR, and Y_t and Y_p the amounts of the target mRNA and synthesized mRNA after PCR. The value of X_s, Y_t, and Y_s are known. Our study gives the asymptotic distribution of Y_t/Y_s given X_t/X_s and some other parameters, including the number of cycles of PCR and the efficiency of PCR, both could be estimated experimentally. Thus, we could find a confidence interval for X_t/X_s, and hence a confidence interval for X_t.

by a noisy procedure, say, a branch process, conditional on the original sample, the relative frequencies of each category in the intermediate sample will be nondegenerate random variables. In this section we shall present the asymptotic distribution of the relative frequencies in the intermediate sample conditional on the original sample. But first, we show that for a specific type of amplification processes, the mean of the relative frequency of any category in the intermediate sample, conditional on the original sample, is exactly the same as the relative frequency in the original sample. This specific process is often used to model the PCR procedure.

Lemma 4. *Let $\{X_t\}$ and $\{Y_t\}$ be two independent branch processes with the following properties:*

1. *$X_{t+1} = X_t + U_t$, where U_t follows a binomial distribution with parameters (X_t, λ), for $0 < \lambda < 1$.*
2. *$Y_{t+1} = Y_t + V_t$, where V_t follows a binomial distribution with parameters (Y_t, λ).*

Let $P_{t+1} = \dfrac{X_{t+1}}{X_{t+1} + Y_{t+1}}$ and $P_t = \dfrac{X_t}{X_t + Y_t}$, then:

$$E[P_{t+1}|P_t] = P_t \tag{3.1}$$

It is easy to see that $\{P_t\}$ for $t = 0, 1, \cdots$ is a martingale, with respect to $\{\sigma(P_0), \sigma(P_0, P_1), \cdots\}$, the sigma fields generated by P_0, (P_0, P_1), etc. Hence for any $r > 0$, we have:

$$\mathrm{E}[P_r|P_0] = P_0 \tag{3.2}$$

From now on we shall make no specific assumptions about the distribution of the amplification factor. In most cases, we only assume that the amplification factor has positive mean and finite variance, as required by the definition of SAR.

To get the asymptotic distribution of the relative frequencies in the intermediate sample, we begin with a simpler case, where the original sample has two categories. Let the mean and the variance of the amplification factor be μ and σ^2 respectively, and the absolute frequencies of the first and the second categories in the original sample be n and r_n respectively. Then the following theorem gives the asymptotic distribution of the relative frequency of the first category in the intermediate sample.

Theorem 5. *Given a sequence of i.i.d. nonnegative random variables X_1, $\cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. Let r_n be a sequence of positive integers such that $n \leq r_n \leq Mn$ for some fixed M. Then:*

$$\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) \implies N\left(0, \frac{\sigma^2}{\mu^2}\right) \tag{3.3}$$

where $p_n = \dfrac{n}{n+r_n}$

If the amplification process is nondecreasing and bounded, then the amplification factor is bounded from below by a positive value, and also bounded from above. It can be shown that in this caes the variance of the relative frequency of the first category also converges.

Corollary 3. *Given the same condition as in Theorem 5, if* $E[X_i^4] < \infty$*, then:*

$$E\left[\frac{(n+r_n)^3}{nr_n}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2\right] \to \frac{\sigma^2}{\mu^2} \tag{3.4}$$

In the proof of Theorem 5, for convenience, we assume that $n \leq r_n \leq Mn$ for some M. This assumption is dropped in the following corollary.

Corollary 4. *If in Theorem 5 and Corollary 3, instead of requiring* $n \leq r_n \leq Mn$*, we require that* $Ln \leq r_n \leq Mn$*, where* L *is some positive real number, the conclusions still hold.*

The following corollary is obvious.

Corollary 5. *In Corollary 3, if we further assume that* $p_n = \dfrac{n}{n+r_n} \to p$*, where* $0 < p < 1$*, then:*

$$E\left[\sqrt{n+r_n}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)\right] \to 0 \tag{3.5}$$

$$E\left[(n+r_n)\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2\right] \to \frac{p(1-p)\sigma^2}{\mu^2} \tag{3.6}$$

Now we can give the asymptotic distribution of the relative frequencies of multiple categories in the intermediate sample, conditional on the original sample.

Theorem 6. *Theorem 5 can be generalized in the following way:*

Given a sequence of independent nonnegative random variables $X_1, \cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. For $n = 1, \cdots$, let $N_{n,1}, \cdots$, $N_{n,k+1}$ be positive integers such that $n = N_{n,1} \leq N_{n,i} \leq Mn$, $i = 1, \cdots$, $k+1$, for some fixed M. Let $N_n = \sum_{i=1}^{k+1} N_{n,i}$, and $p_{n,i} = \dfrac{N_{n,i}}{N_n}$ for $i = 1$, $\cdots$, $k+1$. Define Σ_n as:

$$\Sigma_n = \begin{bmatrix} p_{n,1}(1-p_{n,1}) & -p_{n,1}p_{n,2} & \cdots & -p_{n,1}p_{n,k} \\ -p_{n,2}p_{n,1} & p_{n,2}(1-p_{n,2}) & \cdots & -p_{n,2}p_{n,k} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{n,k}p_{n,1} & -p_{n,k}p_{n,2} & \cdots & p_{n,k}(1-p_{n,k}) \end{bmatrix}$$

With the convention that $N_{n,0} = 0$, for $i = 1, \cdots, k+1$, define:

$$Y_{n,i} = \frac{\sqrt{N_n}\mu}{\sigma}\left(\frac{\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j}{\sum_{j=1}^{N_n} X_j} - p_{n,i}\right)$$

Then:

$$\Sigma_n^{-\frac{1}{2}} \boldsymbol{Y}_n \implies N(\boldsymbol{0}, \boldsymbol{I}_k) \tag{3.7}$$

where $\boldsymbol{Y}_n = (Y_{n,1}, \cdots, Y_{n,k})^T$, and $\boldsymbol{I}_k$ is the $k \times k$ identity matrix.

In Theorems 5 and 6, we assume the amplification factors of all elements are identically distributed. It is possible to generalize the two theorems to allow the amplification factors for elements belonging to different categories to have different distributions. Let μ_i and σ_i^2 be the mean and the variance of the amplification factor for the i^{th} category respectively. Under the new condition, the relative frequency of the i^{th} category converges to $q_{n,i} = \frac{N_{n,i}\mu_i}{\sum_{j=1}^{k+1}(N_{n,j}\mu_j)}$. Define $Y'_{n,i} = \sqrt{N_n}\left(\frac{\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j}{\sum_{j=1}^{N_n} X_j} - q_{n,i}\right)$, then there is a matrix Σ'_n such that $\Sigma'^{-\frac{1}{2}}_n(Y_{n,1},\cdots,Y_{n,k})^T \implies N(\mathbf{0},\boldsymbol{I}_k)$.

Unfortunately, the matrix Σ'_n is much more complicated than Σ_n. For example, the first element of the first column of Σ'_n is:

$(1-q_{n,1})^2 p_{n,1}\sigma_1^2 + q_{n,1}^2 \sum_{i=2}^{k+1} p_{n,i}\sigma_i^2$

and the second element of the first column is:

$$-(1-q_{n,1})q_{n,2}p_{n,1}\sigma_1^2 - (1-q_{n,2})q_{n,1}p_{n,2}\sigma_2^2$$
$$+\ q_{n,1}q_{n,2}\sum_{i=3}^{k+1} p_{n,i}\sigma_i^2$$

The following corollaries are generalizations of corollaries 3, 4 and 5 respectively.

Corollary 6. *In Theorem 6, if $E[X_i^4] < \infty$ and $X_i \geq c > 0$, then the covariance matrix of $\Sigma_n^{-\frac{1}{2}}\boldsymbol{Y}_n^T$ converges to $\boldsymbol{I}_k$, where $\boldsymbol{Y}_n^T = (Y_{n,1},\cdots,Y_{n,k})^T$.*

Corollary 7. *If in Theorem 6 and Corollary 6, instead of requiring $n = N_{n,1} \leq N_{n,i} \leq Mn$, we require that $Ln \leq N_{n,i} \leq Mn$, where L is some positive real number, the conclusions still hold.*

Corollary 8. *If in Theorem 6, we assume that $\Sigma_n \to \Sigma$, then:*

$$(Y_{n,1}, \cdots, Y_{n,k})^T \implies N(\mathbf{0}, \Sigma) \tag{3.8}$$

3.3 Asymptotic distribution of the SAR sample

Theorems 5 and 6 give the asymptotic distribution of the relative frequencies in the intermediate sample conditional on the original sample. The asymptotic distributions for the relative frequencies in the original sample, and the relative frequencies in the final sample conditional on the intermediate sample, are straightforward: Both the relative frequencies in a multinomial sample and the relative frequencies in a multivariate hypergeometric sample converge weakly to multivariate normal. More precisely, let $\boldsymbol{X}$ be a k dimensional random vector following a multivariate hypergeometric distribution with parameters $(N; N_1, \cdots, N_k; n)$, where N is the population size, and n is the sample size. Let $n/N = \beta$, $N_i/N = p_i$ and $\boldsymbol{p} = (p_1, \cdots, p_k)^T$. Fixing $\boldsymbol{p}$ and β, as $n \to \infty$, $(\boldsymbol{X} - n\boldsymbol{p}) \implies N(\mathbf{0}, (1-\beta)n\Sigma_p)$, where Σ_p is the covariance matrix of a multinomial distribution with parameters $(1; p_1, \cdots, p_k)$. (For a general proof, see Hajek (1960).) We need to put these pieces together to get the marginal asymptotic distribution of the relative frequencies in the final sample. The basic idea is to show, under certain conditions, that conditional convergence implies marginal convergence. More precisely, consider two sequences of random variables X_i and Y_i, as well as two random variables X and Y. We say Y_i converges to Y conditional on X_i if 1), X_i converges weakly to X, and 2), there are versions of $P(Y_i \leq y|X_i = x)$ and a Borel set A such that $\mu_X(A) = 1$ and for each fixed $x \in A$, $P(Y_i \leq y|X_i = x) \to P(Y \leq y|X = x)$, where μ_X is the measure induced by X. The goal is to find a sufficient condition to guarantee $Y_i \Longrightarrow Y$.

To do so, we first introduce a new concept called the *dual distribution function* (ddf). The dual distribution functions are defined in a similar way as the distribution functions so that the dual distribution functions could share some properties, such as the uniform convergence, of the distribution functions.

Definition 2. *A nonnegative function G on $\mathbb{R}^k$ is called a dual distribution function if it satisfies the following conditions:*

- *G is continuous from below.*
- *G is decreasing.*
- *Let $\boldsymbol{x} = (x_1, \cdots, x_k)^k$, and $i \in \{1, \cdots, k\}$. If for some i, $x_i \to \infty$, then $G(\boldsymbol{x}) \to 0$. If $x_i \to -\infty$ for all i, then $G(\boldsymbol{x}) \to 1$.*

It is easy to check the following properties of a dual distribution function:

Proposition 1. *A dual distribution function G on $\mathbb{R}^k$ determines uniquely a probability measure μ such that*

$$\mu(\{\boldsymbol{x} : x_1 \geq y_1, \cdots, x_k \geq y_k\}) = G(\boldsymbol{y})$$

for any $\boldsymbol{y} = (y_1, \cdots, y_k)^T \in \mathbb{R}^k$.

Note that if F is the distribution function corresponding to a measure μ, then the dual distribution function G for μ in general is not equal to $1 - F$. More precisely, we have:

Proposition 2. *$G = 1 - F$ if and only if F is a continuous distribution function on $\mathbb{R}$.*

The following lemma is an extension of the well known theorem of the uniform convergence of the distribution functions on $\mathbb{R}$. (For example, see Theorem 7.6.2 of Ash and Doleans-Dade (2000).)

Lemma 5. *Consider a continuous distribution function F defined on $\mathbb{R}^k$. If there is a sequence of distribution functions $\{F_n\}$ converge weakly to F, then F_n converges to F uniformly.*

The following corollary shows the uniform convergence of the dual distribution functions on $\mathbb{R}^k$.

Corollary 9. *If G is a continuous dual distribution function, and a sequence of dual distribution functions G_n pointwise converge to G. Then G_n converges to G uniformly.*

Corollary 9 and the following lemma will be called the lemmas of conditional convergence. Together they give a sufficient condition for conditional convergence, but they can also be used independently.

Lemma 6. *Consider random variables $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}$, $\{\boldsymbol{Y}_n\}$ and $\boldsymbol{Y}$. Let μ_n and μ be the measures induced by $\boldsymbol{X}_n$ and $\boldsymbol{X}$ respectively. Suppose the following* *conditions are satisfied:*

1. $\boldsymbol{X}_n \Longrightarrow \boldsymbol{X}$*, and* X *has a continuous distribution function.*

2. *For any fixed* $\boldsymbol{y}$ *and for all* n*, there is a* μ_n *measurable function* $G_{n,\boldsymbol{y}} = P(\boldsymbol{Y}_n \leq \boldsymbol{y}|\boldsymbol{X}_n = \boldsymbol{x})$ *a.s.*$[\mu_n]$ *such that* $G_{n,\boldsymbol{y}} \rightarrow P(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ *uniformly, and* $P(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ *is continuous in* $\boldsymbol{x}$*.*

Then $\boldsymbol{Y}_n \Longrightarrow \boldsymbol{Y}$*.*

Now we can derive the asymptotic distribution for the relative frequency of a category in the final sample of an SAR scheme.

Theorem 7. *Consider the following SAR scheme: The original sample is a binomial sample with parameters* (m, p)*, where* m *is the sample size, and* p *the relative frequency of the elements belonging to the first category in the population. The mean and the variance of the amplification factor for the amplification process are* μ *and* σ^2 *respectively. Let* M_m *be the intermediate sample size. The final sample of size* N_m *is drawn without replacement from the intermediate sample, where* N_m *is a random variable such that, for some* $0 < \gamma < \mu$*,* $N_m = M_m$ *if* $M \leq \gamma m$ *and* $N_m = \gamma m$ *otherwise. (In this SAR scheme, if the intermediate sample size* M_m *is less than or equal to* γm*, then the whole intermediate sample is taken as the final sample. Otherwise, a final sample of size* γm *will be drawn without replacement from the intermediate sample.) Suppose* Z_m *is the count of elements belonging to the first category in the final sample. Then as* $m \rightarrow \infty$*,*

$$Z_m \implies N\Big(N_m p, N_m c_m p(1-p)\Big) \tag{3.9}$$

where $c_m = 1 - \dfrac{N_m}{m\mu} + \dfrac{N_m}{m}\left(1 + \dfrac{\sigma^2}{\mu^2}\right)$ *is called the normalizing factor of the SAR sample.*

We note that the assumption that the original sample is binomial is not essential to our proof. If it is hypergeometric with the ratio of the sample to population being δ_m, the above result still holds, with the exception that the normalizing factor now is changed to $c_m = 1 - \dfrac{N_m}{m\mu} + \dfrac{N_m}{m}\left(1 - \delta_m + \dfrac{\sigma^2}{\mu^2}\right)$

It seems reasonable to conjecture that Theorem 7 could be extended to the multivariate SAR samples, where the original samples are multinomial or multivariate hypergeometric, and the final samples are multivariate hypergeometric conditional on the intermediate samples. Let $\boldsymbol{X}_m$ and $\boldsymbol{Z}_m$ be the counts of first k categories of elements in the original sample and the final sample respectively. To show the asymptotic normality of $\boldsymbol{Z}_m$, we would only need to show the asymptotic normality of $\boldsymbol{Z}_m$ given $\boldsymbol{X}_m$. One approach would be to prove directly the asymptotic normality by the lemmas of conditional convergence. Another approach would be using the Cramer-Wold's theorem, i.e., showing the appropriate asymptotic normality of $\boldsymbol{u}^T\boldsymbol{Z}$ for an arbitrary $\boldsymbol{u} = (u_1, \cdots, u_k)^T$ conditional on $\boldsymbol{X}_m$. We tried both approaches, but were unable to get the desired result. Here we shall leave the multivariate version of Theorem 7 as a conjecture.

Although it is difficult to extend Theorem 7 to the multivariate SAR sample with the final sample obtained by drawing without replacement, we could show the asymptotic normality of the multivariate SAR sample if the final sample is drawn *with* replacement from the intermediate sample:

Theorem 8. *Consider a multinomial sample of size m drawn from a population of $k+1$ categories of elements with relative frequencies $p_1, \cdots, p_k$, and $1-\sum_{i=1}^k p_i$ respectively. Suppose each element of the multinomial sample is subject to i.i.d. amplification processes such that the mean and variance of the amplification factor are μ and σ^2 respectively. A sample of size n_m is then drawn with replacement from the intermediate sample. Suppose $\boldsymbol{Z}_m = (Z_{m1}, \cdots, Z_{mk})^T$ is the relative frequencies of the first k categories in the final sample. As $m, n_m \to \infty$, we have:*

$$\boldsymbol{Z}_m \sim N(\boldsymbol{p}, \frac{1}{n_m} c_m \Sigma_p) \tag{3.10}$$

where $\boldsymbol{p} = (p_1, \cdots, p_k)^T$, $c_m = 1 + \frac{n_m}{m}\left(1 + \frac{\sigma^2}{\mu^2}\right)$, and Σ_p is a matrix with $\sigma_{p,ii} = p_i(1-p_i)$ and $\sigma_{p,ij} = -p_i p_j$ for $i \neq j$.

Like Theorem 6, Theorem 8 can be generalized to allow different distributions of the amplification factors for elements belonging to different categories. Let μ_i be the mean of the amplification factor for the i^{th} category, and X_i the relative frequency of the i^{th} category in the original sample. We can get the asymptotic distribution of $\left(\frac{\mu_1 X_1}{\sum_{i=1}^{k+1} \mu_i X_i}, \cdots, \frac{\mu_k X_k}{\sum_{i=1}^{k+1} \mu_i X_i}\right)^T$ by the delta method. The generalized version of Theorem 6 is also needed. Otherwise, the proof of Theorem 8 remains largely unchanged. Of course, in this case, the covariance matrix for the asymptotic distribution of the relative frequencies will be extremely complicated.

The asymptotic results of Theorems 7 and 8 imply that the relative frequencies in a SAR sample are consistent estimators of the relative frequencies in the population. In particular, if the amplification factor is the simple type branch process described in Lemma 4, the relative frequencies in a SAR sample are indeed unbiased estimators of the relative frequencies in the population.

3.4 Discussion

In this chapter we have derived the asymptotic distribution of the relative frequencies in the final sample. It is interesting to note that there is a striking similarity between the asymptotic distribution of the relative frequencies of the categories in the final sample of an SAR sample, and the asymptotic distribution of the relative frequencies in a multinomial sample. Let Z be the relative frequencies of the categories in an SAR sample, according to Theorem 8, $(\boldsymbol{Z}_m - \boldsymbol{p}) \Longrightarrow N\left(\boldsymbol{0}, \dfrac{c_m}{n_m}\Sigma_p\right)$, where n_m is the size of the final sample, c_m the normalizing factor, and $\boldsymbol{p}$ the relative frequencies of the categories in the population. Thus, asymptotically Z_m behaves like the relative frequencies in a multinomial sample of size $\dfrac{n_m}{c_m}$ drawn from the same population. This observation leads immediately to tests of whether a category has the same relative frequency in two or more populations. In particular, we can extend the traditional χ^2 test or the G test for association in contingency tables in the following way:

Suppose we have drawn k SAR samples from k populations respectively. Let $n_1, \cdots, n_k$ and $c_1, \cdots, c_k$ be be the sizes and the normalizing factors for the k samples. Suppose $Z_1, \cdots, Z_k$ are the counts of elements belonging to a specific category in the k final samples. Under the null hypothesis that this category has the same relative frequency in all the k populations, the following two test statistics both have asymptotically a χ^2_{k-1} distribution:

$$X^2 = \sum_{i=1}^{k} \frac{\left(Z_i - n_i\hat{p}\right)^2}{c_i n_i \hat{p}}$$

$$G^2 = -2\sum_{i=1}^{k} \frac{Z_i}{c_i} \log \frac{Z_i}{n_i\hat{p}_i}$$

where $\hat{p} = \frac{\sum_{i=1}^{s} Z_i/c_i}{\sum_{j=1}^{k} n_j/c_j}$

The above tests could be easily extended to tests for whether a set of categories all have constant relative frequencies in a set of populations. In the cases where $Z_1, \cdots, Z_k$ are small, it is preferred to use bootstrap method to get the distribution of the test statistics, rather than relying on the asymptotic distribution.

The asymptotic results also show the possible consequences of treating SAR samples as multinomial samples. According to Theorems 7 and 8, asymptotically the covariance matrix of a SAR sample is exactly c times the covariance matrix of a multinomial sample of the same size from the same population, where c is the normalizing factor of the SAR sample.

Therefore, as long as c is close to 1 (c is always greater than 1), it might be just fine, and even convenient, to treat the SAR sample as if it were multinomial. For example, suppose we are given k SAR samples from k populations, and asked to test the null hypothesis that a certain category has the same relative frequency in all the k populations, we could simply use the traditional χ^2 or G^2 tests, and which are approximately χ^2_{k-1} distributed under the null hypothesis. However, when c is large, the multinomial model will significantly underestimate the variance of the SAR sample. In this case, the values of the traditional χ^2 and G^2 statistics are much higher than that of the test statistics modified for the SAR samples, which are indeed χ^2_{k-1} distributed under the null hypothesis. In such a situation, a traditional level α test based on the multinomial model will actually have a type I error rate much higher than α when applied to the SAR samples.

Another interesting consequence of the linear relation between the covariance matrix of the SAR sample and the multinomial sample is that, from an informational point of view, we can estimate the effective size of a SAR sample. Recall that a k dimensional multinomial sample of size n is a collection of n iid random vector conditional on the parameter $\boldsymbol{p} = (p_1, \cdots, p_k)^T$ with $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. The mean of each random vector is $\boldsymbol{p}$, and the covariance matrix is Σ_p with $\sigma_{p,ii} = p_i(1 - p_i)$ and $\sigma_{p,ij} = -p_i p_j$ for $i \neq j$. Let us call such a random vector a unit multinomial vector. Note that the mean and the covariance matrix of a unit multinomial random vector are highly constrained and dependent with each other. This is entirely different than a k dimensional multivariate normal random vector, whose mean could be any k dimensional real vector, and the covariance matrix could be any positive semi-definite matrix, where the mean and the covariance matrix are totally independent.

The strict dependence between the mean and the covariance of a multinomial distribution makes it possible to measure the information carried by a sample. In general, we can define the *effective sample size* of a categorical data set in the following way.

Definition 3. *Consider a sequence of k dimensional random vectors $\{\boldsymbol{X}_n\}$, each of which represents a $k \times 1$ contingency table $\boldsymbol{S}_n$. Suppose that $\boldsymbol{X}_n$ converges weakly:*

$$\sqrt{n}\left(\frac{\boldsymbol{X}_n}{n} - \boldsymbol{p}\right) \implies N\left(\mathbf{0}, c\Sigma_p\right)$$

Then we say that the (asymptotically) effective sample size of the sample $\boldsymbol{S}_n$ is $\frac{n}{c}$.

This definition does not apply to the continuous data, because nothing can prevent us from scaling those data. However, for the contingency tables, there is an unambiguous natural unit: count. Using the above definition, the effective sample size of a multinomial sample of size n is still n. For a hypergeometric sample with parameters (N, M, n), the effective sample size is $n\frac{N}{N-M}$. If M is close to N, the effective sample size of a hypergeometric sample of size n is close to n. On the other hand, if $N \gg M$, the effective sample size will be much larger.

The effective sample size measures the amount of information about the relative frequencies of each cell/category of a contingency table carried by a sample compared to a unit multinomial vector, and it is closely related to the Fisher Information. For example, the Fisher information for a binomial sample with parameters (n, p) is $\frac{n}{p(1-p)}$, which is proportional to the effective sample size n.

3.5 Appendix: Proofs

Proofs for section 2

Lemma 4. *Let $\{X_t\}$ and $\{Y_t\}$ be two independent branch processes with the following properties:*

1. *$X_{t+1} = X_t + U_t$, where U_t follows a binomial distribution with parameters (X_t, λ), for $0 < \lambda < 1$.*
2. *$Y_{t+1} = Y_t + V_t$, where V_t follows a binomial distribution with parameters (Y_t, λ).*

Let $P_{t+1} = \frac{X_{t+1}}{X_{t+1} + Y_{t+1}}$ and $P_t = \frac{X_t}{X_t + Y_t}$, then:

$$E[P_{t+1}|P_t] = P_t \tag{3.11}$$

Proof: Without loss of generality, let $t = 0$. The joint distribution of (X_1, Y_1) given X_0 and Y_0 is:

$$\begin{aligned} & P(X_1 = x, Y_1 = y | X_0, Y_0) \\ = \quad & \binom{X_0}{x - X_0} \lambda^{x - X_0} (1-\lambda)^{2X_0 - x} \binom{Y_0}{y - Y_0} \lambda^{y - Y_0} (1-\lambda)^{2Y_0 - y} \end{aligned}$$

where $X_0 \leq x \leq 2X_0$, and $Y_0 \leq y \leq 2Y_0$.

Let $u = x - X_0$, $v = y - Y_0$, and $P_1 = \frac{X_1}{X_1 + Y_1}$, the conditional mean of P_1 given X_0 and Y_0 is:

$$\begin{aligned} & \mathrm{E}[P_1 | X_0, Y_0] \\ = \quad & \sum_{u=0}^{X_0} \sum_{v=0}^{Y_0} \frac{u + X_0}{u + v + X_0 + Y_0} \binom{X_0}{u} \binom{Y_0}{v} \lambda^{u+v} (1-\lambda)^{X_0 - u + Y_0 - v} \end{aligned}$$

Let $c = u + v$, with the convention that $\binom{k_1}{k_2} = 0$ if $k_1 < k_2$, the above formula can be written as:

$$\begin{aligned} & \mathrm{E}[P_1 | X_0, Y_0] \\ = \quad & \sum_{c=0}^{X_0+Y_0} \sum_{u=0}^{c} \frac{u + X_0}{c + X_0 + Y_0} \binom{X_0}{u} \binom{Y_0}{c-u} \lambda^{c} (1-\lambda)^{X_0 + Y_0 - c} \\ = \quad & \sum_{c=0}^{X_0+Y_0} \frac{\lambda^{c} (1-\lambda)^{X_0 + Y_0 - c}}{c + X_0 + Y_0} \sum_{u=0}^{c} (u + X_0) \binom{X_0}{u} \binom{Y_0}{c-u} \end{aligned}$$

Finally, by the identity $\binom{n}{x} = \sum_{i=0}^{x} \binom{m}{i} \binom{n-m}{x-i}$, and the fact that $\sum_{i=0}^{x} i \frac{\binom{m}{i}\binom{n-m}{x-i}}{\binom{n}{x}} = \frac{m}{n} x$, we have:

$$
\begin{aligned}
&\mathrm{E}[P_1|X_0,Y_0]\\
&= \sum_{c=0}^{X_0+Y_0} \frac{1}{c+X_0+Y_0}\left(X_0 + c\frac{X_0}{X_0+Y_0}\right)\lambda^c(1-\lambda)^{X_0+Y_0-c}\\
&= \frac{X_0}{X_0+Y_0} = P_0
\end{aligned}
$$

Given that $\sigma(P_0) \subset \sigma(X_0, Y_0)$, it then follows that:

$$\mathrm{E}[P_1|P_0] = \mathrm{E}[\mathrm{E}[P_1|X_0,Y_0]|P_0] = P_0$$

□

Theorem 5. *Given a sequence of i.i.d. nonnegative random variables X_1, $\cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. Let r_n be a sequence of positive integers such that $n \le r_n \le Mn$ for some fixed M. Then:*

$$\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) \implies N\left(0, \frac{\sigma^2}{\mu^2}\right) \tag{3.12}$$

where $p_n = \dfrac{n}{n+r_n}$

Proof: Because $n \le r_n \le Mn$, for any n, there is a positive integer m_n such that $m_n n \le r_n \le [m_n + 1]n$, where $1 \le m_n < M$. Moreover, we can find $n+1$ integers:

$$0 = q_{n,0} < q_{n,1} < q_{n,2} \cdots < q_{n,n} = r_n$$

such that $q_{n,i+1} - q_{n,i}$ is either m_n or $m_n + 1$. Create a triangular array of random variables $Y_{i,j}$ such that the n^{th} row of the array has n elements $Y_{n,1}$, $\cdots$, $Y_{n,n}$, where:

$$\begin{aligned} Y_{n,i} &= (1-p_n)X_i - p_n\Bigg(\sum_{j=q_{n,i-1}+n+1}^{q_{n,i}+n} X_j\Bigg) \\ &\quad -[(1-p_n) - p_n(q_{n,i} - q_{n,i-1})]\mu \end{aligned}$$

It follows immediately that for each n, $Y_{n,1}$, $\cdots$, $Y_{n,n}$ are independent, and $\mathrm{E}[Y_{n,i}] = 0$. Let

$$\begin{aligned} S_n &= \sum_{i=1}^{n} Y_{n,i} = \sum_{i=1}^{n} X_i - p_n\Bigg(\sum_{j=1}^{n+r_n} X_j\Bigg) \\ s_n^2 &= \sum_{i=1}^{n} \mathrm{Var}(Y_{n,i}) = \mathrm{Var}\Bigg(\sum_{i=1}^{n} Y_{n,i}\Bigg) = n(1-p_n)\sigma^2 \end{aligned}$$

Let $Z_{n,i} = X_i + \sum_{j=q_{n,i-1}+n+1}^{q_{n,i-1}+n+M+1} X_j + 2\mu$. It is easy to check that $|Y_{n,i}| \leq Z_{n,i} = |Z_{n,i}|$, (because $X_i \geq 0$ and $p_n(q_{n,i} - q_{n,i-1}) \leq 1$). Also, the distribution of $Z_{n,i}$ is independent of n, and is the same as that of $\sum_{j=1}^{M+2} X_j + 2\mu$. Therefore, $Z_{n,i}^2$ is integrable, hence for any $\epsilon > 0$, as $n \to \infty$,

$$\int_{|Z_{n,i}|>\epsilon\sigma\sqrt{n(1-p_n)}} Z_{n,i}^2 dP \to 0$$

(note that $p_n \leq 0.5$). It then follows that:

$$\sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Y_{n,i}|>s_n\epsilon} Y_{n,i}^2\, dP \;\leq\; \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Z_{n,i}|>s_n\epsilon} Y_{n,i}^2\, dP$$

$$\begin{aligned}
&\leq \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Z_{n,i}|>s_n\epsilon} Z_{n,i}^2 \, dP \;=\; \frac{n}{n(1-p_n)\sigma^2} \int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 \, dP \\
&\leq \frac{2}{\sigma^2} \int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 dP
\end{aligned}$$

where $\displaystyle\int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 dP \to 0$ as $n \to \infty$

By the Central Limit Theorem,

$$\frac{S_n}{s_n} \Longrightarrow N(0,1)$$

On the other hand, by the Strong Law of the Large Number,

$$\frac{\sum_{i=1}^{n+r_n} X_i}{n+r_n} \to \mu \text{ w.p.1}$$

Consequently,

$$\begin{aligned}
&\frac{\mu(n+r_n)^{\frac{3}{2}}}{\sigma\sqrt{nr_n}} \left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n \right) \\
= \;& \frac{S_n}{s_n} \frac{\mu}{\dfrac{1}{n+r_n} \displaystyle\sum_{i=1}^{n+r_n} X_i} \Longrightarrow N(0,1)
\end{aligned}$$

□

Corollary 3. *Given the same condition as in Theorem 5, if $E[X_i^4] < \infty$, then:*

$$E\left[\frac{(n+r_n)^3}{nr_n} \left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n \right)^2 \right] \to \frac{\sigma^2}{\mu^2} \tag{3.13}$$

Proof: From Theorem 5, we have:

$$\frac{(n+r_n)^3}{nr_n}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2 \implies \frac{\sigma^2}{\mu^2}\chi_1^2$$

It suffices to show $\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$. Actually, we shall prove the case for $\epsilon = 2$.

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] \\
= \ & \frac{1}{[n+r_n]^2}\mathrm{E}\left[\left((1-p_n)\sum_{i=1}^{n} X_i - p_n \sum_{j=n+1}^{n+r_n} X_j\right)^4\right] \\
= \ & \frac{1}{[n+r_n]^2}\mathrm{E}\left[\left((1-p_n)\sum_{i=1}^{n}(X_i-\mu) - p_n \sum_{j=n+1}^{n+r_n}(X_j-\mu)\right)^4\right] \\
= \ & \frac{1}{[n+r_n]^2}\left\{\sum_{i=1}^{n}(1-p_n)^4\mathrm{E}[(X_i-\mu)^4]\right. \\
& + \sum_{j=n+1}^{n+r_n} p_n^4\mathrm{E}[(X_j-\mu)^4] \\
& + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1-p_n)^4\mathrm{E}[(X_i-\mu)^2]\mathrm{E}[(X_j-\mu)^2] \\
& + 2\sum_{i=n+1}^{n+r_n-1}\sum_{j=i+1}^{n+r_n} p_n^4\mathrm{E}[(X_i-\mu)^2]\mathrm{E}[(X_j-\mu)^2] \\
& \left. + \sum_{i=1}^{n}\sum_{j=n+1}^{n+r_n} p_n^2(1-p_n)^2\mathrm{E}[(X_i-\mu)^2]\mathrm{E}[(X_j-\mu)^2]\right\} \\
= \ & \frac{1}{[n+r_n]^2}\{n(1-p_n)^4\mathrm{E}[(X_1-\mu)^4] + r_n p_n^4\mathrm{E}[(X_1-\mu)^4] \\
& + n(n-1)(1-p_n)^4\sigma^4 + [n+r_n][n+r_n-1]p_n^4\sigma^4 \\
& + n[n+r_n]p_n^2(1-p_n)^2\sigma^4\}
\end{aligned}$$

$$< \quad \mathrm{E}[(X_1-\mu_1)^4]+\sigma^4$$

Given that $\mathrm{E}[(X_1)^4] < \infty$,

$$\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] \leq \mathrm{E}[(X_1-\mu_1)^4]+\sigma^4 < \infty$$

Note that $X_i \geq c > 0$, hence $\dfrac{\sum_{i=1}^{n+r_n} X_i}{n+r_n} \geq c$. Also we have $(n+r_n)^2 \leq 2(M+1)nr_n$. It then follows that:

$$\begin{aligned} &\sup_n \mathrm{E}\left[\left(\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left|\frac{S_n}{\sum_{i=1}^{n+r_n} X_i}\right|\right)^4\right] \\ \leq &\quad \frac{4(M+1)^2}{c^4}\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] < \infty \end{aligned}$$

Therefore $\dfrac{(n+r_n)^3}{nr_n}\left(\dfrac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2$ is uniformly integrable, hence its mean converges to $\dfrac{\sigma^2}{\mu^2}$.

□

Corollary 4. *If in Theorem 5 and Corollary 3, instead of requiring $n \leq r_n \leq Mn$, we require that $Ln \leq r_n \leq Mn$, where L is some positive real number, the conclusions still hold.*

Proof: First we note that if $r_n < n$, then:

$$\begin{aligned} &\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) \\ = &\quad -\frac{(r_n+n)^{\frac{3}{2}}}{\sqrt{r_n n}}\left(\frac{\sum_{i=n+1}^{n+r_n} X_i}{\sum_{j=1}^{n+r_n} X_j} - \frac{r_n}{r_n+n}\right) \end{aligned}$$

We also note that if $X \sim N(0,1)$, then $-X \sim N(0,1)$.

□

Theorem 6. *Theorem 5 can be generalized in the following way:*

Given a sequence of independent nonnegative random variables X_1, $\cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. For $n = 1$, $\cdots$, let $N_{n,1}$, $\cdots$, $N_{n,k+1}$ be positive integers such that $n = N_{n,1} \leq N_{n,i} \leq Mn$, $i = 1$, $\cdots$, $k+1$, for some fixed M. Let $N_n = \sum_{i=1}^{k+1} N_{n,i}$, and $p_{n,i} = \dfrac{N_{n,i}}{N_n}$ for $i = 1$, $\cdots$, $k+1$. Define Σ_n as:

$$\Sigma_n = \begin{bmatrix} p_{n,1}(1-p_{n,1}) & -p_{n,1}p_{n,2} & \cdots & -p_{n,1}p_{n,k} \\ -p_{n,2}p_{n,1} & p_{n,2}(1-p_{n,2}) & \cdots & -p_{n,2}p_{n,k} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{n,k}p_{n,1} & -p_{n,k}p_{n,2} & \cdots & p_{n,k}(1-p_{n,k}) \end{bmatrix}$$

With the convention that $N_{n,0} = 0$, for $i = 1$, $\cdots$, $k+1$, define:

$$Y_{n,i} = \frac{\sqrt{N_n}\mu}{\sigma}\left(\frac{\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j}{\sum_{j=1}^{N_n} X_j} - p_{n,i}\right)$$

Then:

$$\Sigma_n^{-\frac{1}{2}}\boldsymbol{Y}_n \implies N(\boldsymbol{0}, \boldsymbol{I}_k) \tag{3.14}$$

where $\boldsymbol{Y}_n = (Y_{n,1}, \cdots, Y_{n,k})^T$, and $\boldsymbol{I}_k$ is the $k \times k$ identity matrix.

Proof: Σ_n is the covariance matrix for a k-dimensional vector $(V_1, \cdots, V_k)$ where $(V_1, \cdots, V_k, 1 - \sum_{i=1}^{k} V_i)$ has a multinomial distribution with parameters $\left(1; \dfrac{N_{n,1}}{N_n}, \cdots, \dfrac{N_{n,k+1}}{N_n}\right)$. Thus, Σ_n is positive definite, and $\Sigma_n^{-\frac{1}{2}}$ exists.

Now define random vectors $\boldsymbol{Z}_n = (Z_{n,1}, \cdots Z_{n,k})$ by:

$$\begin{aligned} Z_{n,i} &= Y_{n,i}\frac{\sum_{j=1}^{N_n} X_j}{N_n\mu} \\ &= \frac{1}{\sigma\sqrt{N_n}}\left(\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j - p_{n,i}\sum_{j=1}^{N_n} X_j\right) \end{aligned}$$

It is easy to check that Σ_n is the covariance matrix of the random vector $(Z_{n,1},\cdots,Z_{n,k})^T$. Now let $\boldsymbol{u}=(u_1,\cdots,u_k)^T$ be any k-dimensional vector. Using the similar method used in the proof of Theorem 1, we can decompose $\sqrt{N_n}\boldsymbol{u}^T\Sigma_n^{-\frac{1}{2}}\boldsymbol{Z}_n$ into the sum of n independent random variables U_1, $\cdots$, U_n with zero mean such that the Lindeberg's condition is satisfied. This is possible because the absolute value of each entry of $\Sigma_n^{-\frac{1}{2}}$ is bounded from above by 1. The basic idea is:

First, write $\sqrt{N_n}\boldsymbol{u}^T\Sigma_n^{-\frac{1}{2}}\boldsymbol{Z}_n$ as:

$$\sqrt{N_n}\boldsymbol{u}^T\Sigma_n^{-\frac{1}{2}}\boldsymbol{Z}_n = \sum_{j=1}^{N_n} c_jX_j$$

Given that entries of $\Sigma_n^{-\frac{1}{2}}$ are bounded between -1 and 1, it can be shown that $\sigma|c_j|\leq 2\sum_{i=1}^k|u_i|$. Suppose $r_nn\leq N_n\leq(r_n+1)n$. Clearly, $r_n\leq M$. Now we can find a sequence of $n+1$ integers

$$0=q_{n,0}<q_{n,1}<\cdots<q_{n,n}=N_n$$

such that $r_n\leq q_{n,i+1}-q_{n,i}\leq r_n+1$. Define U_i as:

$$U_i=\sum_{j=q_{n,i-1}+1}^{q_{n,i}} c_jX_j-\sum_{j=q_{n,i-1}+1}^{q_{n,i}} c_j$$

Then it is easy to check that the Lindeberg's condition is satisfied. That is, let $s_n^2 = \mathrm{Var}(\sum_{j=1}^n U_j)$, as $n \to \infty$,

$$\sum_{i=1}^n \frac{1}{s_n^2} \int_{|U_i|>s_n\epsilon} U_i^2 dP \to 0$$

It then follows:

$$\frac{\sqrt{N_n}\boldsymbol{u}^T \Sigma_n^{-\frac{1}{2}} \boldsymbol{Z}_n}{s_n} \implies N(0,1)$$

Now because the covariance matrix for $\Sigma_n^{-\frac{1}{2}}\boldsymbol{Z}_n$ is the identity matrix $\boldsymbol{I}_k$,

$$s_n^2 = \mathrm{Var}\left(\sqrt{N_n}\boldsymbol{u}^T \Sigma_n^{-\frac{1}{2}} \boldsymbol{Z}_n\right) = N_n \sum_{i=1}^k u_i^2$$

Therefore,

$$\boldsymbol{u}^T \Sigma_n^{-\frac{1}{2}} \boldsymbol{Z}_n \implies N\left(0, \sum_{i=1}^k u_i^2\right)$$

which is the same as the distribution of $\boldsymbol{u}^T\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_k)$.

Thus,

$$\Sigma_n^{-\frac{1}{2}} \boldsymbol{Z}_n \implies N(\boldsymbol{0}, \boldsymbol{I}_k)$$

Because $\dfrac{\sum_{j=1}^{N_n} X_j}{N_n\mu} \to 1$ w.p.1., it then follows that, for any k-dimensional vector $\boldsymbol{u} = (u_1, \cdots, u_k)^T$,

$$\boldsymbol{u}^T \Sigma_n^{-\frac{1}{2}} \boldsymbol{Y}_n = \boldsymbol{u}^T \frac{\sum_{j=1}^{N_n} X_j}{N_n\mu} \Sigma_n^{-\frac{1}{2}} \boldsymbol{Z}_n \implies \boldsymbol{u}^T N(\boldsymbol{0}, \boldsymbol{I}_k)$$

□

Corollary 6. *In Theorem 6, if* $E[X_i^4] < \infty$ *and* $X_i \geq c > 0$*, then the covariance matrix of* $\Sigma_n^{-\frac{1}{2}} \boldsymbol{Y}_n^T$ *converges to* $\boldsymbol{I}_k$*, where* $\boldsymbol{Y}_n^T = (Y_{n,1}, \cdots, Y_{n,k})^T$*.*

Proof: From Corollary 3, $\mathrm{E}[Y_{n,i}^4] < \infty$ for $i = 1, \cdots, k+1$. Therefore, for any $\boldsymbol{u} = (u_1, \cdots, u_k)^T$, $\mathrm{E}[(\boldsymbol{u}^T \boldsymbol{Y}_n)^4] < \infty$. Thus the variance of $\boldsymbol{u}^T \boldsymbol{Y}_n$ converges to the variance of the distribution μ if $\boldsymbol{u}^T \boldsymbol{Y}_n \implies \mu$.

□

Corollary 7. *If in Theorem 6 and Corollary 6, instead of requiring* $n = N_{n,1} \leq N_{n,i} \leq Mn$*, we require that* $Ln \leq N_{n,i} \leq Mn$*, where* L *is some positive real number, the conclusions still hold.*

Proof: The proofs in Theorem 6 and Corollary 6 depend only on the assumption that there is a fixed number M such that $M \min(N_{n,1}, \cdots, N_{n,k+1}) \geq \min(N_{n,1}, \cdots, N_{n,k+1})$.

□

Proofs for section 3

Lemma 5. *Consider a continuous distribution function* F *defined on* $\mathbb{R}^k$*. If there is a sequence of distribution functions* $\{F_n\}$ *converge weakly to* F*, then* F_n *converges to* F *uniformly.*

Proof: Let the measures corresponding to F and F_n be μ and μ_n respectively. Define a compact set C_a as $C_a = \{(x_1, \cdots, x_k)^T : |x_1| \leq a, \cdots, |x_k| \leq a\}$. Note that F is uniformly continuous on C_a. For any $\epsilon > 0$, choose an a such that $\mu(C_a^c) < \epsilon$. Then we can find a finite number of compact sets B_1, $\cdots$,

B_m such that $\bigcup_{i=1}^m B_i = C_a$ and that $\max_{\boldsymbol{x},\, \boldsymbol{y} \in B_i}(|F(\boldsymbol{x}) - F(\boldsymbol{y})|) \leq \epsilon$ for all $1 \leq i \leq m$.

Let $\boldsymbol{x}_{i,max}$ and $\boldsymbol{x}_{i,min}$ be the maximum and the minimum points in B_i. Because $F_n \implies F$, we can find an $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, and for all $1 \leq i \leq m$,

$$\begin{aligned} |F_n(\boldsymbol{x}_{i,max}) - F(\boldsymbol{x}_{i,max})|) &\leq \epsilon \\ |F_n(\boldsymbol{x}_{i,min}) - F(\boldsymbol{x}_{i,min})|) &\leq \epsilon \\ |\mu_n(C_a) - \mu(C_a)| &\leq \epsilon \end{aligned}$$

It then follows that, for all $n \geq N(\epsilon)$, $|F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \leq 3\epsilon$ for any $\boldsymbol{x} \in C_a$, and $\mu_n(C_a^c) \leq 2\epsilon$.

For any $\boldsymbol{x} = (x_1, \cdots, x_k)^T \in \mathbb{R}^k$, define a set

$$L_{\boldsymbol{x}} = \{\boldsymbol{y} = (y_1, \cdots, y_k) : y_1 \leq x_1, \cdots, y_k \leq x_k\}$$

Note that for any $\boldsymbol{x}$, $\mu_n(L_{\boldsymbol{x}}) = F_n(\boldsymbol{x})$, and $\mu(L_{\boldsymbol{x}}) = F(\boldsymbol{x})$. Let $\boldsymbol{a} = (a, \cdots, a)^T$. Now let us consider the following two situations:

- Suppose $C_a \cap L_{\boldsymbol{x}} = \emptyset$, then we have:

  $|F_n(\boldsymbol{x}) - F(\boldsymbol{x})| = |\mu_n(L_{\boldsymbol{x}}) - \mu(L_{\boldsymbol{x}})| \leq 2\epsilon.$

- Suppose $C_a \cap L_{\boldsymbol{x}} = C_{\boldsymbol{a},\boldsymbol{x}} \neq \emptyset$. Clearly, $C_{\boldsymbol{a},\boldsymbol{x}}$ is compact, hence has a maximum point $\boldsymbol{x}_{a,max}$. It is easy to see that $L_{\boldsymbol{x}_{a,max}} \subset L_{\boldsymbol{x}}$, and $(L_{\boldsymbol{x}} \setminus L_{\boldsymbol{x}_{a,max}}) \cap C_a = \emptyset$. Now we have:

$$
\begin{aligned}
&|F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \\
&\quad = \quad |[\mu_n(L_{\boldsymbol{x}} \backslash L_{\boldsymbol{x}_{a,max}}) + F_n(\boldsymbol{x}_{a,max})] - [\mu(L_{\boldsymbol{x}} \backslash L_{\boldsymbol{x}_{a,max}}) + F(\boldsymbol{x}_{a,max})]| \\
&\quad \leq \quad |\mu_n(L_{\boldsymbol{x}} \backslash L_{\boldsymbol{x}_{a,max}}) - \mu(L_{\boldsymbol{x}} \backslash L_{\boldsymbol{x}_{a,max}})| + |F_n(\boldsymbol{x}_{a,max}) - F(\boldsymbol{x}_{a,max})| \\
&\quad \leq \quad 2\epsilon + 3\epsilon = 5\epsilon
\end{aligned}
$$

□

Corollary 9. If G is a continuous dual distribution function, and a sequence of dual distribution functions G_n pointwise converge to G. Then G_n converges to G uniformly.

Proof: Similar as the proof for Lemma 5. Note that $G_n \to G$ pointwisely implies that $\mu_n \Longrightarrow \mu$, where μ_n and μ are the measures determined by G_n and G.

□

Lemma 6. *Consider random variables* $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}$, $\{\boldsymbol{Y}_n\}$ *and* $\boldsymbol{Y}$. *Let* μ_n *and* μ *be the measures induced by* $\boldsymbol{X}_n$ *and* $\boldsymbol{X}$ *respectively. Suppose the following conditions are satisfied:*

1. $\boldsymbol{X}_n \Longrightarrow \boldsymbol{X}$, *and* X *has a continuous distribution function.*

2. *For any fixed* $\boldsymbol{y}$ *and for all* n, *there is a* μ_n *measurable function* $G_{n,\boldsymbol{y}} = P(\boldsymbol{Y}_n \leq \boldsymbol{y}|\boldsymbol{X}_n = \boldsymbol{x})$ *a.s.*$[\mu_n]$ *such that* $G_{n,\boldsymbol{y}} \to P(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ *uniformly, and* $P(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ *is continuous in* $\boldsymbol{x}$.

Then $\boldsymbol{Y}_n \Longrightarrow \boldsymbol{Y}$.

Proof: It suffices to show that for all $\boldsymbol{y}$,

$$\int P(\boldsymbol{Y}_n \leq \boldsymbol{y} | \boldsymbol{X}_n = \boldsymbol{x})\, d\mu_n(\boldsymbol{x}) \rightarrow \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu(\boldsymbol{x})$$

or equivalently,

$$\int G_{n,\boldsymbol{y}}\, d\mu_n(\boldsymbol{x}) \rightarrow \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu(\boldsymbol{x})$$

Fixed $\boldsymbol{y}$, because $G_{n,\boldsymbol{y}}$ converges to $P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})$ uniformly, for any $\epsilon > 0$, there is an $N(\epsilon)$ such that for any $n \geq N(\epsilon)$,

$$\sup_{\boldsymbol{x}} |G_{n,\boldsymbol{y}} - P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})| \leq \epsilon$$

Because $\mu_n \rightarrow \mu$ and $P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})$ is bounded and continuous, we can choose M_ϵ such that for all $n \geq M_\epsilon$,

$$\left| \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu_n(\boldsymbol{x}) - \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu(\boldsymbol{x}) \right| < \epsilon$$

Therefore, for all $n > \max(N_\epsilon, M_\epsilon)$,

$$\begin{aligned} & \left| \int G_{n,\boldsymbol{y}}\, d\mu_n(\boldsymbol{x}) - \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu(\boldsymbol{x}) \right| \\ \leq & \quad \left| \int G_{n,\boldsymbol{y}}\, d\mu_n(\boldsymbol{x}) - \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu_n(\boldsymbol{x}) \right| \\ & \quad + \left| \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu_n(\boldsymbol{x}) - \int P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})\, d\mu(\boldsymbol{x}) \right| \end{aligned}$$

$$\leq \int |G_{n,\boldsymbol{y}} - P(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})| \, d\mu_n(\boldsymbol{x}) + \epsilon \leq 2\epsilon$$

□

Theorem 7. *Consider the following SAR scheme: The original sample is a binomial sample with parameters (m, p), where m is the sample size, and p the relative frequency of the elements belonging to the first category in the population. The mean and the variance of the amplification factor for the amplification process are μ and σ^2 respectively. Let M_m be the intermediate sample size. The final sample of size N_m is drawn without replacement from the intermediate sample, where N_m is a random variable such that, for some $0 < \gamma < \mu$, $N_m = M_m$ if $M \leq \gamma m$ and $N_m = \gamma m$ otherwise. (In this SAR scheme, if the intermediate sample size M_m is less than or equal to γm, then the whole intermediate sample is taken as the final sample. Otherwise, a final sample of size γm will be drawn without replacement from the intermediate sample.) Suppose Z_m is the count of elements belonging to the first category in the final sample. Then as $m \to \infty$,*

$$Z_m \implies N\Big(N_m p, N_m c_m p(1-p)\Big) \tag{3.15}$$

where $c_m = 1 - \dfrac{N_m}{m\mu} + \dfrac{N_m}{m}\left(1 + \dfrac{\sigma^2}{\mu^2}\right)$ is called the normalizing factor of the SAR sample.

Proof: Let X_m be the count of elements belonging to the first category in the original sample, and Y_m and R_m be the counts of elements belonging to the first and the second categories in the intermediate sample respectively.

By the central limit theorem, as $m \to \infty$:

$$U_m = \frac{1}{\sqrt{m}}(X_m - mp) \implies N(0, p(1-p))$$

Conditional on $\frac{X_m}{m} = x$:

$$\begin{aligned} V_m &= \frac{1}{\sqrt{m}}(Y_m - m\mu x) \implies N(0, \sigma^2 x) \\ W_m &= \frac{1}{\sqrt{m}}(R_m - m\mu(1-x)) \implies N(0, \sigma^2(1-x)) \end{aligned}$$

Note that Y_m and R_m are independent conditional on X_m, hence conditional on $\frac{X_m}{m} = x$:

$$\frac{1}{\sqrt{m}}[(1-x)(Y_m - \mu m x) - x(R_m - \mu m(1-x))] \implies N(0, \sigma^2(1-x)x)$$

Conditional on $\frac{Y_m}{m} = y$ and $\frac{R_m}{m} = r$:

$$\frac{1}{\sqrt{N_m}}\left(Z_m - N_m \frac{y}{y+r}\right) \Longrightarrow N\left(0, \left(1 - \frac{N_m}{m(y+r)}\right)\frac{yr}{(y+r)^2}\right)$$

Now we are going to prove the asymptotic normality of Z_m in four steps.

1. Conditional on $\frac{X_m}{m} = x$:

$$\frac{1}{\sqrt{N_m}}\left(Z_m - \frac{N_m}{m} X_m\right)$$

$$\begin{aligned}
&= \frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{Y_m}{Y_m+R_m}\right) + \frac{1}{\sqrt{N_m}}\left(N_m\frac{Y_m}{Y_m+R_m} - \frac{N_m}{m}X_m\right)\\
&= \frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{Y_m}{Y_m+R_m}\right)\\
&\quad + \frac{\sqrt{N_m}}{Y_m+R_m}[(1-x)(Y_m - xm\mu) - x(R_m-(1-x)m\mu)]
\end{aligned}$$

Conditional on $\frac{X_m}{m} = x$, $V_m = \frac{1}{\sqrt{m}}(Y_m - m\mu x) = v$, and $W_m = \frac{1}{\sqrt{m}}(R_m - m\mu(1-x)) = w$, (here x, v, and w are constant, hence independent of m), as $m \to \infty$,

$$\begin{aligned}
N_m &\to \gamma m\\
\frac{1}{\sqrt{\gamma m}}(N_m - \gamma m) &\to 0\\
\frac{\sqrt{N_m m}}{Y_m+R_m} = \frac{\sqrt{N_m m}}{m\mu + \sqrt{m}(v+w)} &\to \frac{\sqrt{\gamma}}{\mu}\\
\frac{N_m}{Y_m+R_m} = \frac{N_m}{m\mu + \sqrt{m}(v+w)} &\to \frac{\gamma}{\mu}\\
\frac{Y_m R_m}{(Y_m+R_m)^2} &\to x(1-x)
\end{aligned}$$

Thus, conditional on $\frac{X_m}{m} = x$, $V_m = v$, and $W_m = w$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \implies N\left(\frac{\sqrt{\gamma}}{\mu}[(1-x)v - xw], \left(1-\frac{\gamma}{\mu}\right)x(1-x)\right)$$

2. Next we show that conditional on $\frac{X_m}{m} = x$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \implies N\left(0, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2}\right)x(1-x)\right)$$

First we show that given $\dfrac{X_m}{m} = x$, the conditional distribution of $\dfrac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$ given $V_m = v$ and $W_m = w$ is a dual distribution function in $(v, -w)$. This is equivalent to showing that the conditional distribution of Z_m given Y_m and $-R_m$ is a dual distribution function in Y_m and $-R_m$. (The equivalence holds because $\dfrac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$, V_m, and W_m are linear in Z_m, Y_m, and R_m respectively, and the slopes of the linear transformations are all positive.)

- First we show $P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$ is decreasing in $(y_m, -r_m)$. Because the minimal change in y_m or r_m is 1, it suffices to show that, for any z,

$$\begin{aligned}
&P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) \\
&\quad \geq \quad P(Z_m \leq z | Y_m = y_m + 1, -R_m = -r_m) \\
&P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) \\
&\quad \geq \quad P(Z_m \leq z | Y_m = y_m, -R_m = -r_m + 1)
\end{aligned}$$

Note that conditional on $Y_m = y_m$ and $R_m = r_m$, Z_m has a hypergeometric distribution with parameters $(y_m + r_m, y_m, N_m)$, where $y_m + r_m$ is the population size, y_m the number of elements in population belonging to the first category, and N_m the sample size. Let $h_{N,M,n}$ be the distribution function for a hypergeometric function with parameters (N, M, n), we need to show, for any

(N, M, n), and any z:

$$\begin{aligned} h_{N,M,n}(z) &\geq h_{N+1,M+1,n} \\ h_{N,M,n}(z) &\geq h_{N-1,M,n} \end{aligned}$$

– Let $F_1 = h_{N,M,n}(z)$, $F_2 = h_{N+1,M+1,n}$, and f_1 and f_2 be the probability mass functions corresponding to F_1 and F_2 respectively:

$$\begin{aligned} F_1(z) &= \sum_{x=0}^{z} f_1(z) = \sum_{x=0}^{z} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \\ F_2(z) &= \sum_{x=0}^{z} f_2(z) = \sum_{x=0}^{z} \frac{\binom{M+1}{x}\binom{N-M}{n-x}}{\binom{N+1}{n}} \end{aligned}$$

Solve the inequality $f_1(x) > f_2(x)$ for $0 \leq x \leq \min(M, n)$, ($f_1(x) = 0$ for $x > \min(M, n)$, and $f_2(x) = 0$ for $x > \min(M + 1, n)$,) we get:

$f_1(x) > f_2(x)$ if and only if $x < \dfrac{n}{N+1}(M + 1)$. Thus, $F_1(z) \geq F_2(z)$ when $z \leq \dfrac{n}{N+1}(M + 1)$.

Also note that if $n \leq M$, $F_1(n) = 1 = F_2(n)$, and if $n > M$, then

$$F_1(M) = 1 > F_2(M) = 1 - \frac{\binom{N-M}{n-M-1}}{\binom{N+1}{n}}$$

Thus, $F_1(z) \geq F_2(z)$ for $z \geq \min(M, n)$. For $\dfrac{n}{N+1}(M+1) < z < \min(n, M)$,

$$\begin{aligned} F_1(z) - F_2(z) &> F_1(z+1) - F_2(z+1) > \cdots \\ &> F_1(\min(n, M)) - F_2(\min(n, M)) \geq 0 \end{aligned}$$

because $f_1(z) > f_2(z)$ for $\frac{n}{N+1}(M+1) < z < \min(n, M)$.

- Let $F_3 = h_{N-1,M,n}$, and f_3 be the probability mass function for F_3. To prove $F_1(z) \geq F_3(z)$, we first solve the inequality $f_1(x) > f_3(x)$ for $0 \leq x \leq \min(M, n)$, and get: $f_1(x) > f_3(x)$ if and only if $x < \frac{nM}{N}$.

  Thus, $F_2(z) \geq F_3(z)$ for $z \leq \frac{nM}{N}$. It is also easy to see that $F_1(\min(n, M)) = 1 = F_3(\min(n, M))$. Therefore, by a similar argument as above, we can show that $F_1(z) \geq F_3(z)$ for any z.

- The continuity and the proper convergences to 0 and 1 are easy to satisfy.

  Fix a z such that $0 \leq z < N_m$, define:

$$G_{m,z}(y_m, -r_m) = P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$$

  Clearly $G_{m,z}$ is defined for only a countable number of (nonnegative, nonpositive) integer valued pairs of $(y_m, -r_m)$. First extend $G_{m,z}$ to allow negative integer values for Y_m and/or positive integer values for $-R_m$ by defining

$$G'_{m,z}(y_m, -r_m) = G_{m,z}(\max(0, y_m), \min(0, -r_m))$$

> Clearly $G'_{m,z}$ is still decreasing in $(y_m, -r_m)$. Now extend $G'_{m,z}$ to a continuous decreasing function $G''_{m,z}$. It is easy to check that $G''_{m,z}$ has the desired limit behavior. That is, it goes decreasingly to 0 when either $y_m \to \infty$ or $-r_m \to \infty$, and goes increasingly to 1 when both $y_m \to -\infty$ and $-r_m \to -\infty$.

$G''_{m,z}$ is a dual distribution function in $(y_m, -r_m)$ such that $G''_{m,z} = P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$ a.s. with respect to the measure induced by $(Y_m, -R_m)$. Thus, $P\left(\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \leq z | V_m, W_m\right)$ can also be extended to a function $H_{m,z}$ such that:

$$H_{m,z} = P\left(\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \leq z | V_m, W_m\right) \text{ a.s.}$$

with respect to the measure introduced by $(V_m, -W_m)$, and $H_{m,z}$ is a dual distribution function in $(v, -w)$.

It is easy to check that the distribution function for a $N(\mu, \sigma^2)$ random variable is a dual distribution function in μ. Moreover, if $\mu = x - y$, this distribution function is a dual distribution function in $(x, -y)$. Given that the joint distribution of V_m and W_m converges to a bivariate normal with mean $\mathbf{0}$, and the covariance matrix:

$$\begin{bmatrix} \sigma^2 x & 0 \\ 0 & \sigma^2(1-x) \end{bmatrix}$$

By the lemmas of conditional convergence, the asymptotic normality of the distribution of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$ conditional on $\frac{X_m}{m} = x$ follows immediately.

3. Now we show conditional on $U_m = \frac{1}{\sqrt{m}}(X_m - mp) = u$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp) \Longrightarrow N\left(\sqrt{\gamma}u, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2}\right)p(1-p)\right)$$

The argument is similar to the one used in step 1. Basically, conditional on $U_m = u$, as $m \to \infty$

$$\frac{X_m}{m}\left(1 - \frac{X_m}{m}\right) = \left(p + \frac{u}{\sqrt{m}}\right)\left(1 - p + \frac{u}{\sqrt{m}}\right) \to p(1-p)$$

Then we notice that:

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp) = \frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) + \sqrt{\gamma}u$$

4. Finally, given the asymptotic conditional normality of Z_m given X_m, and the asymptotic normality of X_m, to derive the asymptotic distribution of Z_m using the lemmas of conditional convergence, we only need to show that the conditional distribution of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp)$ given $U_m = u$ is a dual distribution function in u. This is equivalent

to showing that, conditional on $X_m = x_m$, the distribution function of Z_m is a dual distribution function in x_m.

- First, we show that the conditional distribution function of Z_m given $X_m = x_m$ is decreasing in x_m.

  Imagine that we mark each of the m elements in the original sample with a unique label, say, l_i for the i^{th} element. Then after the amplification and the resampling steps, a final sample of size N_m of the m distinct labels is obtained. The probability of getting a specific sample is uniquely determined by the size m of the original sample. Let $C_{x_m,z}$ be the set of possible final samples where the total number of labels $l_1, \cdots, l_{x_m}$ is less than or equal to z, and $C_{x_m+1,z}$ the set of possible final samples where the total number of labels $l_1, \cdots, l_{x_m+1}$ is less than or equal to z. Now $P(Z_m \leq z | X_m = x_m)$ is simply the probability of getting a possible final sample s such that $s \in C_{x_m,z}$, and $P(Z_m \leq z | X_m = x_m + 1)$ is the probability of getting a possible final sample s' such that $s' \in C_{x_m,z}$. Clearly $C_{x_m+1,z} \subset C_{x_m,z}$, which implies:

$$P(Z_m \leq z | X_m = x_m + 1) \leq P(Z_m \leq z | X_m = x_m)$$

- As a function of x_m, $P(Z_m \leq z | X_m = x_m)$ is defined at finite points. We can interpolate linearly between these points, and extend smoothly and increasingly below the smallest point, which is 0, so that the function converges to 1 as $x_m \to -\infty$, and extend smoothly and increasingly above the largest point, which is m, so that the function converges to 0 as $x_m \to \infty$. Call this extended function H_m. Clearly, $H_m = P(Z_m \leq z|X_m = x_m)$ a.s. with respect to the measure introduced by X_m.

Now we have shown that the conditional distribution function of Z_m given $X_m = x_m$ is a dual distribution function in x_m, hence the conditional distribution function of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp)$ given $U_m = u$ is a dual distribution function in u. Given that a normal distribution function is a dual distribution function in its mean, and the fact that $U_m \Longrightarrow N(0, p(1-p))$, by the lemmas of conditional independence,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp) \Longrightarrow N\left(0, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2} + \gamma\right)p(1-p)\right) \tag{3.16}$$

Let $c_m = 1 - \frac{N_m}{m\mu} + \frac{N_m}{m}\left(1 + \frac{\sigma^2}{\mu^2}\right)$, it is easy to check that $\frac{1}{\sqrt{\gamma m}}(N_m - \gamma m) \to 0$ w.p.1. as $m \to \infty$.

□

Theorem 8. *Consider a multinomial sample of size m drawn from a population of $k+1$ categories of elements with relative frequencies p_1, $\cdots$, p_k, and $1 - \sum_{i=1}^{k} p_i$ respectively. Suppose each element of the multinomial sample is subject to i.i.d. amplification processes such that the mean and variance of the amplification factor are μ and σ^2 respectively. A sample of size*

n_m is then drawn with replacement from the intermediate sample. Suppose $\boldsymbol{Z}_m = (Z_{m1}, \cdots, Z_{mk})^T$ is the relative frequencies of the first k categories in the final sample. As $m, n_m \to \infty$, we have:

$$\boldsymbol{Z}_m \sim N(\boldsymbol{p}, \frac{1}{n_m} c_m \Sigma_p) \tag{3.17}$$

where $\boldsymbol{p} = (p_1, \cdots, p_k)^T$, $c_m = 1 + \dfrac{n_m}{m}\left(1 + \dfrac{\sigma^2}{\mu^2}\right)$, and Σ_p is a matrix with $\sigma_{p,ii} = p_i(1 - p_i)$ and $\sigma_{p,ij} = -p_i p_j$ for $i \neq j$.

Proof: The general idea is similar to the proof of Theorem 7. Below is an outline.

Let $\boldsymbol{X}_m = (X_{m1}, \cdots, X_{mk})^T$ and $\boldsymbol{Y}_m = (Y_{m1}, \cdots, Y_{mk})^T$ be the relative frequencies of the first k categories in the original sample and the intermediate sample respectively. Let $\dfrac{n_m}{m} = \gamma_m$. By the central limit theorem for the multinomial data, and Theorem 6 of the asymptotic distribution of the ratio in amplification, as $m, n_m \to \infty$:

$$\sqrt{m}(\boldsymbol{X}_m - \boldsymbol{p}) \implies N(\boldsymbol{0}, \Sigma_p)$$

Conditional on $X_m = \boldsymbol{x}$:

$$\sqrt{m}(\boldsymbol{Y}_m - \boldsymbol{x}) \implies N\left(\boldsymbol{0}, \frac{\sigma^2}{\mu^2}\Sigma_x\right)$$

where Σ_x is a matrix with $\sigma_{x,ii} = x_i(1 - x_i)$, and $\sigma_{x,ij} = -x_i x_j$ for $i \neq j$.

Conditional on $\boldsymbol{Y}_m = \boldsymbol{y}$:

$$\sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{y}) \implies N(\boldsymbol{0}, \Sigma_y)$$

where Σ_y is a matrix with $\sigma_{y,ii} = y_i(1-y_i)$, and $\sigma_{y,ij} = -y_i y_j$ for $i \neq j$. Then we can show that:

- Conditional on $X_m = \boldsymbol{x}$ and $V_m = \sqrt{m}(\boldsymbol{Y}_m - \boldsymbol{x}) = \boldsymbol{v}$, we have $\boldsymbol{Y}_m \to \boldsymbol{x}$, hence:

$$\begin{aligned}\sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{x}) &= \sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{Y}_m) + \sqrt{\frac{n_m}{m}}\boldsymbol{v} \\ &\implies N(\sqrt{\gamma_m}\boldsymbol{v}, \Sigma_x)\end{aligned}$$

- Let $G_{m,z} = P(\boldsymbol{Z}_m \leq z|\boldsymbol{Y}_m = \boldsymbol{y})$. From the probability mass function of the multinomial distribution, it is easy to check that $G_{m,z}$ is continuous and decreasing in $\boldsymbol{Y}_m$ for $0 \leq y_i \leq 1$. Extend $G_{m,z}$ to a function $G'_{m,z}$ defined on $\mathbb{R}^k$ such that when $y_i \to \infty$ for some $1 \leq i \leq k$, $G'_{m,z} \to 0$ decreasingly and continuously, while when $y_i \to \infty$ for all $1 \leq i \leq k$, $G'_{m,z} \to 1$ increasingly and continuously. Because $G'_{m,z} = P(\boldsymbol{Z}_m \leq z|\boldsymbol{Y}_m = \boldsymbol{y})$ a.s. with respect to the measure induced by $\boldsymbol{Y}_m$, by the lemmas of conditional convergence,

$$\sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{x}) \implies N\left(\boldsymbol{0}, \left(1 + \gamma_m \frac{\sigma^2}{\mu^2}\right)\Sigma_x\right)$$

- Given the above result, it can be shown that conditional on $U_m = \sqrt{m}(\boldsymbol{X}_m - \boldsymbol{p}) = \boldsymbol{u}$:

$$\begin{aligned}\sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{p}) &= \sqrt{n_m}(\boldsymbol{Z}_m - \boldsymbol{X}_m) + \sqrt{\frac{n_m}{m}}\boldsymbol{u} \\ &\Longrightarrow N\left(\sqrt{\gamma_m}\boldsymbol{u}, \left(1 + \gamma_m \frac{\sigma^2}{\mu^2}\right)\Sigma_p\right)\end{aligned}$$

It is easy to show that, for any $\boldsymbol{z}$, $P(\boldsymbol{Z}_m \leq z|\boldsymbol{X}_m = \boldsymbol{x})$ is decreasing in $\boldsymbol{x}$, hence there is a dual distribution function $H'_{m,z}$ of $\boldsymbol{x}$ such that $H'_{m,z} = P(\boldsymbol{Z}_m \leq z|\boldsymbol{X}_m = \boldsymbol{x})$ almost surely with respect to the measure induced by $\boldsymbol{X}_m$. Then we can finish the proof by the lemmas of conditional convergence.

□

Chapter 4

SAGE data analysis

In Chapter 3, we have developed the statistical theory for the sampling, amplification, and resampling (SAR) procedure. With the help of this theoretical result, in this chapter, we shall construct a new statistical model of the SAGE data through a detailed analysis of each step of the SAGE protocol. It turns out that two important parameters of the new model are not measured in the current protocol. However, we are going to show that, because of the large number of cells used in the SAGE experiment, the new statistical model can be approximated very well, in practice, by the traditional multinomial model, which does not require those two parameters.

The first section is the discussion of the new SAR model for the SAGE data. In section 2, we present a test for differentially expressed genes based on the SAR model, and show the connection between the new test and the test based on the multinomial model. In section 3, we shall study possible definitions of housekeeping genes, and present a test for housekeeping genes based on our new definition of the housekeeping gene. Section 4 is a discussion of the various approaches to the clustering of the SAGE gene expression level data. Finally, we propose an algorithm for identifying gene markers for cell populations from SAGE data.

4.1 Distribution of the SAGE data

We have yet to find from the literature an explicitly stated statistical model of SAGE data. But it seems reasonable to assume that the current view holds that each SAGE library is a sample following a multinomial distribution, with the number of all tags found in the library being the sample size, and the number of distinct tags being the number of categories. This is confirmed by the fact the only statistical test for differentially expressed genes we found in the SAGE literature assumes that the distribution of the count a type of tag in a SAGE library has a Poisson distribution (Audic and Claverie 1997), and we know that conditional on the sum, a set of k independent Poisson random variables has a k dimensional multinomial distribution.

The choice of the multinomial distribution for the frequencies of n trials with k possible outcomes seems so natural that people may take it granted. However, strictly speaking, a sample following a multinomial distribution usually is either randomly collected from a population of infinite size, or drawn randomly with replacement from a population that could be finite. However SAGE data are not generated in either way. In this section, we shall examine closely the SAGE experiment protocol, and derive a new statistical model for the SAGE data.

From the statistical point of view, a typical SAGE experiment consists of 7 steps. Here we shall discuss each step briefly, and propose appropriate statistical models for each of them.

Step 1: *Get a sample of mRNA transcripts from a sample of cells.*

A sample of cells for a SAGE experiment consists of $10^5 \sim 10^8$ cells, and the sample mRNA transcripts are extracted from these $10^5 \sim 10^8$ cells. Strictly speaking, the sample mRNA transcripts are the sum of $10^5 \sim 10^8$ random samples of the mRNA transcripts. However, from the discussion in Chapter 2, we know that in practice we are unable to discover the statistical associations among the mRNA transcripts of various genes within a single cell. Therefore, for simplicity, we shall model the sampling of the original mRNA transcripts as drawing without replacement from a finite population of mRNA transcripts, hence the sample mRNA transcripts should be a multivariate hypergeometric sample.

If the size of the population is much larger than the size of the sample, the multinomial distribution is a good approximation to the multivariate hypergeometric distribution. Asymptotically, both the multinomial and the multivariate hypergeometric will converge weakly to a multivariate normal, with the latter having a slightly smaller variance (Hajek 1960). If we choose the multinomial distribution to model the sample mRNA transcripts, suppose the relative frequencies of gene

g in the cell population is p, the size of the original mRNA sample is m, then the count of the gene g in this sample follows a binomial distribution with parameters (m, p).

Step 2: *Synthesize cDNA clones from the mRNA transcripts.*

Usually the cDNA synthesis is highly efficient, hence we shall assume that in this step one mRNA transcript generates one and only one cDNA clone. That is, no gene is lost, altered, or added.

Step 3: *Generate tags, which are 10 base long segments cut from a certain site of the cDNAs.*

The relation between the set of tags and the set of genes is not necessarily a bijection. Suppose a tag t can be generated from k distinct unigenes $g_1, \cdots, g_k$, and given a gene g_i, the chance of getting a copy of t is d_i, then the count of tag t after this step is a binomially distributed with parameter (m, q), where $q = \sum_{i=1}^{k} p_i d_i$, and p_i is the relative frequency of gene g_i in the cell population. Note that it is often the case that $q \approx p_j d_j$ for some $1 \leq j \leq k$. That is, the tag t is mainly generated from a single type of gene g_j. Therefore, although the SAGE libraries are counts of tags, we can still learn the information about the expression levels of the genes. For convenience, we shall call q the expression level of tag t in the cell population.

Step 4: *Apply the PCR procedure to boost the counts of the tags.*

For the sake of simplicity, the PCR procedure for each type of tag will be modeled as a single type branching processes with the mean of the offspring distribution equal to $1 + \lambda$, where λ is the efficiency of the PCR procedure. Strictly speaking, though, because of the mutation, the multitype branching process would be a better model for the PCR procedure. Here by mutation we mean that the PCR duplicate of tag t_i might be another tag t_j that differs from t_i by one or more bases. However, the probability distribution for a multitype process is much more complicated. Moreover, because each tag is only 10-base long, the proportion of the mutated duplicates for a tag t_i is small relative to the total PCR duplicates for t_i (Sun 1999). We are also not worried about the mutation caused by deletion or addition, for tags other than 10-base long will not be counted. Therefore, the single type branching process model should be acceptable for our purpose.

Step 5: *Link the tags to form long sequences.*

For simplicity, here we assume that the linkage is perfect. That is, no tag is lost, altered, or added. Note that even if some tags are lost after this step, as long as the loss is random, it will not affect the final model.

Step 6: *Choose those sequences with certain length.*

The tags found in the selected sequences could be modeled as a random sample collected by drawing without replacement from the tags found in the PCR product after step 4. Thus, conditional on the tags in the PCR product after step 4, the tags after step 6 should be a multivariate hypergeometric sample from a population which is the set of all tags found in the PCR product.

Step 7: *Read off tag counts by sequencing the chosen sequences.*

It is claimed that there is a 1% error rate per base for a single pass sequencing. A tag t_i could be read as t_j, which differs from t_i in one or more bases, thus reducing the count of t_i by one and and raising count of t_j by one. This problem is similar to the one caused by the mutation in PCR. Although it only changes the variance of the SAGE data slightly, it does introduce bias to the estimation of the relative frequencies, which could be a big problem for those tags with low relative frequencies. However, in this chapter, we shall ignore the sequencing error, mostly because it is tag and library dependent, hence would be better handled by a preprocess of the SAGE data. For example, we could adjust the count of tag t according to the counts of the tags that differ from that t by only one base.

Theoretically speaking, we have given a statistical model for the SAGE data by specifying a model for each step of the SAGE experiment. Unfortunately, this model is too complicated to be useful. Thus it is useful to know the asymptotic distribution of the SAGE data. Let n be the size of the SAGE library, q the relative frequency of tag t in the cell population (see the discussion of step 3), m the number of mRNA transcripts in the original sample, λ the PCR efficiency, and r the number of PCR cycles. If we ignore the sequencing error in step 7, then a SAGE library is simply the final sample of a SAR procedure. By Theorem 7 of Chapter 3, it is easy to see that the asymptotic distribution of the count Z of a tag t in the SAGE library is:

$$Z \sim N\left(nq, ncq(1-q)\right)$$

where

$$c = 1 - \frac{n}{m(1+\lambda)^r} + \frac{n}{m}\left(1 + \frac{(1+\lambda)^r - 1}{(1+\lambda)^r}\frac{1-\lambda}{1+\lambda}\right)$$

is called the normalizing factor for the sample. Note that c is equal to the ratio of the variance of Z under the new statistical model, to the variance of Z under the multinomial model.

The exact mean of the count of t in the final SAGE data is:

$$\mathrm{E}[Z] = nq$$

When $(1+\lambda)^r \gg \max\left(1, \frac{n}{m}\right)$, in practice, we could assume:

$$c = 1 + \frac{n}{m}\frac{2}{1+\lambda}$$

The distribution of the estimation of the relative frequency of q, then, is:

$$\hat{q} = \frac{Z}{n} \sim N\left(q, \frac{1}{n}cq(1-q)\right)$$

Note that according to the multinomial model, with a similar setting, the mean of the count of tag t in the SAGE library is also nq, though asymptotically the sample relative frequency $\hat{q} = Z/n$ has a normal distribution with mean q and variance $q(1-q)/n$. Both the new model and the multinomial model agree that $\hat{q}$ is an unbiased estimator of the concentration level q of the tag t, though they differ on the variance of $\hat{q}$. The value of $\mathrm{Var}(\hat{q})$ according to the new SAR model is c times the value implied by the multinomial model. Thus, to compare the estimation of the relative frequency q of tag t based on the new SAR model with the estimation based on the multinomial model, we need to know the values of the normalizing factors c.

For a given SAGE library, the value of the normalizing factor c is determined by the values of the following 4 parameters: the number m of all mRNA transcripts in the sample cells, the efficiency λ of the PCR procedure, the number r of the PCR cycles, and the number n of all tags in the SAGE library, i.e., the library size. Unfortunately, the values of two of these parameters, m and λ, are not available under the current SAGE protocol. However, we do know that a typical SAGE experiment needs at least 100,000 cells as input, and the number of mRNA transcripts in each cell is on the order of 10^4. Given the size of a SAGE library is on the order of 10^4, the value of n/m is usually less than 0.0001, which implies that the value of c is less than 1.0002. The fact that the value of c is so close to 1 suggests that, asymptotically, the distribution of $\hat{q}$ is approximately normal with variance $q(1-q)/n$ under the SAR model. Moreover, we can even say that, in a certain sense, the multinomial model is also a good approximation of the exact SAR model.

To show this, let X be the number of tag t after PCR (step 4), and N the number of all tags after PCR, then X/N becomes the relative frequency of tag t after PCR. According to Theorem 1 of Chapter 2, the distribution of X/N is approximately normal with mean q and variance $2q(1-q)/[m(1+\lambda)]$, where m is the number of all mRNA transcripts in the original sample cells, and λ is the efficiency of the PCR. In practice, because $\sqrt{\mathrm{Var}(X/N)} \ll q$, with a very high probability q would be a good approximation of X/N, hence X/N could be treated as the constant q. [1] Recall that ignoring the sequencing error, the final SAGE library is a sample drawing without replacement from the PCR product after step 4. We know that a sample drawing without replacement from a finite population could be treated as a multinomial sample if the the population size is much much larger than the sample size. Thus, given that $N \gg n$, where n is the library size, it is safe to say that Z, the count of tag t in the SAGE library, has approximately a Binomial$(n, X/N)$ distribution. Replacing X/N by the constant q, Z then

[1] Let $n = 1.5 \times 10^9$, which is pretty low for a SAGE experiment, $q = 10^{-4}$, which is also pretty low for an expressed tag, and $\lambda = 0.5$, which is not very efficient either, the standard error of X/N will be approximately 2×10^{-7}, which is much smaller than q.

is approximately Binomial(n, q) distributed.

The above discussion shows that the multinomial distribution could be a good approximation of the SAR model of the SAGE data. However, we should point out that this does not imply that any analysis of the SAGE data based on the multinomial model is also justified. Instead, whenever possible, we should argue directly that the analysis based on the multinomial model is valid, not simply because the multinomial model itself is a good approximation of the SAR model, but more importantly, because the analysis based on the multinomial model can be shown, in practice, also a good approximation of the analysis based on the SAR model.

4.2 Differentially expressed genes

One of the major applications of the SAGE technology is to determine whether the expression levels of the genes have changed over a sequence of time points and/or a series of experimental treatments. Strictly speaking, by the expression level of a gene G in a cell population, we could either mean the expectation of the number of mRNA transcripts of gene G in a single cell in that cell population, or refer to the ratio of the number of the mRNA transcripts of gene G relative to the number of the mRNA transcripts of all the genes in that cell population. However, the SAGE technology provides us only with the information about the relative frequencies of the various tags, which can be mapped to various genes. Thus, given a set of SAGE libraries for each time point and/or each treatment, to test whether a gene is differentially expressed is the same as testing whether the relative frequency of a tag is *not* constant over all the libraries.

Tests based on the multinomial model

The only test we could find in the literature of SAGE data analysis is a pseudo Bayesian test (Audic and Claverie 1997) which assumes that the counts of tags in a library to be independent Poisson random variables. This test can be used to determine whether the relative frequencies of a tag are the same over two SAGE libraries. Consider a tag t. Suppose the library L_1 has $X = x$ copies of t, library L_2 has $Y = y$ copies of t, and the sizes of L_1 and L_2 are n_1 and n_2 respectively. Assume that X and Y are independent random variables such that $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$, by imposing an improper prior on the distribution of λ_1, we can compute the posterior of λ_1 given $X = x$. Then, under the null hypothesis $\lambda_2 = \lambda_1$, the predictive distribution of Y given $X = x$ will be:

$$P(Y = y|X = x) = \left(\frac{n_2}{n_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\dfrac{n_2}{n_1}\right)^{x+y+1}}$$

Given a significance level α, the null hypothesis is rejected if $y < c_1$ or $y > c_2$, where c_1 and c_2 are the $\dfrac{\alpha}{2}$ and $1 - \dfrac{\alpha}{2}$ quantiles of Y.

While this test is fine for comparing the expression levels of a gene between two libraries, it cannot be applied to the cases where we have more than two libraries. As a much more versatile test based on the multinomial model, we would recommend the classic Pearson's χ^2 test. Given s SAGE libraries $L_1, \cdots, L_s$ generated independently from s cell populations, let p_i be the expression level of tag t in the ith cell population, X_i be the count of t in library L_i, n_i be the size of library L_i, and $\hat{p} = \frac{\sum_{i=1}^{s} X_i}{\sum_{i=1}^{s} n_i}$. Under the null hypothesis that the expression levels of t are the same in all the s cell populations, the χ^2 test statistic is defined as:

$$T_1(\boldsymbol{S}) = \sum_{i=1}^{s} \frac{(X_i - n_i\hat{p})^2}{n_i\hat{p}}$$

Note that under the null hypothesis that $p_1 = p_2 = \cdots = p_s$, assuming the multinomial model, conditional on $n_1, \cdots, n_s$ and $\sum_{i=1}^{s} X_i$, $X_1, \cdots, X_s$ are independent of p_1, and jointly have a multivariate hypergeometric distribution. Thus we can compute the exact distribution of T_1 either directly, or by Monte Carlo simulation. Better yet, T_1 has asymptotically a χ^2_{s-1} distribution, and it turns out that for the SAGE data, because the sizes of the libraries are relatively large, assuming the multinomial model, the asymptotic distribution of T_1 is usually a very good approximation of the exact distribution.

For example, let us look at the 6 SAGE libraries from Dr. Peters' SAGE lab at the Graduate School of Public Health of the University of Pittsburgh (see `http://www.genetics.pitt.edu/sage/`). These libraries are the measurements of the gene expression levels of the arterial endothelial tissue under shear stress at 6 time points. The sizes of these 6 libraries are 33876,

Probabilities	Quantiles for T_1	Quantiles for χ_5^2
0.05	1.24	1.15
0.25	2.80	2.67
0.50	4.40	4.35
0.75	6.54	6.63
0.95	10.64	11.07
0.99	14.77	15.09
0.999	20.76	20.52
0.9999	27.80	25.74

Table 4.1: Quantiles of T_1: exact distribution vs χ_5^2

35197, 28684, 38732, 29759, and 27906 respectively. Suppose the total count of a tag t over the 6 libraries is 12. Table 4.1 gives the quantiles for both the exact distribution and the asymptotic distribution of the T_1 statistic for tag t under the null hypothesis that the expression levels of t are the same in the 6 tissues. It is easy to see that, in term of the distribution function, the exact distribution of T_1 is close to that of the χ_5^2 distribution, which is the asymptotic distribution of T_1. Given that a total count of 12 over 6 libraries represents a relatively low expression level, and the fact that the higher the total count, the better the asymptotic distribution as an approximation of the exact distribution for T_1, it is safe to use the asymptotic distribution of T_1 in the tests of differentially expressed genes.

Test based on the new SAR model

From the discussion in Chapter 3, we know that a statistic similar to the Pearson's χ^2 test can be used to test for differentially expressed genes under the new SAR statistical model.

Suppose s SAGE libraries $L_1, \cdots, L_s$ are generated from s cell populations, where $m_1, \cdots, m_s$ are the numbers of mRNA transcripts in the s cell populations respectively. Let n_i be the size of library L_i, λ_i and r_i be the efficiency and the number of cycles of the PCR for L_i, the normalizing factor c_i for the library L_i then is:

$$c_i = 1 - \frac{n_i}{m_i(1+\lambda_i)^{r_i}} + \frac{n_i}{m_i}\left(1 + \frac{(1+\lambda_i)^{r_i} - 1}{(1+\lambda_i)^{r_i}}\frac{1-\lambda_i}{1+\lambda_i}\right)$$

Let Z_i be the count of tag t in library L_i, and $\hat{p} = \frac{\sum_{i=1}^{s} \frac{Z_i}{c_i}}{\sum_{i=1}^{s} \frac{n_i}{c_i}}$. The new statistic is defined as:

$$T_2(\boldsymbol{S}) = \sum_{i=1}^{s} \frac{(Z_i - n_i\hat{p})^2}{c_i n_i \hat{p}}$$

Let p_i be the relative frequency of the tag t in the ith cell population. From the discussion in Chapter 3, we know that, under the null hypothesis that $p_1 = p_2 = \cdots = p_s$, $T_2(\boldsymbol{S})$ asymptotically has a χ^2_{s-1} distribution. Thus, a level α test for the null hypothesis that t has constant expression level over all the s cell populations will be: reject the null hypothesis if and only if $T_2(\boldsymbol{S})$ is greater than the $1-\alpha$ quantile of the χ^2_{s-1} distribution.

It is easy to see that the T_1 and T_2 statistics have similar functional form. The only difference between T_1 and T_2 is that the latter requires the normalization of each SAGE library L_i by the corresponding normalizing factor c_i. Note that when c is very close to 1 (c is always greater than 1), the values of T_1 and T_2 will be also very close, hence it does not matter whether we use T_1 or T_2 to test for differentially expression genes over s cell populations, (assuming we always reject the null if the value of the test statistic is greater than the $1-\alpha$ quantile of χ^2_{s-1}). On the other hand, when c is large, the choice between T_1 and T_2 will lead to significantly different test results. From the previous section, we know that in practice c is very close to 1, which implies that there is little difference between T_1 and T_2. Given that we cannot get the exact values of m and λ under the current SAGE protocol, hence are unable to compute the exact value of T_2, in the remaining part of this chapter, we shall use T_1 as the statistic for testing differentially expression genes, while bearing in mind that T_1 is used as a very good approximation of T_2. [2]

Control of the false discovery rate

To identify genes differentially expressed over two or more cell populations, we need to conduct the statistical tests for all the genes in the whole genome. This means that we are testing the expression levels of thousands of tags simultaneously. On the other hand, it is often the case that only a small fraction of the genes are believed to be expressed differently over the different cell populations. It is thus important to control the false discovery rate (FDR), i.e., the percentage of the tags that are not differentially expressed

[2] Note here we are only claiming that the distribution of T_1 under the SAR model is close the distribution of T_2 under the SAR model. Nevertheless, given the fact that the multinomial distribution could be used to approximate the SAR distribution, it seems plausible to say that the distribution of T_1 under the SAR model is close the distribution of T_1 under the multinomial model. Therefore, it is reasonable to assume that, even under the SAR model, the asymptotic distribution of T_1 is a good approximation of the exact distribution of T_1, given that this is the case under the multinomial model.

but are wrongly determined to be differentially expressed.

One FDR controlling method that is suitable to the testing of differentially expressed genes is Benjamini and Hochberg's *Linear Step Up Multiple Comparison Procedure* (Benjamini & Hochberg, 1995). This procedure requires that the test statistics for all the tests to be either independent or weakly positively dependent. [3] (The T_1 statistics for the genes are not independent, but they are approximately independent, and possibly with positive dependency.) The Benjamini and Hochberg's (BH) procedure first sorts the p-values of the test statistics $p_{(1)} \leq \cdots \leq p_{(k)}$ in ascending order, where k is the number of tests. Then for a given level α, we find the largest i such that $p_{(i)} \leq \alpha i/k$. This procedure will reject all the null hypotheses for the tests with p-values lower than $p_{(i)}$. [4]

Note that when applying the BH procedure, we should only test those tags with at least a moderate total count over all the libraries. The reason is that the vast majority of the genes are barely expressed, and we know apriori that the barely expressed genes cannot at the same time be differentially expressed genes. [5] For example, among the 38651 distinct tags detected in the 6 shear stress libraries, only 1718 of them have a total count of 15 or more (sum over all the 6 libraries). Using the BH procedure, if we only look

[3] This condition has been relaxed in Benjamini & Yekutieli (2001).

[4] There are some new FDR control methods have improved upon the BH procedure. See Benjamini & Yekutieli (2001) and Genovese & Wasserman (2002).

[5] The other reason why it does not make sense to conduct tests on those barely expressed genes is that, because of the sequencing error, it is unreliable to estimate the expression levels of those barely expressed tags, and hence the tests based on these data are also not reliable.

at the tags with a total count of at least 15, at the level of 0.01, 222 tags are determined to be differentially expressed. However, if we include in our tests all the tags regardless of their total counts, at the level of 0.01, only 146 tags would be determined to be differentially expressed.

4.3 Housekeeping genes[6]

Housekeeping genes are usually defined as those genes whose expression levels vary little in different cell populations belonging to the same species of organisms. This definition, however, is not rigorous enough for us to identify housekeeping genes via the statistical analysis of the gene expression level data. Here we are going through a list of rigorous definitions, and choose the ones that seem reasonable.

Global housekeeping genes

The stochastic nature of the chemical processes within a cell means that the number of mRNA transcripts of a gene in a cell must be a random number. Given that housekeeping genes are often used as references to estimate the expression levels of other genes, ideally, we would like the expression levels of the housekeeping genes in a cell to be independent of the specific cell population where the cell comes from. However, the current technology does not allow us to measure the expression level of a gene in a single cell, hence our definitions must take into consideration the limitation of current

[6] This and the next sections are based on a joint work with David Peters

technology.

Let X be the number of mRNA transcripts for a gene H, and N the number of the mRNA transcripts of all genes, in a single cell. Both X and N are random variables whose distributions may depend on another random variable C, the cell population where the single cells come from. Let $\mathrm{E}_c[X]$, $\mathrm{E}_c[N]$, and $\mathrm{E}_[X/N]$ be the expectations of X, N, and X/N in a cell from a population $C = c$ respectively. In general, $\mathrm{E}_c[X]$, $\mathrm{E}_c[N]$, and $\mathrm{E}_c[X/N]$ will be functions of c.

Because of the limitation of current technology, it is impossible to measure X and N directly. However, in most experiments, the sample will contain hundreds of thousands of cells such that, for a gene H, the counting of H per cell in a sample cells from a population c will be extremely close to $\mathrm{E}_c[X]$, and the relative frequency of H in the sample cells will be extremely close to $\mathrm{E}_c[X/N]$. Thus, in practice, we shall assume that the counting of H per cell in the sample cells is equal to $\mathrm{E}_c[X]$, and the relative frequency of H in the sample cells is equal to $\mathrm{E}_c[X/N]$.

Now we could define housekeeping genes according to either of the following two conditions:

(1a). The counting of the mRNA transcripts of a housekeeping gene H per cell in a sample is independent of the cell population where the cells come from, i.e. $\mathrm{E}_c[X]$ is a constant that does not depend on c,

or, alternatively,

(1b). The relative frequency of the mRNA transcripts of a housekeeping gene H in a sample is independent of the cell population where the cells come from, i.e., $\mathrm{E}_c[X/N]$ is a constant that does not depend on c.

For the test of housekeeping genes according to definition (1b), all we need to do is to generate, or collect, a large number of measurements of the expression levels of the genes in various cell populations. Then a gene that is *not* differentially expressed over all the cell populations would be a housekeeping gene. However, the test of housekeeping genes according to definition (1a) is much harder, because it is not always possible to measure the counting of a gene per cell in a sample. First, we usually do not know exactly how many cells are contained in the sample. Second, even if we know the exact number of cells, the experiment procedure may also make the measurement of the counting per cell impossible. For example, in a SAGE experiment, the final result is a library of tags obtained by a SAR procedure. This makes it hard to say, in a SAGE library, the tags are equivalent to how many cells of mRNA transcripts.

Although we may be unable to measure the counting per cell for any gene, it is still possible to test whether it is a housekeeping gene with constant counting per cell, provided we know that some other gene is a housekeeping gene in the same sense. The idea is that the ratio of the countings of two housekeeping genes should be constant over all cell populations (belonging to the same species of organisms). Of course, if we are not sure about any gene being a housekeeping gene, this idea would not work, because the ratio of the countings of any pair of co-expressed genes might be constant for all cell populations.

We first search for the housekeeping genes by identifying the tags that have constant relative frequencies in 121 publicly available SAGE libraries. These libraries consist of a wide varieties of human tissues, tumor or normal, bulk or cell line. We compute the T_1 statistic for 1447 tags whose total counts in all the 121 libraries are at least 448. (Note that $448 \approx 5600160 \times 0.00008$, where 5600160 is the total number of tags in the 121 libraries.) There are several reasons why we set this threshold:

- The tags with low counts are more susceptible to sequencing error.
- Relative high counts ensure that the χ^2_{120} is very close to the exact distribution, and allow us to get p-values without Monte Carlo simulation.
- The test for differentially expressed genes will have higher power for tags with higher counts.

Among the 1447 tags, the best candidate for housekeeping gene is the tag `AAGTGATTCT`, which represents the unigene cluster `Hs. 180677`, splicing factor 1. The T_1 statistic for this tag is 198.89, which gives the p-value of 0.000008. Clearly it is almost impossible for this gene to be a housekeeping gene.

Given the fact that even the tag with best (lowest) T_1 score is unlikely to be a housekeeping gene satisfying definition (1b), it is unlikely that there is any housekeeping gene satisfying definition (1b). The question is then whether some genes satisfying definition (1a). Here we shall confine our search to those genes with the best T_1 scores. The idea is that, while the tags with constant countings per cell are not likely to have constant relative frequencies, nor should their relative frequencies vary too much over different cell populations. Because we do not know any gene satisfying definition (1a) yet, it is impossible to test for housekeeping gene satisfying definition (1a) directly. What we are going to do is to search for a group of the tags with best scores that are co-expressed in the sense that the ratios of the countings of these tags to each other are constant in all the 121 libraries. Here is the test for co-expressed genes:

Consider k tags. Let $X_{i,j}$ be the count of the ith tag in the jth library. The test statistic for the null hypothesis that the ratios of the countings of these k tags to each other are constant in all libraries is a Pearson's χ^2 statistic:

$$T_a = \sum_{i=1}^{k} \sum_{j=1}^{121} \frac{(X_{i,j} - \hat{x}_{i,j})^2}{\hat{x}_{i,j}}$$

where $\hat{x}_{i,j} = (\sum_{h=1}^{k} x_{h,j})(\sum_{l=1}^{121} x_{i,l})/(\sum_{h=1}^{k} \sum_{l=1}^{121} x_{h,l})$.

Although asymptotically T_a has a $\chi^2_{120(k-1)}$ distribution under the null hypothesis, the asymptotic distribution may not be always very close to the exact distribution. Based on the discussion in section 1, we assume that the SAGE data have a multinomial distribution, so that we could compute the exact distribution of T_a. We confine our search to the 40 tags with the lowest T_1 scores. Among the 40 tags, the best, i.e., highest, p-value we get for a pair of tags is for tags `CTGTGCATTT` and `TGTAAGTCTG`, which is about 0.15. The best p-value for a triple of tags is for tags `AATTTGCAAC`, `GTTTCTTCCC`, and `AAAGTCAGAA`, which is about 0.00025. The best p-value we get for a quadruple of tags is for tags `CTCCACAAAT`, `TGCCTTACTT`, `GTCTTAACTC`, and `AGGGGATTCC`, which is less than 0.00001. Based on these results, it is clear that, among the 40 tags with best T_1 scores, no more than 2 could possibly be housekeeping genes in the sense of having constant counting per cell.

Coefficient of variation

From the above analysis, it seems that our definitions of housekeeping genes are too stringent. Here we shall relax the definition to allow some variation in the relative frequencies of the housekeeping genes. (Because of the difficulty in estimating the counting per cell of a gene, from now on we will only consider the definition of housekeeping genes based on the relative frequency of a gene.)

Before giving a new definition, we would first introduce a new concept: the coefficient of variation. Suppose the relative frequencies of the gene H in the 121 SAGE samples are $\boldsymbol{q} = (q_1, \cdots, q_{121})$ such that $\sum_{i=1}^{121} q_i n_i = p \sum_{i=1}^{121} n_i$, then asymptotically T_1 has a noncentral $\chi(\gamma)^2_{121}$ distribution with noncentrality parameter $\gamma = \sum_{i=1}^{121}[n_i(q_i - p)^2/(p - p^2)]$. The *coefficient of variation* of the relative frequency of H then is defined as:

$$d(p, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{121}[(n_i/\sum_{i=1}^{121} n_i)(q_i - p)^2/p^2]} \approx \sqrt{\gamma/(p\sum_{i=1}^{121} n_i)} \approx \sqrt{\gamma/(\sum_{i=1}^{121} X_i)}$$

The coefficient of variation roughly tells on average how much the relative frequencies of a gene in different cell populations deviate from their weighted mean. A new definition of housekeeping gene based on the concept of coefficient of variation is:

(2). A gene is a housekeeping gene if the coefficient of variation of the relative frequency of this gene is less than or equal to 20%.

A level α statistic test for housekeeping genes based on the above definition and the 121 public SAGE libraries would be: rejecting the null hypothesis that the gene H is a housekeeping gene in the sense that the coefficient of variation of the relative frequency of H is less than or equal to 20% if and only if the value of T_1 statistic for H is greater than the $1 - \alpha$ quantile of the $\chi^2_{120}(\gamma)$ distribution, where $\gamma \approx 0.04\sum_{i=1}^{121} X_i$, with X_i being the count of the tag T for gene H in the ith library.

Here we should note that, according to our definition, the coefficient of variation depends on the sizes of the SAGE libraries. This means that the deviation of the relative frequency from the weighted mean in a small library does not contribute to the coefficient of variation as much as the deviation

Tag	T_1 score	Count	γ	p-value
AAGTGATTCT	198.89	521	20.84	0.0017837
GCGGTGAGGT	227.04	615	24.60	0.0000508
CTGTGCATTT	238.67	711	28.44	0.0000184
CGCACCATTG	235.39	639	25.56	0.0000155
GGCCAAAGGC	234.85	523	20.92	0.0000048

Table 4.2: The highest 5 p-values for $d(p, \boldsymbol{q}) = 0.2$

in a large library. In other words, the test is not sensitive to the variation of the relative frequency of a gene in a small library. On the other hand, this insensitivity does make some sense, because a small library provides less information about the relative frequency of a tag than a large library.

Table 4.2 gives the 5 tags with the highest p-values under the null hypothesis that coefficient of variation of their relative frequencies is 0.2. As we mentioned earlier, under the null hypothesis, the T_1 score has asymptotically a noncentral $\chi^2_{120}(\gamma)$ distribution. From the table, it is clear that it is highly unlikely for the coefficient of variation of the relative frequency of any tag to be equal to or less than 0.2. If we modify the null hypothesis by increasing the coefficient of variation to 0.3 and 0.4, the p-values for `AAGTGATTCT` would 0.067 and 0.56 respectively under these two new hypotheses, which suggest that the coefficient of variation of `AAGTGATTCT` probably is around 0.4. [7] Thus, unless we are ready to modify our definition to accept a tag with a coefficient of variation of 0.4 as a housekeeping gene, there would be no housekeeping gene.

[7] As a matter of fact, the tag `AAGTGATTCT` also has the highest p-values for the null hypotheses that the coefficients of variation are 0.3 and 0.4 respectively.

Local housekeeping genes

While there seems to be no gene varying little over all cell populations, we note that in the cases where we may need housekeeping genes to serve as references, we usually compare cell populations closely related to each other. For this kind of applications, it suffices for the reference genes to have constant expression levels among a small class of cell populations. This leads to the definition of weak and strong local housekeeping genes.

(3a). A gene is a weak local housekeeping gene relative to a class of cell populations if the coefficient of variation of the relative frequency of this gene is less than or equal to 20% for all populations in that class.

(3b). A gene is a strong local housekeeping gene relative to a class of cell populations if the relative frequency of this gene is constant over all populations in that class.

The statistics for testing whether a gene is a weak or a strong local housekeeping gene are both the T_1 statistic, though T_1 will have different distributions under different null hypotheses. Let k be the number of libraries, X_i the count of the tag in the ith library, and $\gamma \approx 0.04 \sum_{i=1}^{k} X_i$. If the null hypothesis is that the gene is a strong local housekeeping gene, T_1 will have asymptotically a χ^2_{k-1} distribution; if the null hypothesis is that the gene is a weak housekeeping gene, T_1 will have roughly a $\chi^2_{k-1}(\gamma)$ distribution.

Tag	T_1 Score	Count	γ	p-value
AAAGTCAGAA	25.04	155	6.20	0.8762
CCCTGGCAAT	26.10	149	5.96	0.8383
CTGTGCATTT	27.51	211	8.44	0.8546
GGCCAAAGGC	32.22	180	7.20	0.6401

Table 4.3: The highest 4 p-values for $d(p, \boldsymbol{q}) \leq 0.2$, brain libraries

Among the 121 public SAGE libraries we mentioned before, there are 30 libraries for normal brain tissues, bulk brain tumors, and brain tumor cell lines, and 26 libraries for normal breast tissues, bulk breast tumors, and breast tumor cell lines. Tables 4.3 and 4.4 give the 4 tags with the highest p-values under the null hypothesis that the coefficients of variation of the tags are less than or equal to 0.2 for the 30 brain libraries and 26 breast libraries respectively. In our study, only the tags whose mean relative frequency over the 121 libraries are at least 0.00008 are tested. These include 1487 tags for the brain tissues, and 1494 tags for the breast tissues. From these two tables, there seem to be some genes that are weak local housekeeping genes for the brain tissues and breast tissues respectively. As a matter of fact, there are even 13 and 38 tags for the brain libraries and breast libraries respectively that could pass the test of level 5% for strong local housekeeping genes. Moreover, there is no intersection between the 13 tags for the brain libraries and the 38 tags for the breast libraries. This seems to suggest that when we focus on a class of libraries, we might be able to find some genes whose expression levels vary little within this class.

Tag	T_1 score	Count	γ	p-value
TAAGTAGCAA	25.18	137	5.48	0.7162
GTGCTTGTAC	25.89	94	3.76	0.6083
CATCATTCCT	26.68	87	3.48	0.5544
TGCCTTACTT	27.09	86	3.44	0.5313

Table 4.4: The highest 4 p-values for $d(p, \boldsymbol{q}) \leq 0.2$, breast libraries

However, the claims of the existence of weak local housekeeping genes for the brain libraries and the breast libraries are not conclusive. In short, the data we now have are not sufficient to exclude the possibility of the non-existence of weak local housekeeping genes, especially for the breast libraries. To illustrate, we shall conduct a power analysis of the tests for weak local housekeeping genes in the brain tissues and the breast tissues. We shall test the tags with a new null hypothesis: the coefficient of variation of a tag is at least 30%. The idea is that, if our tests are powerful, then the tags that pass the test for weak local housekeeping genes should have extremely low p-values under the new null hypothesis.

Tables 4.5 and 4.6 give the 4 tags with the lowest p-values under the null hypothesis that the coefficients of variation of the tags are greater than or equal to 0.3, among the 1487 tags for the brain libraries, and the 1494 tags for the breast libraries. The test, at a level of α, will reject the null hypothesis if the value of the T_1 statistic is less than the α quantile of the $\chi^2_{29}(\gamma)$ for the brain tissues, or $\chi^2_{25}(\gamma)$ for the breast tissues. Not surprisingly, these are the same set of tags shown in tables 4.3 and 4.4. Table 4.6

Tag	T_1 Score	Count	γ	p-value
CTGTGCATTT	27.51	211	18.99	0.0227
AAAGTCAGAA	25.04	155	13.95	0.0294
CCCTGGCAAT	26.10	149	13.41	0.0443
GGCCAAAGGC	32.22	180	16.20	0.1119

Table 4.5: The lowest 4 p-values for $d(p, \boldsymbol{q}) \geq 0.3$, brain libraries

Tag	T_1 score	Count	γ	p-value
TAAGTAGCAA	25.18	137	12.33	0.0995
GTGCTTGTAC	25.89	94	8.46	0.2104
CATCATTCCT	26.68	87	7.83	0.2620
TGCCTTACTT	27.09	86	7.74	0.2823

Table 4.6: The lowest 4 p-values for $d(p, \boldsymbol{q}) \geq 0.3$, breast libraries

shows that, even the best candidate for the weak local housekeeping gene for the breast tissues, the integral membrane protein 2B gene that is represented by tag `TAAGTAGCAA`, has a substantial probability (9.95%) of having a coefficient of variation of 0.3. The tags for the brain libraries are doing better. The non-POU domain containing, octamer-binding gene and the ubiquinol-cytochrome c reductase core protein II gene, represented by the tags `CTGTGCATTT` and `AAAGTCAGAA` respectively, seem to be good candidates for the weak local housekeeping genes for the brain tissues.

False discovery rate

So far we have shown that there is no gene whose expression level is constant for all human tissues, and that it is highly unlikely that any gene would have a coefficient of variation of less than 0.2 for all the human tissues. On the other hand, we have also identified a few genes whose coefficients of variation are possibly less than 0.2 for the brain tissues and the breast tissues respectively, and the two tags `CTGTGCATTT` and `AAAGTCAGAA` are especially promising as candidates of weak local housekeeping genes for the brain tissues. Strictly speaking, however, it is possible that even these two tags for the brain tissues are not weak local housekeeping genes. This is because their p-values, under the null hypotheses that the tag's coefficients of variation are at least 0.3, are the lowest two among the 1487 tags. If we would likely to control the false discovery rate of our tests, i.e., the percentage of the tags that have coefficients of variation greater than 0.3 but are wrongly determined to have smaller coefficients of variation by the test, we are not able to claim with reasonable confidence that the coefficients of variation of these two tags are truly less than 0.3.

On the other hand, we would argue that the current methods for controlling false discovery rate may be too conservative for our study. Consider the test for the brain tissues. We agree that,in our study, most of the 1487 tags are not weak local housekeeping genes. However, this does not necessarily imply that when testing these tags for the null hypothesis that the coefficient of variation of each tags is at least 0.3, the statistics for a significant number of the genes whose coefficients of variation are at least 0.3 will have low p-values simply by chance. In other words, while the test may be not very powerful when applied to other data sets, it could very powerful when applied to the SAGE gene expression level data. The reason is that, among those tags whose coefficients of variation are at least 0.3, the vast majority of them actually have coefficients of variation much larger than 0.3. The expression levels of these tags vary so much over the 30 brain libraries that the chance is virtually 0 for these tags to have low enough T_1 scores that would yield low p-values. Therefore, if a tag does have a low T_1 score and a low p-value, it is likely that it is because the tag also has a low coefficient of variation, rather than that it simply gets a low T_1 score by chance.

4.4 Clustering the SAGE data

Clustering of the genes based on their expression levels over different experiments could help us to identify genes that are co-regulated. Theoretically, the clustering algorithm could be applied to the gene expression level data obtained from a set of iid samples of cells from the same cell population. However, for the reason mentioned in Chapter 2, in practice, the clustering algorithm can only be applied to the data obtained from distinct cell populations that are closely related. For example, gene clustering is often applied to the data generated from the tissues under different treatments to identify groups of genes reacting in similar ways under the various treatments.

There are mainly two types of clustering algorithms: the k-means algorithms and the hierarchical algorithms. The hierarchical algorithms require only the existence of a distance function that measures the distance between any two data points. The k-means algorithms, on the other hand, assumes that the data sets live in an Euclidean space.

The choice of distance function is crucial to the results of a clustering study, because the basic idea of the both algorithms is trying to put data points close to each other into the same cluster. In the case of gene clustering, we would like to use the distance functions that measure the similarity between the patterns of the expression levels of two genes. For example, by clustering the shear stress data, which consist of 6 measurements of the gene expression levels made at 6 time points after shear stress is applied to the aortic endothelial tissue, we would like to identify groups of genes that respond to the shear stress in similar ways. For this type of task, the traditional Euclidean distance is not a good choice. Suppose $\boldsymbol{x}_1 = (10, 5, 10, 5, 10, 5)$, $\boldsymbol{x}_2 = (10, 10, 10, 10, 10, 10)$, and $\boldsymbol{x}_3 = (20, 10, 20, 10, 20, 10)$ are the expression levels of three genes, G_1, G_2, and G_3, at the 6 time points in the shear stress experiment. The Euclidean distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is $5\sqrt{3}$, which is much closer than Euclidean distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$, which is $5\sqrt{15}$. However, we would like to say that genes G_1 and G_3 show the same pattern of response to the shear stress, which is quite different from the pattern shown by gene G_2. This implies that the k-means algorithm, which requires the Euclidean distance function, is not suitable for our task. [8]

A better distance function could be derived from the correlation between the expression levels of two genes. For example, we can define the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ as $d_c(\boldsymbol{x}, \boldsymbol{y}) = 1 - \mathrm{Corr}(\boldsymbol{x}, \boldsymbol{y})$. With this distance function, using the above example, we have $d_c(\boldsymbol{x}_1, \boldsymbol{x}_3) = 0 < d_c(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1$. Using

[8] Here we assume no transformation is performed to the SAGE data.

the hierarchical algorithm with this distance function, we have clustered 243 genes that are differentially expressed over the 6 shear stress libraries. As expected, when plotting the genes expression levels against the time points, we found that genes clustered into a tight cluster respond to the shear stress in a similar way.

However, the clustering based on the correlation distance still has two drawbacks. First, it is not clear how tight a cluster need to be in order for us to consider the tags in this cluster to be co-regulated. Second, given the fact that the count of a tag in a SAGE library is approximately binomial distributed, we know that the SAGE data for the highly expressed tags are more accurate than the SAGE data for the barely expressed tags. By using correlation distance we have ignored this piece of information. To incorporate the information about the distribution of the SAGE data into our clustering study, we propose a new distance function d_g. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the expression levels of two tags over k libraries, then the new distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is exactly the log likelihood ratio statistic G^2 for a $2 \times k$ contingency table with first row being $\boldsymbol{x}$ and the second row $\boldsymbol{y}$:

$$d_g(\boldsymbol{x}, \boldsymbol{y}) = 2 \sum_{i=1}^{k} [x_i \log(x_i/\hat{x}_i) + y_i \log(y_i/\hat{y}_i)]$$

where $\hat{x}_i = (x_i + y_i) \sum_{j=1}^{k} x_j / [\sum_{j=1}^{k} (x_j + y_j)]$.

We could also use any other statistics designed for testing association in a two dimensional contingency table, e.g., the Pearson's χ^2 statistics. Or

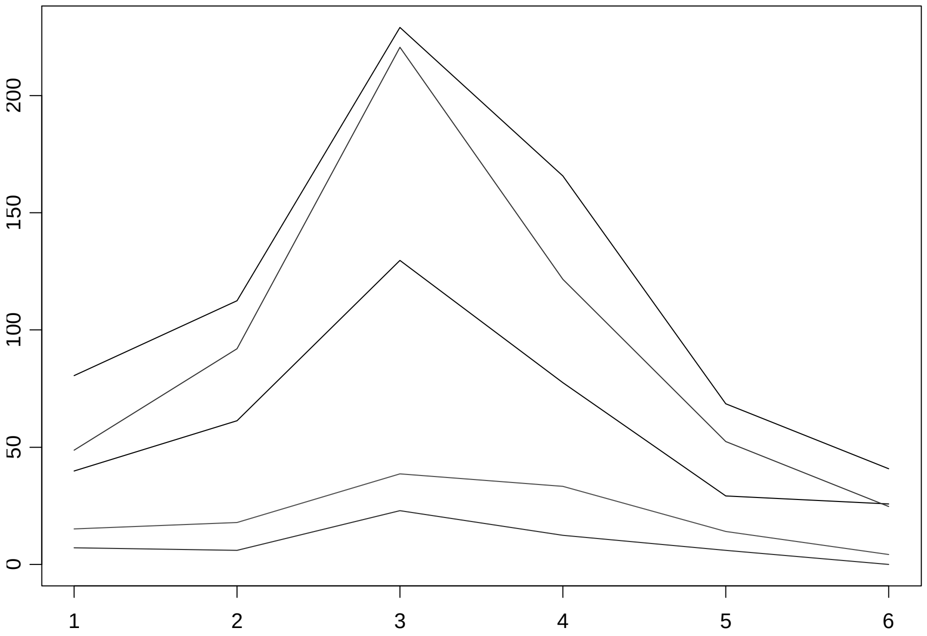


Figure 4.1: Expression levels of 5 tags in the same cluster

even better, we could use $1 - p(T)$ as the distance function, where T is any statistic for testing association in a two dimensional contingency table, and $p(T)$ is the p-value of that statistic under null hypothesis.

It is easy to implement the gene clustering algorithm using the d_g distance function. All we need to do is plug in the new distance function to the hierarchical clustering algorithm. We cluster the expression levels of the 243 genes in the 6 shear stress SAGE experiment using the d_g distance function, and identify 71 small clusters such that within each cluster, the distance between any pair of tags is less than 11.07. [9] While the small-

[9] The hierarchical clustering algorithm implemented here uses the maximum distance between a data point in one cluster and a data point in the other cluster as the distance between the two clusters.

est clusters consist of only a single tag, the largest ones consist of 10 to 14 tags. Note that 11.07 is the 95% quantile of the χ^2_5 distribution, which is the asymptotic distribution of d_g when the number of libraries is 6. Thus the statistical interpretation of the clustering result is: within each cluster, for any pair of tags, if we run a level 5% test with the null hypothesis that these two tags respond to the shear stress in exactly the same way, we are not going to reject the null hypothesis. Figure 4.1 plots the expression levels of the 5 tags, `ATAATTCTTT`, `TAATAAAGGT`, `AAAAATAAAG`, `ATCTTGTTAC`, and `GAAAAATGGT`, which belong to the same cluster.

Although the gene clustering algorithm using distance function d_g allows an intuitive statistical interpretation of the clustering results, it has yet to take into account the fact that the measured expression levels of the highly expressed tags are more reliable than those of the barely expressed tags. This suggests a new type of hierarchical clustering algorithm.

Note that in a typical hierarchical algorithm, to cluster a data set of n data points, we begin with n clusters, where each cluster contains one and only one data point. Then we proceed by merging the closest pair of clusters into a single cluster, until, after $n - 1$ merges, all data points are put into a single cluster. A distinct feature of the traditional hierarchical algorithm is that the distance between two clusters is uniquely determined by the pairwise distances between the data points in one cluster and the data points in the other cluster. [10] This makes the hierarchical algorithm

[10] Here we have a few choices. The distance between two cluster could be:

very easy to use, because it can handle any data set, provided the pairwise distances between the data points are give. However, at the same time, the users are stuck with the few choices of the distance functions between two clusters, which may not always the best choices for the data to be clustered.

Based on our knowledge about the distribution of the SAGE data, we would like to recommend a small modification of the hierarchical clustering algorithm so that it could work better for our data set. The suggested modification to the traditional algorithm is that, instead of using the minimum, or the maximum, or the mean, or the median of the distances between the points in one cluster and the points in the other cluster as the distance between the two clusters, we compute the distance in the following way:

Suppose the first cluster A consists of a data points, $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_a$, the second cluster B consists of b data points, $\boldsymbol{x}_{a+1}, \cdots, \boldsymbol{x}_{a+b}$, and each data point is a k-dimensional non-negative integer valued vector. First we compute the log likelihood ratio statistic $d_g(A, B)$: [11]

$$d_g(A, B) = 2 \sum_{i=1}^{a+b} \sum_{j=1}^{n} x_{i,j} \log(x_{i,j}/\hat{x}_{i,j})$$

a. the minimum distance between a point in one cluster and a point in the other cluster;

b. the maximum distance between a point in one cluster and a point in the other cluster;

c. the mean of the distances between the points in one cluster and the points in the other cluster;

d. the median of the distances between the points in one cluster and the points in the other cluster.

[11] Here we could also use Pearson's χ^2 statistic.

where $\hat{x}_{i,j} = (\sum_{k=1}^{n} x_{i,k})(\sum_{l=1}^{a+b} x_{l,j})/(\sum_{l=1}^{a+b} \sum_{k=1}^{n} x_{l,k})$. [12] The distance between A and B then is $1 - p(d_g(A, B))$, where $p(d_g(A, B))$ is the p-value of the statistic $d_g(A, B)$ under the null hypothesis that there is no association in the $(a+b) \times k$ contingency table, where each row of the table is a data point in A or B. Note that under this null hypothesis, $d_g(A, B)$ has asymptotically a $\chi^2_{(k-1)(a+b-1)}$ distribution.

We can choose a threshold, say, 0.95, and stop merging clusters when the newly defined distance between any two cluster is larger than 0.95. Suppose we use the modified algorithm, with threshold set to 0.95, to cluster the shear stress gene expression level data. The statistical interpretation of the clustering results would be: for each cluster, the null hypothesis that the all the tags in the cluster respond to the shear stress in the same way cannot be rejected at the level of 5%.

4.5 Identifying marker genes

Suppose we have two general groups of cell populations, for example, benign and vicious breast tumor tissues. It would be interesting to know whether the difference in fatality between these two general groups of cell populations can be traced to the difference at the genetic level. In particular, we would like to know whether there is a set of signature genes that are expressed at one level in one cell population, and at a different level in another cell population (van de Vijver et al 2002).

[12] It is assumed that $0 \times \infty = 0$.

To identify such a set of genes, we cannot simply choose a benign breast tumor tissue and a vicious breast tumor tissue, generate two SAGE libraries, and identify those genes that are differentially expressed over the two tissues. For example, suppose we have k types of benign tumors, call them B_1, $\cdots$, B_k, and l types of vicious tumors, call them V_1, $\cdots$, V_l. If we only generate two libraries from B_1 and V_1 respectively, and find that gene G is differentially expressed, we cannot conclude that G is a marker gene, because it is possible that the expression levels of G in B_1 and V_2 are close to each other. Thus, to search for the marker genes, we have to generate SAGE libraries from each subtype of benign tumors and each subtype of vicious tumors. [13]

Suppose we already have the k libraries for the k types of benign tissues, and l libraries for the l types of vicious tissues. How should we proceed to find out the list of marker genes? One approach is based on the test for differentially expressed genes. First, using the test for differentially expressed genes, we could find a list of genes that are *not* differentially expressed over the benign tissues, and a list of genes not differentially expressed over the vicious tissues. Then we collapse the libraries for the benign tumors into a single library where the count of a tag t is the sum of the counts of t over all the benign libraries. Similarly, we collapse the vicious libraries into a single library. Then we use the test for differentially expressed genes again to find

[13] Ideally we would like to have, for each subtype of tumor tissue, a SAGE library. However, in practice, we may have to be satisfied with the libraries generated from some, but not all, subtypes of the tumor tissues.

a list of genes that are differentially expression over these two new libraries. The intersection of these three lists would be a list of marker genes.

The main problem of the above approach is that it is too strong a condition to ask a marker gene to have constant expression levels over the benign tissues and the vicious tissues respectively. For example, suppose we have 3 benign libraries and three vicious libraries, each of size 30000. Let the counts of tag t, which represents gene G, be 0, 10, and 20 respectively in the benign libraries, and 100, 200, and 300 respectively in the vicious libraries. It is obvious that G is a marker gene. However, because the tag t would certainly fail to make the first two lists in the above approach, gene G would not be returned as a maker gene by the above approach.

In this section, we advance a different approach to identifying marker genes. Similar to what we did with the housekeeping genes, our approach begins with an effort to clarify the statistical interpretation of marker genes.

First, we should realize that the expression levels of a marker gene/tag may be neither constant over the benign tissues, nor constant over the vicious tissues. Thus, it would be convenient to treat the expression levels of a gene G or a tag t in the k benign tissues as a random sample of size k from a certain distribution μ_b, and the expression levels of G or t in the l vicious tissues as a random sample of size l from another distribution μ_v. [μ_b (μ_v) will be called the distribution of the expression level of tag t in the benign (vicious) tissues.] Let $F_{q,n}$ be the conditional distribution of the count of a tag in a library of size n given that the library is generated from a tissue where the expression level of that tag is q. Let X be the count of a tag t in a SAGE library of size n generated from a benign tissue, and μ_b the distribution of the expression level of t in the family of benign tissues. It is easy to see that the distribution of X is $F_b(x;n) = \int F_{q,n}(x)d\mu_b(q)$. Similarly, if Y is the count of tag t in a library of size n generated from a vicious tissue, then the distribution of Y would be $F_v(y;n) = \int F_{q,n}(y)d\mu_v(q)$. [$F_b(x;n)$ ($F_v(y;n)$) will be called as the distribution of the count of tag t in a library of size n generated from a benign (vicious) tissue.]

Second, we note that what we expect from a marker gene G is that, given the expression level of G in a tissue, we could determine, with more or less confidence, whether this is a benign tissue or a vicious tissue. This is equivalent to determining whether the count of tag t, which represents G, is a sample from the distribution $F_b(x;n)$ or the distribution $F_v(y;n)$. Obviously, the larger the difference between $F_b(x;n)$ and $F_v(y;n)$, the more confident we are in determining whether the library is generated from a benign tissue or not.

Now we are ready to give a statistical interpretation of the concept of the marker gene. First, we need to choose a distance function to to measure the distance between two distributions. Our choice is the total variation distance, because it can be interpreted in the following way: Suppose a data point is drawn randomly and with equal chance from one of the two distributions μ_1 and μ_2. Given the value of the data point, we are asked to tell from which distribution the data point is drawn. We agree that knowing the distribution functions of μ_1 and μ_2 will help us. Let s be the expected success rate of guessing after knowing the distribution functions. The question is, how much better can we perform than random guessing, which has an expected success rate of 0.5? It turns out that $(s - 0.5)$ is exactly half of the total variation distance between μ_1 and μ_2. Because of the above interpretation, in the context of searching for marker genes, we shall call the total variation distance between two distributions μ_1 and μ_2 the contrast between μ_1 and μ_2, and denote it by $c(\mu_1, \mu_2)$. The statistical interpretation of marker gene, based on the concept of contrast, is then:

Definition 4. *Let* $\boldsymbol{A}$ *be a general group of* k *cell populations:* $A_1, \cdots, A_k$*, and* $\boldsymbol{B}$ *another general group of* l *cell populations:* $B_1, \cdots, B_l$*. Let* t_1 *and* t_2 *be two tags representing two genes* G_1 *and* G_2 *respectively. For* $i = 1, 2$*, let* $F_A^i(x; n)$ *(*$F_B^i(y; n)$*) be the distribution of the count of tag* t_i *in a library of size* n *generated from a sample of cells coming from a cell population in* $\boldsymbol{A}$ *(*$\boldsymbol{B}$*). Then we say* t_1 *is a better marker gene than* t_2 *at size* n *for the two general groups of cell populations if and only if* $c(F_A^1(x; n), F_B^1(y; n)) > c(F_A^2(x; n), F_B^2(y; n))$*.* $c(F_A^i(x; n), F_B^i(y; n))$*, the contrast between* $F_A^i(x; n)$ *and* $F_B^i(y; n)$*, will be called the degree of separation of tag* t_i *or gene* G_i *at size* n*.*

Based on the above interpretation, to search for the marker genes for the two types of breast tumors from the SAGE libraries obtained from the k benign tumors and l vicious tumors, all we need to do is to estimate the degrees of separation at a certain size for all the genes expressed in those tissues, then put them in descending order. To estimate the degree of separation for a tag t at size n, we will need to estimate $F_b(x; n)$ and $F_v(y; n)$.

Given that we already have a statistical model for the SAGE data, it is reasonable to adopt a parametric approach in estimating $F_b(x; n)$ and $F_v(y; n)$. The parametric family of $F_b(x; n)$ and $F_v(y; n)$ depends on the family of $F_{q,n}$, the conditional distribution of the count of tag t in a SAGE library of size n given that the expression level of tag t is q, and the family of μ_b and μ_v, the distributions of the expression levels of t in the two cell populations. Based the discussion in section 1 of this chapter, we shall assume that $F_{q;n}$ is a binomial distribution with parameters (n, q). This suggests that, for the sake of computational simplicity, we shall assume that μ_b and μ_v are beta distributions with parameters (α_b, β_b) and (α_v, β_v) respectively. It then follows that $F_b(x; n)$ and $F_v(y; n)$ are beta binomial distributions with parameters (n, α_b, β_b) and (n, α_v, β_v) respectively.

There are five parameters, n, α_b, β_b, α_v, and β_v, for $F_b(x; n)$ and $F_v(y; n)$. Among them, n is actually an arbitrary constant, and does not needed to be estimated from the SAGE data. Given that the sizes of most SAGE libraries are around 30,000, we shall simply set n to 30,000. We only need to estimate (α_b, β_b) from the benign libraries, and (α_v, β_v) from the vicious libraries.

Let us look at the estimation of (α_b, β_b). Let $\boldsymbol{x} = (x_1, \cdots, x_k)$ be the counts of t in the k SAGE libraries that are generated from the k types of benign tumors respectively. Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ be the sizes of the k libraries. According to the beta-binomial model, the log-likelihood function for the counts of tag t in the k libraries is:

$$\log P(\boldsymbol{x}|\boldsymbol{n}, \alpha_b, \beta_b) = \sum_{i=1}^{k} \log P(x_i|n_i, \alpha_b, \beta_b)$$

It seems straightforward to estimate α_b and β_b using the maximum likelihood estimation (MLE) method. Unfortunately, in practice, we might encounter the cases where x_i/n_i is almost a constant for $i = 1, \cdots, k$. In these cases, the maximum likelihood estimations of α_b and β_b are likely to diverge. To solve this problem, we transform the parameters α_b and β_b into $\theta_b = \alpha_b/(\alpha_b + \beta_b)$ and $\eta_b = 1/(\alpha_b + \beta_b)$. (Note that θ_b and $\theta_b(1-\theta_b)\eta_b$ are the mean and the variance of the beta distribution with parameters (α_b, β_b).) Under the new parameterization, the log-likelihood function of the Beta-Binomial distribution is: [14]

$$\begin{aligned}\log P(\boldsymbol{x}|\boldsymbol{n}, \theta_b, \eta_b) \quad = \quad & \sum_{i=1}^{k} \Bigg[\log \binom{n_i}{x_i} + \log\Gamma(x_i + \theta_b/\eta_b) - \log\Gamma(\theta_b/\eta_b) \\ & + \log\Gamma(n_i - x_i + (1-\theta_b)/\eta_b) - \log\Gamma((1-\theta_b)/\eta_b) \\ & - \log\Gamma(n_i + 1/\eta_b) + \log\Gamma(1/\eta_b)\Bigg]\end{aligned}$$

[14] Here we give two expressions of the log-likelihood function. The second expression may appear to be computationally more efficient, because it does no use the log Gamma function. However, from our experience, the algorithms based on the first expression, which uses the log-likelihood function, could be much faster than those based on the second expression.

$$\begin{aligned} = & \sum_{i=1}^{k}\Bigg[\log\binom{n_i}{x_i} + \sum_{j=0}^{x_i-1}\log(j+\frac{\theta_b}{\eta_b}) \\ & + \sum_{j=0}^{n_i-x_i-1}\log(j+\frac{1-\theta_b}{\eta_b}) - \sum_{j=0}^{n_i-1}\log(j+\frac{1}{\eta_b})\Bigg]\end{aligned}$$

The first derivatives of the log-likelihood function are:

$$\begin{aligned} \frac{\partial}{\partial\theta_b}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) &= \sum_{i=1}^{k}\left[\sum_{j=0}^{x_i-1}\frac{1}{\theta_b+\eta_b j} - \sum_{j=0}^{n_i-x_i-1}\frac{1}{1-\theta_b+\eta_b j}\right] \\ \frac{\partial}{\partial\eta_b}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) &= \sum_{i=1}^{k}\Bigg[\sum_{j=0}^{x_i-1}\frac{j}{\theta_b+\eta_b j} + \sum_{j=0}^{n_i-x_i-1}\frac{j}{1-\theta_b+\eta_b j} \\ &\quad -\sum_{j=0}^{n_i-1}\frac{j}{1+\eta_b j}\Bigg]\end{aligned}$$

In our algorithm, we use a version of gradient descent to get the MLE estimation of θ_b and η_b from the SAGE data. [15] To improve the speed of

[15] Theoretically, the Newton's method may be also a good choice. To use Newton's method, we need the second derivatives of the log-likelihood function:

$$\begin{aligned} \frac{\partial^2}{\partial\theta_b^2}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) &= -\sum_{i=1}^{k}\left[\sum_{j=0}^{x_i-1}\frac{1}{(\theta_b+\eta_b j)^2} + \sum_{j=0}^{n_i-x_i-1}\frac{1}{(1-\theta_b+\eta_b j)^2}\right] \\ \frac{\partial}{\partial\eta_b^2}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) &= -\sum_{i=1}^{k}\Bigg[\sum_{j=0}^{x_i-1}\frac{j^2}{(\theta_b+\eta_b j)^2} + \sum_{j=0}^{n_i-x_i-1}\frac{j^2}{(1-\theta_b+\eta_b j)^2} \\ &\quad -\sum_{j=0}^{n_i-1}\frac{j^2}{(1+\eta_b j)^2}\Bigg] \\ \frac{\partial^2}{\partial\eta_b\partial\theta_b}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) &= \frac{\partial^2}{\partial\theta_b\partial\eta_b}\log P(\boldsymbol{x}|\boldsymbol{n},\theta_b,\eta_b) \\ &= -\sum_{i=1}^{k}\left[\sum_{j=0}^{x_i-1}\frac{j}{(\theta_b+\eta_b j)^2} - \sum_{j=0}^{n_i-x_i-1}\frac{j}{(1-\theta_b+\eta_b j)^2}\right]\end{aligned}$$

However, a comparison of the running times shows that, for our data sets, the Newton's method is much slower than gradient descent.

our algorithm, we use the moment estimations of θ_b and η_b as the starting point (Tamura and Young 1987).

Let $N = \sum_{i=1}^{k} n_i$, the moment estimation of $\theta_b = \alpha_b/(\alpha_b + \beta_b)$ is:

$$\hat{\theta}_b = \frac{\sum_{i=1}^{k} x_i}{N} \tag{4.1}$$

Let $SS = \sum_{i=1}^{k}[(x_i - n_i\hat{\theta}_b)^2/n_i]$.

If $\hat{\theta}_b(1-\hat{\theta}_b)\sum_{i=1}^{k} n_i(1-n_i/N) > SS > \hat{\theta}_b(1-\hat{\theta}_b)(k-1)$, the moment estimation of $\eta_b = 1/(\alpha_b + \beta_b)$ is:

$$\hat{\eta_b} = \frac{SS - \hat{\theta}_b(1-\hat{\theta}_b)(k-1)}{\hat{\theta}_b(1-\hat{\theta}_b)\sum_{i=1}^{k} n_i(1-n_i/N) - SS} \tag{4.2}$$

Otherwise, $\hat{\eta_b} = 0$.

We apply our algorithm to the comparison of 3 SAGE libraries generated from different types of human aortic endothelial tissues with 35 libraries of other human tissues, including venous and lymphatic endothelial tissues. The resulting 10 best marker genes for the aortic endothelial tissues, the tags representing them, and the degrees of separation at size 30000 for these tags are given in Table 4.7.

The concept of marker genes can also be defined for a set of more than two general groups of cell populations. To do so, we only need to extend the definition of contrast to measure the difference among a set of more than two distributions:

Definition 5. *Consider k distributions $\mu_1, \cdots, \mu_k$ on the space $(X, \mathcal{X})$. Let*

Tag	Separation	Gene
CCACCCTCAC	0.99	heparan sulfate proteoglycan 2 (perlecan)
GCCACCACCA	0.99	intercellular adhesion molecule 2
TTTGCACCTT	0.99	connective tissue growth factor
GCAGAGCAGT	0.98	lymphoblastic leukemia derived sequence 1
CTTTGTTTTG	0.98	chromosome 20 open reading frame 43
TGCTGACTCC	0.98	hypothetical protein FLJ21841
AAACCAAAAA	0.97	endoglin (Osler-Rendu-Weber syndrome 1)
GAAGCAGGAC	0.97	cofilin 1 (non-muscle)
TGTCATCACA	0.97	lysyl oxidase-like 2
CCTAGCTGGA	0.96	peptidylprolyl isomerase A (cyclophilin A)

Table 4.7: The 10 best marker genes for aortic endothelium

f_1, $\cdots$, f_k be their derivatives with respect to a common measure ν. The contrast among $\mu_1, \cdots, \mu_2$ is given by:

$$c(\mu_1, \cdots, \mu_k) = \frac{1}{k} \int_X \max_{1 \le i \le k} \left\{ f_i - \frac{1}{k-1} \sum_{j \ne i} f_j \right\} d\nu \tag{4.3}$$

The search algorithm for the marker genes between two general groups of cell populations could be used, with virtually no modification, to identify marker genes among more than two general groups of cell populations.

Chapter 5

Conclusion

In this thesis I first discussed the problem of aggregation, and argued that, as long as gene expression level data are obtained from aggregations of genes, we could not learn the conditional independence relations, and to some extent, even the correlations, among the expression levels of the genes in a single cell. Then I showed how to model the SAGE data, and how to learn from the SAGE data. Algorithms were proposed for searching for housekeeping genes, clustering genes, and identifying marker genes. Not surprisingly, the SAGE data model, and all these algorithms, are based only on the expectations of the gene expression levels, the only piece of information we could learn reliably given our current state of technology.

However, I am not claiming that it is impossible to learn anything about the gene regulating mechanism beyond the expected expression levels. As suggested at the end of chapter 2, there are currently two approaches to the learning of the gene regulatory network that could, under certain conditions, bypass the problem of aggregation. One is the intervention approach, where the expected expression levels of some genes are experimentally manipulated, usually the genes are totally knocked out. It is expected that such dramatic changes in the expected expression levels of the regulating genes could lead to changes in the expected expression levels of the regulated genes, which could be measured using current technology, including SAGE and microarray. While this approach might be not practical in the discovery of the gene regulatory network (Danks et al, 2002), it certainly could be used to confirm the regulatory networks obtained via other approaches.

The other approach is the genome-wide location analysis. The idea is to determine, for each transcriptional factor f, the set of genes whose promoter regions are bound to a compound containing f. Here again only the expected relative frequencies of the promoter regions are needed to identify genes regulated by a transcriptional factor, hence it is not affected by the problem of aggregation. Suppose gene G_1 encodes protein f, and f is found in the compound bound to the promoter region of gene G_2, then obviously G_1 regulates G_2. Based on pieces of such information, we can build directly a gene regulatory network. This technology is very promising because the number of measurements needed to construct a global regulatory network is linear in the number of transcriptional factors. Indeed, a global regulatory network for yeast has been build using this approach (Lee et al, 2002).

Finally, I would like to point out that the technology for analyzing the expression levels of a small set of genes in a single cell has been around for a few years (Emmert-Buch et al 1996). We should not ignore the possibility that the development of the technology will soon make the measurement of the expression levels of all genes in a single cell efficient. If this is the case, the causal inference algorithms mentioned in chapter 2 could be again used to discover gene regulatory networks. Of course, we need to be aware of the effect of PCR on the variance of the measured gene expression levels, because now a relatively small number of mRNAs are used as input. For example, if the new technology is an improved version of the SAGE technology that requires only a single cell as input, the distributions of the counts of the tags in the final data could be no longer approximated by multinomial distributions. Instead, we should use the SAR model described in section 1 of chapter 4, and make sure that the protocol of the experiment includes the measurement of the parameters of the SAR model, especially the number of all mRNAs in the input cell and the PCR efficiency.

Reference

Akutsu, T., Miyano, S., Kuhara, S. (1998), Identification Of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model, *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms*, 695-702

Ash, R., Doleans-Dade, C. (2000), *Probability and Measure Theory* (2nd ed.), San Diego, CA: Academic Press.

Audic, S. & Claverie, J. (1997), "The Significance of Digital Gene Expression Profiles", *Genome Research* **7:** 986–995.

Bar-Joseph, Z., Gifford, D., & Jaakkola, T., (2001), "Fast optimal leaf ordering for hierarchical clustering", *Bioinformatics 17 Suppl. 1*: pp. S22-S29.

Benjamini, Y. & Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society, Series B*, **57**, pp. 289-300.

Benjamini, Y. & Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency", *The Annals of Statistics* **29**.

Bishop, Y., Fienberg, S., & Holland, P. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: The MIT Press.

Chen, Y., Dougherty, E., & Bittner, M., (1997), “Ration-based Decisions and the Quantitative Analysis of cDNA Microarrya Images”, *Journal of Biomedical Optics 2(4)*: pp. 364-374

Chu, T. (2002), “Sampling, Amplifying, and Resampling”, *Tech Report.*

Danks D., & Glymour, C., (2002), “Linearity Properties of Bayes Nets with Binary Variables”, *Proceedings of the Conference on Uncertainty in Artificial Intelligence 2001*, Seattle.

Danks, D., Glymour, C., & Spirtes, P. (2002), “Inference and Experimental Design for the Discovery of Genetic Regulatory Structure through Experimental Interventions: Statistical Realism and Combinatorial Complexity”, forthcoming, *Bioinformatics.*

Datson, N., de Jong, J., van den Berg, M., de Kloet, E., Vreugdenhil, E., (1999), ”MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue”, *Nucleic Acids Research, 27 (5)*: pp.1300-1307.

Davidson, E., Rast, J., Oliveri, P., Ransick, A., Calestani, C., Yuh, C., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C., Livi, C., Lee, P., Revilla, R., Rust, A., Pan, Z., Schilstra, M., Clarke, P., Arnone, M., Rowen, L., Cameron, R., McClay, D., Hood, L, & Bolouri, H. (2002), A Genomic Regulatory Network for Development, *Science*, **295**, 1669-1678.

D’haeseleer, P. (2000), Reconstructing Gene Networks from Large Scale Gene Expression Data, Ph.D Thesis, University of New Mexico

D'haeseleer, P., Liang, S., & Somogyi, R., (2000) Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering, *Bioinformatics*, **16(8)**,707-26.

Dudoit, S., Yang, Y., Callow, M., & Speed, T., (2000), "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Technical report 578*, Department of Statistics, University of California, Berkeley

Eisen, M., Spellman, . P., Brown, P., & Botstein, D., (1998), "Cluster analysis and display of the genome-wide expression patterns", *Proceedings of the National Academy of Sciences 95*: pp. 14863-14868.

Emmert-Buck, M., Bonner, R., Smith, P., Chuaqui, R., Zhuang, Z., Goldstein, S., Weiss, R., & Liotta, L. (1996), "Laser Capture Microdissection", *Science* **274**: 998–1001

Friedman, N., Nachman I., & Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data", *Recomb 2000*, Tokyo

Genovese, C. & Wasserman, L. (2002), "Operating Characteristics and Extensions of the FDR Procedure", *Journal of the Royal Statistical Society, Series B* **64**: 499-518

Hajek, J. (1960), "Limit Distributions in Simple Random Sampling from a Finite Population", *Publ. Math. Inst. Hungar. Acad. Sci.* **5**: 361–374.

Hartemink, A. (2001), *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*, Ph.D Thesis, MIT,

Hereford, L., Rosbash, M., (1977), "Number and distribution of polyadenylated RNA sequences in yeast", *Cell 10*: pp.453-462

Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., & Hood, L. (2001), Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network, *Science*, **292**, 929-934

Kerr, M. & Churchill, G., (2001a), "Experimental design for gene expression microarrays", *Biostatistics 2*: pp.183-201.

Kerr, M. & Churchill, G., (2001b), "Statistical design and the analysis of gene expression microarrays", *Genetical Research 77*: pp.123-128.

Liang, S., Fuhrman, S., Somogyi, R. (1998), REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures, *Pacific Symposium on Biocomputing*, **3**, 18-29

Orlando, V. (2000), "Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation", *Trends in Biochemical Sciences 25 (3)* pp.99-104

Richardson, T., (1996), *Models of Feedback: Interpretation and Discovery*, PhD Thesis, Department of Philosophy, Carnegie Mellon University.

Robins, J., Scheines, R., Spirtes, P., & Wasserman, L. (2000), "Uniform Consistency In Causal Inference", Technical Report, Department of Statistics, Carnegie Mellon University.

Shrager, J., Langley, P., & Pohorille, A. (2002), Guiding Revision of Regulatory Models with Expression Data, *Proc. of the Pacific Symposium on BioComputing*, **7**, 486-497

Spirtes, P., Glymour, C., & Scheines, R. (2001) *Causation, Prediction and Search*, Cambridge, MIT Press.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., & Wimberly, F., (2001), "Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data", *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology*, Duke University.

Stein, E. & Weiss, G. (1971) *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton: Princeton University Press

Sun, F. (1995), "The Polymerase Chain Reaction and Branching Processes", *Journal of Computational Biology* **23:** 3034–3040

Tamura, R. and Young, S. (1987), A Stablilized Moment Estimator for the Beta-Binomial Distribution, *Biometric 43*: pp813-824

Tusher, V., Tibshirani, R., & Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response", *Proceedings of the National Academy of Sciences 98 (9)*: pp. 5116-5121.

van de Vijver, M., He, Y., van 't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., & Bernards, R., (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer", *The New England Journal of Medicine 347 (25)*: pp.1999-2009

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge, UK: Cambridge

University Press.

Velculescu, V., Zhang, L., Vogelstein, B., & Kinzler, K. (1995), "Serial Analysis of Gene Expression", *Science* **270**: 484–487

Velculescu, V., Zhang, L., Zhou, W., Traverso, G., St. Croix, B., Vogelstein, B., & Kinzler, K. (2000), "Serial Analysis of Gene Expression Detailed Protocol", version 1.0e, John Hopkins Oncology Center and Howard Hughes Medical Center.

Velculescu, V., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M., Bassett, D., Hieter, P., Vogelstein, B., Kinzler, W., (1997), "Characterization of the Yeast Transcriptome", *Cell 88*: pp.243-251.

Warrington, J., Nair, A., Mahadevappa, M., & Tsyganskaya, M. (2000), "Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes", *Physiological Genomics* **2:** 143–147

Yang, Y., Buckley, M., Dudoit, S., & Speed, T., (2000), "Comparison of methods for image analysis on cDNA microarray data", *Technical report 584*, Department of Statistics, University of California, Berkeley

Yang, Y. & Speed, T., (2002), "Design issues for cDNA microarray experiments", *Nature Reviews 3*: pp.579-588.

Yoo, C., Thorsson V., & Cooper, G.F., (2002), Discovery of Causal Relationships in a Gene-Regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data, *Proc. of the Pacific Symposium on BioComputing*, **7**, 498-509

Yuh, C., Bolouri, H., & Davidson, E. (1998), Genomic Cis-Regulatory Logic:

Experimental and Computational Analysis of a Sea Urchin Gene, *Science*, **279**, 1896-1902.

Web Site References

GeneEd Biotechnology Glossary

—`http://www.geneed.com/glossary/index.html`

SAGE Anatomic Viewer at the website of the Cancer Genome Anatomy Project (CGAP), National Cancer Institute (NCI)

—`http://cgap.nci.nih.gov/SAGE/AnatomicViewer`

SAGE lab at the Graduate School of Public Health of the University of Pittsburgh

—`http://www.genetics.pitt.edu/sage/`

SAGEmap at National Center for Biotechnology Information (NCBI)

—`http://www.ncbi.nlm.nih.gov/SAGE/`

SAGEmap's FTP site at NCBI

—`ftp://ftp.ncbi.nih.gov/pub/sage`

Schilstra, M. (2002), NetBuilder

—`http://strc.herts.ac.uk/bio/maria/NetBuilder`

Glossary

A great online source for the biological terminology is the GeneEd Biotechnology Glossary at `http://www.geneed.com/glossary/index.html`.

FDR: False Discovery Rate, the expectation of the ratio of the number of tests where the null hypothesis is falsely rejected to the number of tests where the null hypothesis is rejected.

Microarray: A technology for measuring gene expression levels. See Chapter 1, pp. 1–3.

SAGE: Serial Analysis of Gene Expression, a technology for measuring gene expression levels. See Chapter 1, pp. 3–5.

SAR: Sampling, Amplification, and Resampling, a sampling scheme abstracted from the SAGE protocol. See Chapter 3. pp. 58–59.

Tag: In this thesis by a tag we mean a SAGE tag, a 10-base long fragment of cDNA sequence that could be used to represent the mRNA transcript of a gene. See Chapter 1, pp. 3–4.