

Limitations of Statistical Learning from Gene Expression Data

Tianjiao Chu
Institute for Human and Machine Cognition
University of West Florida
Pensacola, FL 32502

Abstract

Current technologies for measuring gene expression levels, such as microarray and SAGE, measure the summed expression levels of the genes from a large aggregate of cells, rather than the expression levels of the genes in an individual cell. This paper discusses, from the statistical point of view, what we could learn, both in principle and in practice, from the microarray and SAGE gene expression level data. We show that, when the summed gene expression levels are measured from a large number of cells, the conditional independence relations among the summed gene expression levels are essentially determined by the correlation matrix among the gene expression levels, and are very unlikely to be the same as the conditional independence relations among the expression levels of the genes in an individual cell. 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. Furthermore, we show that, in practice, we probably could not even get an accurate estimation of the correlation matrix of the gene expression levels, for the number of experiments required to estimate the correlation matrix is too large to be feasible. Therefore, the only piece of information we can learn reliably from the current gene expression level data is the expected gene expression levels.

1 Introduction

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 figuring 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. However, there are some statistical difficulties to the causal inference approach of deriving gene regulatory network from gene expression data. Some of these difficulties—such as the presence of latent common causes and cycles—have, in principle, been overcome (Spirtes, et al, 2001). But the effort of making causal inference from gene expression data reveals another elementary statistical difficulty: the problem of aggregation. That is, 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. We have shown, in previous study, that in general 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 (Chu et al 2003). We also gave two sufficient conditions for the conditional independencies to be preserved under aggregation. In this paper, we 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 result 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 next section is a short introduction of the basic ideas of causal inference. In section 3, we discuss the problem of aggregation and give a brief review of the results of our previous study (Chu et al 2003). We then present the main result of this study and its implications. The proofs of the theorems in this paper are given in the appendix.

2 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 1, we could specify the following functional relations:

$$\begin{aligned} Z &= f(Y, W) + \epsilon_z \\ Y &= g(X) + \epsilon_y \\ W &= h(X) + \epsilon_w \end{aligned} \tag{1}$$

Where f , g , h are any functions and ϵ_z , ϵ_y , ϵ_w are independently distributed noises. It follows that the joint probability density of Z , Y , W , X admits a Markov

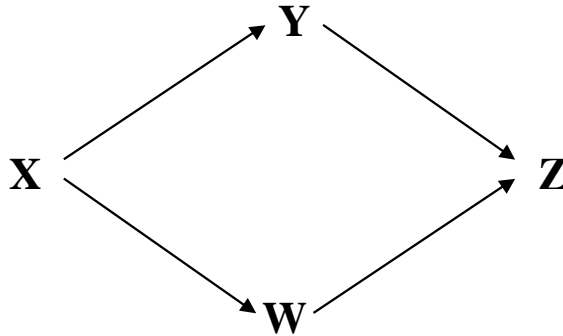


Figure 1: A simple causal graph

factorization

$$d(X, Y, Z, W) = d(Z|Y, W)d(Y|X)d(W|X)d(X) \quad (2)$$

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 , \mathbf{Y} be the set of parents of X in G , and \mathbf{Z} a set of variables that are neither parents nor descendants of X . Then conditional on \mathbf{Y} , X and \mathbf{Z} are independent.*

The Markov condition is a sufficient condition for conditional independence relation in the sense that a conditional independence relation predicted by the Markov condition must be observed. However, it is possible that in a causal model, for some special combinations of parameter values, some observed conditional independence relations are not predicted by the Markov condition. 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 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint sets of variables in a causal model G . Then \mathbf{X} and \mathbf{Y} are independent given \mathbf{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 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 .

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

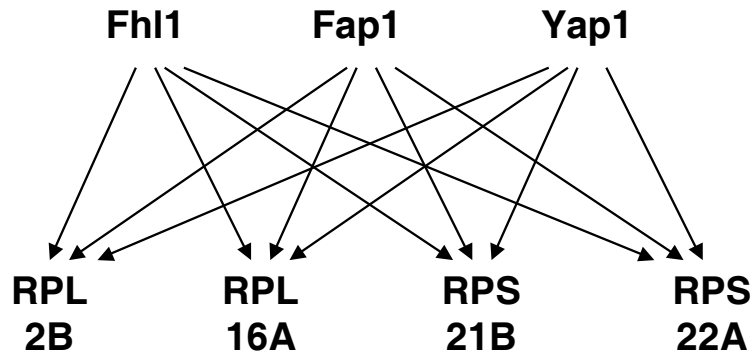


Figure 2: A yeast gene regulatory network

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

3 Conditional independence under aggregation

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 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 i th 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 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 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 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 1 (Danks and Glymour, 2001).

A graphical heuristic explanation of the problem of aggregation is shown in figure 3. Here we consider a simple regulatory network: Gene X regulates gene Y , and Y regulates gene Z . Figure 3(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 3(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 3(b) does not imply that $\sum_{i=1}^2 X_i$ and $\sum_{i=1}^2 Z_i$ are independent conditional on $\sum_{i=1}^2 Y_i$.

There are some special cases where the conditional independencies are invariant under aggregation. For example, although the graph in figure 3(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).

Another more interesting special is when the causal system can be represented as

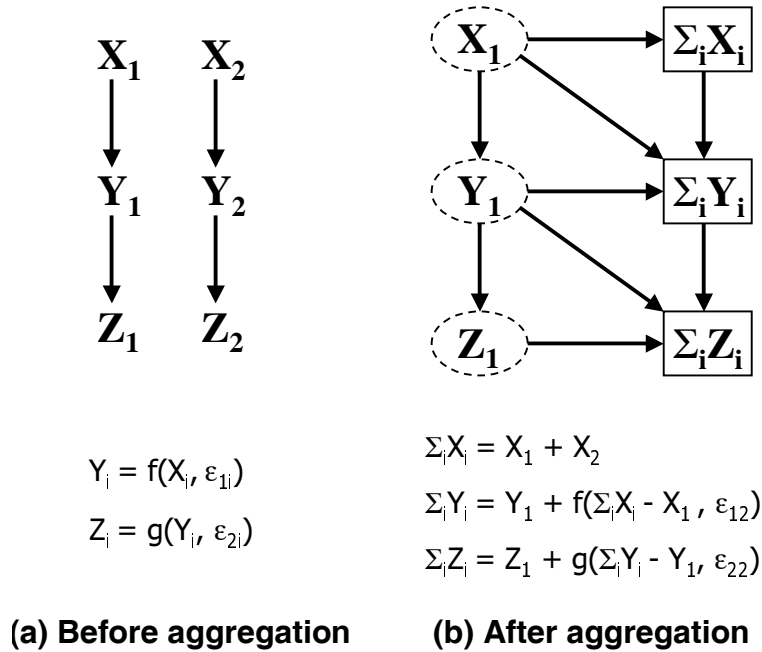


Figure 3: Problem of aggregation

linear model. That is, the noise terms, as in equations 1, are normally distributed and each variable is a linear function of its parents and an independent Gaussian noise. Under this condition, for any directed acyclic graph, the set of conditional independencies implied by the graph will hold for the summed variables. This is because the conditional independence for linear model is equivalent to vanishing partial correlation, and partial correlation is invariant under aggregation.

Chu et al (2003) gives another two less restrictive sufficient conditions for conditional independence of variables to be the same as the conditional independence of their sums. Not surprisingly, both conditions are related to linear model. These two conditions 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 \mathbf{V} of causal sufficient random variables.² Let $Y, X^1, \dots, X^k \in \mathbf{V}$, and $\mathbf{X} = \{X^1, \dots, X^k\}$ be the set of parents of Y in G . If $Y = \mathbf{c}^T \mathbf{X} + \epsilon$,³ where $\mathbf{c}^T = (c^1, \dots, c^k)$, and ϵ is a noise term independent of all non-descendants of Y , then Y is independent of all its non-parents and non-descendants conditional on its parents \mathbf{X} , and this relation holds under aggregation.*

The above theorem states that, under the local linearity condition, the condi-

²A set \mathbf{V} of random variables are causal sufficient if, for any $X, Y \in \mathbf{V}$, if Z is a common cause of X and Y , then $Z \in \mathbf{V}$.

³In 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.

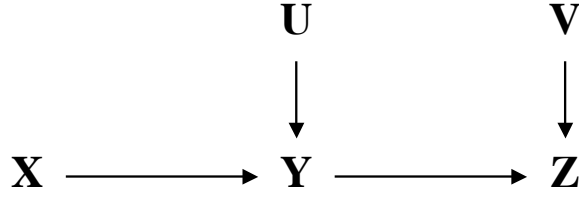


Figure 4: A Sea Urchin type regulatory network

tional independence relation between a random variable and its non-descendant 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 \mathbf{V} of random variables. Let $\mathbf{X} = \{X^1, \dots, X^h\}$, $\mathbf{Y} = \{Y^1, \dots, Y^k\}$, $\mathbf{W} = \{W^1, \dots, W^m\}$, and $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W} = \mathbf{V}$. Suppose that the following three conditions hold:*

1. *The joint distribution of $X^1, \dots, X^h, Y^1, \dots, Y^k$ is multivariate normal with nonsingular covariance matrix.*
2. *For $i = 1, \dots, k$, Y^i is neither a parent, nor a child, of any variable $W^j \in \mathbf{W}$. That is, there is no direct edge between a variable in \mathbf{Y} and a variable in \mathbf{W} .*
3. *For $i = 1, \dots, h$, X^i is not a child of any variable $W^j \in \mathbf{W}$. That is, if there is an edge between a variable in \mathbf{X} and a variable in \mathbf{W} , the direction of the edge must be from the variable in \mathbf{X} to the variable in \mathbf{W} .*

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

Unfortunately, there is no evidence supporting that linear models could be used as good approximation for the known gene regulatory networks. Consider the network regulating the expression of Endo16 gene of sea urchin (Yuh, et al., 1998), which is arguably one of the best-established regulatory network with known functional relations. In this network, the expression level of the Endo16 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. The causal structure shown in figure 4 is a simplification of the proposed Sea Urchin endo16 gene regulatory network. Let $Y = UX$ and $Z = VY$. Suppose X has a Poisson distribution with parameter λ , U and V are Bernoulli random variables with parameters p_1 and p_2 respectively.

Then it can be shown that, although X and Z are independent conditional on Y , as long as U and V are not degenerate, that is, neither U nor V is a constant, this conditional independence relation is not preserved under aggregation (Chu et al 2003).

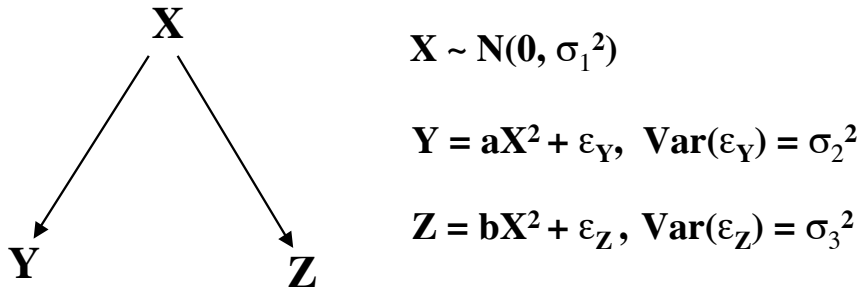


Figure 5: Covariance matrix and conditional independence

4 Conditional independence among large sample means

The discussion in section 3 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 4) 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, Z) has a multivariate normal distribution, then X and Y are independent given Z if and only if the partial covariance/correlation of X and Y with respect to 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 5. The covariance matrix for (X, Y, Z) is:

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

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:

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

However, it is easy to see, from figure 5, that Y and Z are *independent* given X , and that X and Z are *dependent* given Y .

Nevertheless, 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, 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 $\sum_i Z_i$ converges in total variation distance to the product of the densities of $\sum_i X_i$ given $\sum_i Z_i$ and $\sum_i Y_i$ given $\sum_i Z_i$ if and only if the partial correlation of X and Y with respect to Z is 0. To prove this, we need a few lemmas about the characteristic functions of multivariate distributions.

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

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

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

With the above lemmas, we can prove the first main theorem of this paper, 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 \mathbf{X}_n be i.i.d. random vectors with 0 mean and non-singular covariance matrix Σ_X , and $\overline{\mathbf{X}}_n = \sum_{i=1}^n \mathbf{X}_i / \sqrt{n}$. Suppose the characteristic function $\phi(\mathbf{t}) = E[\exp(\mathbf{t}^T \mathbf{X})]$ is integrable, then $\overline{\mathbf{X}}_n$ have bounded continuous densities that converge uniformly to the density of a multivariate normal distribution with 0 mean and covariance matrix Σ_X .*

The following corollary is a direct consequence of Theorem 3.

Corollary 1. *Let $\{(X_n, Y_n, \mathbf{Z}_n)\}$ be a sequence of i.i.d. $k+2$ dimensional random vectors with mean $\mathbf{0}$ and nonsingular covariance matrix Σ . Suppose (X_n, Y_n, \mathbf{Z}_n) and \mathbf{Z}_n both have bounded densities (with respect to the Lebesgue measure). Let $\bar{X}_n = (\sum_{i=1}^n X_i)/\sqrt{n}$, $\bar{Y}_n = (\sum_{i=1}^n Y_i)/\sqrt{n}$, and $\bar{\mathbf{Z}}_n = (\sum_{i=1}^n \mathbf{Z}_i)/\sqrt{n}$, and (U, V, \mathbf{W}) be a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix Σ . Then the total variation distance between the conditional distribution of (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$ and the product of the conditional distributions of \bar{X}_n given $\bar{\mathbf{Z}}_n$ and \bar{Y}_n given $\bar{\mathbf{Z}}_n$ converges to the total variation distance between the conditional distribution of (U, V) given \mathbf{W} and the product of the conditional distributions of U given \mathbf{W} and V given \mathbf{W} almost surely with respect to the measure induced by \mathbf{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:

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

Therefore, $\text{Var}(\bar{X}_n, \bar{Y}_n, \bar{\mathbf{Z}}_n) = \text{Var}(U, V, \mathbf{W})$. Now consider the case where X and Y are independent given \mathbf{Z} , but the partial correlation of X and Y with respect to \mathbf{Z} is not 0. Clearly, the partial correlation of U and V with respect to \mathbf{W} cannot be 0 either, hence U and V must be dependent given \mathbf{W} . As we aggregate more and more cells, the total variation distance between the conditional distribution of (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$ and the product of the conditional distributions of \bar{X}_n given $\bar{\mathbf{Z}}_n$ and \bar{Y}_n given $\bar{\mathbf{Z}}_n$ converges to a positive value.⁴ Therefore, there must be a number N such that for all $n \geq N$, \bar{X}_n and \bar{Y}_n are dependent given $\bar{\mathbf{Z}}_n$. On the other hand, if the partial correlation of X and Y with respect to \mathbf{Z} is 0, then total variation distance between the conditional distribution of (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$ and the product of the conditional distributions of \bar{X}_n given $\bar{\mathbf{Z}}_n$ and \bar{Y}_n given $\bar{\mathbf{Z}}_n$ 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 (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$ from the product of the conditional distributions of \bar{X}_n given $\bar{\mathbf{Z}}_n$ and \bar{Y}_n given $\bar{\mathbf{Z}}_n$. Hence the power of any general test of conditional independence for the conditional distribution of (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$ will be too poor to be useful.

As a special case of Corollary 1, when \mathbf{Z} is empty, we can show that whether \bar{X}_n and \bar{Y}_n are independent is also determined by the covariance matrix, or more precisely, by the value of $\text{Cov}(X, Y)$. The relation between independence and covariance is less complicated, thanks to the fact that if X and Y are independent, then $\text{Cov}(X, Y) = 0$. Basically, if $\text{Cov}(X, Y) \neq 0$, then \bar{X}_n and \bar{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 (\bar{X}_n, \bar{Y}_n) and the product of the marginal distributions of \bar{X}_n and \bar{Y}_n will converge to a non-zero value, which is the total

⁴This value is the total variation distance between the conditional distribution of (U, V) given \mathbf{W} and the product of the conditional distributions of U given \mathbf{W} and V given \mathbf{W} .

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,⁵ the total variation distance between the joint distribution of (\bar{X}_n, \bar{Y}_n) and the product of the marginal distributions of \bar{X}_n and \bar{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 (\bar{X}_n, \bar{Y}_n) and the product of the marginal distributions of \bar{X}_n and \bar{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 general independence test will likely return that \bar{X}_n and \bar{Y}_n are independent.

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 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 \mathbf{X} is a discrete distribution that only assigns non-zero probabilities to points $\mathbf{x} = (x_1, \dots, x_k)$ 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 $\mathbf{X} = (X_1, \dots, X_k)$ has a lattice distribution, and h_i is the span of the i th coordinate X_i . Then \mathbf{X} has a regular lattice distribution if, for any $1 \leq i \leq k$, there are at least two vectors $\mathbf{x}^i = (x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_k)$ and $\mathbf{y}^i = (x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_k)$, such that $|y_i - x_i| = h_i$, $P(\mathbf{X} = \mathbf{x}^i) > 0$, and $P(\mathbf{X} = \mathbf{y}^i) > 0$.*

Let $\phi(\mathbf{t})$ be the characteristic function of \mathbf{X} , define $T = \{\mathbf{t} : |\phi(\mathbf{t})| = 1, \mathbf{t} \neq \mathbf{0}\}$. By Lemma 1, $|\phi(\mathbf{t})| = 1$ implies that $\mathbf{t}^T \mathbf{X} = b + 2m\pi$ a.s. for $m = 0, \pm 1, \dots$. In particular, $\mathbf{t}^T (\mathbf{x}^i - \mathbf{y}^i) = t_i(y_i - x_i) = 2m_1\pi$ 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 \mathbf{X} has a regular lattice distribution, $|\phi(\mathbf{t})| < 1$ 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 \mathbf{X}_n be i.i.d. discrete random vectors with a lattice distribution that satisfies the regularity condition given above. Suppose \mathbf{X}_n has mean $\mathbf{0}$ and a non-singular covariance matrix Σ_X . Let h_i be the span of the marginal distribution of the i th coordinate of \mathbf{X}_n , $\phi(\mathbf{t})$ be the characteristic function of \mathbf{X}_n , and $\bar{\mathbf{X}}_n = \sum_{i=1}^n \mathbf{X}_i / \sqrt{n}$. Then the probability mass functions $p_n(\mathbf{x})$ of $\bar{\mathbf{X}}_n$ converge uniformly*

⁵This 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, \dots, X_n) and (Y_1, \dots, Y_n) are also independent, hence $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n Y_i$ are independent.

to the density g of a multivariate normal distribution with 0 mean and covariance matrix Σ_X in the following way:

$$\sup_x \left[\frac{n^{k/2}}{\prod_{i=1}^k h_i} p_n(\mathbf{x}) - g(\mathbf{x}) \right] \rightarrow 0 \quad (3)$$

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

5 Estimate correlation matrix from noisy aggregation data

In section 3 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 4, 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 i th cell, and S and T the counts of tags for A and B in the resulting SAGE library. Suppose that $E[X_i] = E[Y_i] = 15$, $\text{Var}(X_i) = \text{Var}(Y_i) = 225$, and $\text{Cov}(X_i, Y_i) = 112.5$, (hence $\text{Corr}(X_i, Y_i) = 0.5$). 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: ⁶

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

Therefore, we have:

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

which implies that $\text{Corr}(S, T) \approx -8.5 \times 10^{-4}$. On the other hand, if we assume that $\text{Cov}(X_i, Y_i) = 0$, and everything else remains the same, the correlation between S and T would be -1×10^{-3} . Thus to test the null hypothesis that $\text{Corr}(X_i, Y_i) = 0$ versus the alternative that $\text{Corr}(X_i, Y_i) = 0.5$, we have to test $\text{Corr}(S, T) = -1 \times 10^{-3}$ versus $\text{Corr}(S, T) = -8.5 \times 10^{-4}$. 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%. ⁷ 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 $\text{Corr}(X_i, Y_i) = 0$ or $\text{Corr}(X_i, Y_i) = 0.5$, 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 $\text{Corr}(X_i, Y_i) \neq 0$.

⁶For the details of the proof, see Chapter 3.

⁷Let z be the test statistic. Under the null, $E_0[z] \approx -0.001$, under the alternative, $E_1[z] \approx -0.00085$. The variance of z is approximately $1/(n - 3)$, where n is the sample size. The level 15% test will reject the null if $z > -0.000925$.

It is difficult to estimate how many microarray measurements are required so that we can test reliably whether $\text{Corr}(X_i, Y_i) = 0$, 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 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 $\text{Corr}(X_i, Y_i) = 0$.

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.

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.

Acknowledgment

I would like to thank Prof. Clark Glymour for first pointing to me the problem of aggregation and giving helpful suggestions on this paper. This paper is supported in part by NASA grant NCC2-1227.

6 Appendix: Proofs

Lemmas 1, 2 and 3 are multivariate versions of some well known facts about the univariate characteristic functions.

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

$$f_n(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^k \int \exp(-i\mathbf{t}^T \mathbf{x}) \phi(\mathbf{t}/\sqrt{n})^n dt \quad (4)$$

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

$$\int \left| \exp(-i\mathbf{t}^T \mathbf{x}) \left(\phi(\mathbf{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\mathbf{t}^T \Sigma_X \mathbf{t}) \right) \right| dt \rightarrow 0 \quad (5)$$

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

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

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

$$\int_{|\mathbf{t}| \leq c} \left| \exp(-i\mathbf{t}^T \mathbf{x}) \left(\phi(\mathbf{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\mathbf{t}^T \Sigma_X \mathbf{t}) \right) \right| dt < \epsilon \quad (6)$$

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

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

□

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

$$\sup_{\mathbf{x}} \left[\frac{n^{k/2}}{\prod_{i=1}^k h_i} p_n(\mathbf{x}) - g(\mathbf{x}) \right] \rightarrow 0 \quad (7)$$

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

$$p_n(\mathbf{x}) = \frac{\prod_{i=1}^k h_i}{(2\sqrt{n}\pi)^k} \int_{-\frac{\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{-\frac{\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \exp(-i\mathbf{t}^T \mathbf{x}) \phi(\mathbf{t}/\sqrt{n})^n dt \quad (8)$$

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

$$\int_{-\frac{\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{-\frac{\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \exp(-i\mathbf{t}^T \mathbf{x}) [\phi(\mathbf{t}/\sqrt{n})^n dt - \int \exp(-\frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma}_X \mathbf{t}) dt] \rightarrow 0 \quad (9)$$

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

$$\int_{-\frac{\sqrt{n}\pi}{h_k}}^{\frac{\sqrt{n}\pi}{h_k}} \cdots \int_{-\frac{\sqrt{n}\pi}{h_1}}^{\frac{\sqrt{n}\pi}{h_1}} \left| \phi(\mathbf{t}/\sqrt{n})^n - \exp(-\frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma}_X \mathbf{t}) \right| dt \rightarrow 0 \quad (10)$$

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, $|\phi(\mathbf{t})| < 1$ if $0 < |t_i| < h_i$ for all $1 \leq i \leq k$. □

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

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

Let $F_{X,Y|\mathbf{Z}}^n$, $F_{X|\mathbf{Z}}^n$, and $F_{Y|\mathbf{Z}}^n$ be the conditional distributions of (\bar{X}_n, \bar{Y}_n) given $\bar{\mathbf{Z}}_n$, \bar{X}_n given $\bar{\mathbf{Z}}_n$, and \bar{Y}_n given $\bar{\mathbf{Z}}_n$ respectively. Clearly they are also lattice distributions. We are going to approximate these three conditional distributions with three continuous distributions $G_{X,Y|\mathbf{Z}}^n$, $G_{X|\mathbf{Z}}^n$, and $G_{Y|\mathbf{Z}}^n$. The basic idea is to transform the probability mass functions of the lattice 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 Euclidean 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:

$$\begin{aligned} g_{X,Y|\mathbf{z}}^n(x, y|\mathbf{z}) &= \frac{n}{h_1 h_2} \sum_{(m_1, m_2) \in \mathbb{Z}^2} I_{C_{m_1, m_2}^{x, y, n}}(x, y) \\ &\quad P((\bar{X}_n, \bar{Y}_n) \in C_{m_1, m_2}^{x, y, n} \mid \bar{\mathbf{Z}}_n = d_n(\mathbf{z})) \\ g_{X|\mathbf{z}}^n(x|\mathbf{z}) &= \frac{\sqrt{n}}{h_1} \sum_{m_1 \in \mathbb{Z}} I_{C_{m_1}^{x, n}}(x) P(\bar{X}_n \in C_{m_1}^{x, n} \mid \bar{\mathbf{Z}}_n = d_n(\mathbf{z})) \\ g_{Y|\mathbf{z}}^n(y|\mathbf{z}) &= \frac{\sqrt{n}}{h_2} \sum_{m_2 \in \mathbb{Z}} I_{C_{m_2}^{y, n}}(y) P(\bar{Y}_n \in C_{m_2}^{y, n} \mid \bar{\mathbf{Z}}_n = d_n(\mathbf{z})) \end{aligned}$$

where

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

and $d_n(w_3, \dots, w_{k+2}) = (v_3, \dots, v_{k+2})$, with $v_i = \lceil \text{ceil}((\sqrt{n}w_i - nc_i)/h_i - 0.5)h_i + nc_i \rceil / \sqrt{n}$ for $3 \leq i \leq k+2$.⁸

Let $p_{X,Y|Z}^n$, $p_{X|Z}^n$, and $p_{Y|Z}^n$ be the probability mass functions for $F_{X,Y|Z}^n$, $F_{X|Z}^n$, and $F_{Y|Z}^n$ respectively. Let $q_{X,Y|Z}^n = p_{X|Z}^n p_{Y|Z}^n$, and $Q_{X,Y|Z}^n$ be the corresponding distribution function. Let $h_{X,Y|Z}^n = g_{X|Z}^n g_{Y|Z}^n$, and $H_{X,Y|Z}^n$ be the corresponding distribution function. It is easy to see that the total variation of the signed measure $Q_{X,Y|Z}^n - F_{X,Y|Z}^n$ is the same as the total variation of the signed measure $H_{X,Y|Z}^n - G_{X,Y|Z}^n$, i.e., $|Q_{X,Y|Z}^n - F_{X,Y|Z}^n| = |H_{X,Y|Z}^n - G_{X,Y|Z}^n|$. As a direct consequence of Theorem 4, we have $h_{X,Y|Z}^n - g_{X,Y|Z}^n \rightarrow 0$ as $n \rightarrow \infty$, although the convergence may be not uniform. By the bounded convergence theorem, we have $|H_{X,Y|Z}^n - G_{X,Y|Z}^n| \rightarrow 0$, hence $|Q_{X,Y|Z}^n - F_{X,Y|Z}^n| \rightarrow 0$, as $n \rightarrow \infty$. \square

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 Microarray Images", *Journal of Biomedical Optics* *2(4)*: pp. 364-374
- Chu, T. (2002), "Sampling, Amplifying, and Resampling", *Tech Report*.
- Chu, T., Glymour, G., Scheines, R., and Spirtes, P. (2003) "A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarray," *Bioinformatics* *19(9)*; pp. 1147-1152.
- Danks D., & Glymour, C., (2001), "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. (2003), "The Computational and Experimental Complexity of Gene Perturbations for Regulatory Network Search," *Proceedings of IJCAI-2003 Workshop on Learning Graphical Models for Computational Genomics*: pp. 22-31.
- 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.

⁸ $\text{ceil}(x)$ returns the smallest integer that is greater than or equal to x . Thus $d_n(\mathbf{z})$ returns the lattice point of $\bar{\mathbf{Z}}_n$ closest to \mathbf{z} . In cases of tie, it will return the smallest. This way we have defined the conditional distribution for the cases where $P(\bar{\mathbf{Z}}_n = \mathbf{z}) = 0$.

- 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
- Pearl, J. (2000) *Causality*, Cambridge, UK: Cambridge University Press.
- 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*,

- 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 Stabilized 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.