uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Shenggang Li
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Mathematics)
GRADE / DEGREE

Department of Mathematics and Statistics
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Stem Cell Differentiation Study: Statistical Analysis and Application to Microarray Data

TITRE DE LA THÈSE / TITLE OF THESIS

Sankoff David
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Mayer Alvo

Miklós Csűrös

Stéphane Aris-Brosou

Sanjoy Sinha

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Stem Cell Differentiation Study: Statistical Analysis and Application to Microarray Data

Shenggang Li

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1] The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-48658-0
Our file Notre référence
ISBN: 978-0-494-48658-0

Abstract

This thesis has two main aims:

1. To develop new statistical and data mining procedures for modeling stem cell genes during the differentiation process,

2. To present biological interpretations of "stemness" genes based on time-course expression patterns and genetic networks.

Stem cells have been of much interest over the past few years. Up to now, scientists have recognized many crucial features that distinguish stem cells from other types of cells: first, stem cells are capable of renewing themselves, serving as the organism's repair system. Second, they can differentiate into many types of adult cells that finally form over 200 kinds of specialized cells in the body.

Given the vast information in gene expression data, it is possible to develop simple and reliable methods to investigate the differentiation of stem cells. However, very little is known about the mechanisms that characterize the genome-wide dynamic regulation of gene expression controlling stem cell differentiation. Relatively few research efforts have tackled stem cell differentiation problems using statistical procedures on time-course expression data. There is no doubt that understanding the time course of differentiation is of great significance since it should enable scientists to artificially induce stem cells to generate the appropriate intended specialized cells, such as the beating cells of the heart muscle or the insulin-producing cells of the pancreas.

In this thesis, we approach the time-course differentiation of stem cells, as measured by expression data, through some new statistical and data mining procedures. Our framework can be outlined as following:

- Develop new detection methods for differential gene expression to identify "stemness" genes that drive and repress stem cell differentiation.
- Develop time-course expression clustering analysis techniques and apply them to stem cells.
- Explore new efficient analyzes for generalized linear models (GLM) that can be potentially applied in gene interaction network.

Validations of computational performance are based on the experimental data available from StemBase, a well-cited public collection of DNA microarray data on stem cells designed for the purpose of facilitating stem cell research.

Key words: stem cells, differentiation, embryonic stem cells (ESC), microarray, gene expression, clustering, gene regulatory network, time course, differential genes, Gene Ontology(GO), similarity measures, quasi-likelihood, mixture model

Acknowledgements

I would like to sincerely appreciate my respected supervisor Professor David Sankoff for giving me so much freedom to explore and discover new field in stem cell microarray analysis over the years during my Ph.D. program as well as the critical reviews and suggestions on my papers, this thesis and additional research directions.

I am very grateful to Professor Miguel Andrade for his close verification on my interpretations of the gene ontology and search results. His excellent work and rich experience in stem cell research have guided one of my best learning experiences. I would say I could hardly have finished my thesis without his contribution.

I would like to thank my other thesis proposal committee members. They have really been very supportive. Professor Alvo Mayer, Sanjoy K. Sinha, Stehane Aris-Brosou, and Miklos Csuros brought much useful suggestions on my work, and kept me honest about all the details, which enlightened me to reach the best solution. I also deeply thank other committee members for their careful reviews and invaluable suggestions on my thesis.

Finally I would like to thank my classmates Zhenyu Yang, Wei Xu, Chunfang Zheng and Huiling Xong for their friendship and encouragement during my university life.

Dedication

To my dear wife Yanchen for her love and devoted support and to my lovely newborn son Yichuan for giving me new dreams to pursue

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Stem cells have the potential to develop into many different types of cells in the body. An undifferentiated stem cell will go through mitotic cell division and differentiate into various kinds of specialized offspring cells such as heart, lung, muscle, skin and brain cells. Stem cells can be categorized in terms of differentiation stage:

- Totipotent cells are capable of forming every type of body cell through replicating and differentiating themselves,
- Pluripotent stem cells may give rise to any type of cell in the body except those needed to develop a fetus,
- Multipotent stem cells will produce only cells in some family associated cells of cells, for instance, the hematopoietic stem cell can only differentiate into a blood cell of one type or another,
- Unipotent stem cells can produce only one cell type in adult organisms, but can also self-renew.

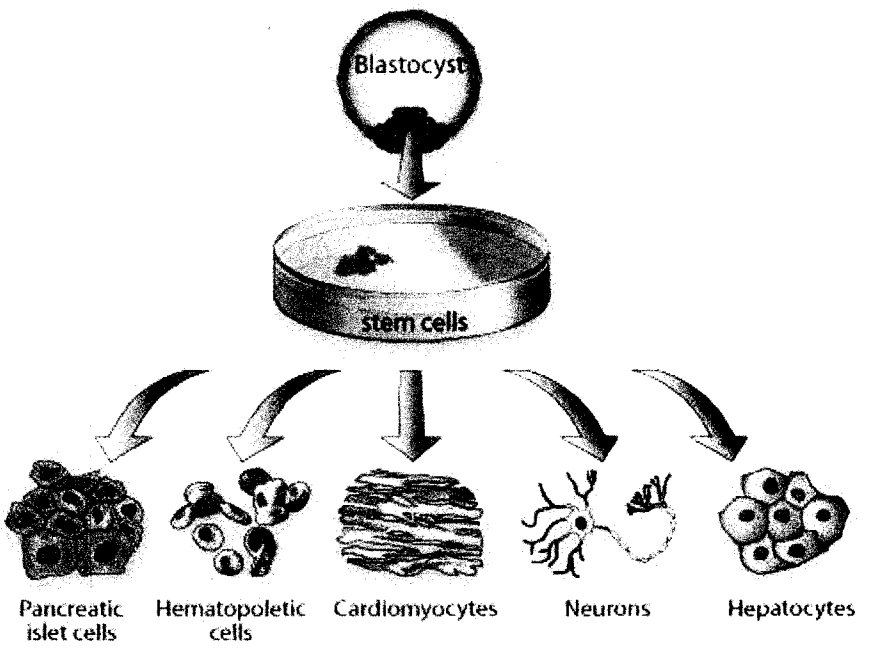

Figure 1.1: Embryonic Stem Cells and Derived Cells

Because of its special biological functions, the stem cell has earned great popularity in many therapies for replacing disordered or damaged cells resulting from various types. Recent studies show that ESC (embryonic stem cells) can be induced to differentiate into cells with multiple functions by adding specific growth factors. It has been reported recently that successful hematopoietic stem cell transplantation has achieved this potential of stem cell treatment. For instance, scientists have infused stem cells into the patient's bone marrow to produce new blood cells [Barker JN. et al. 2005].

There are two major types of stem cells in terms of development period: embryonic and adult stem cells (also namely tissue stem cells). Embryonic stem cells derive from three to five days old embryo developing from fertilized eggs. The embryos used in these studies were created for infertility purposes through *in vitro* fertilization procedures. The embryos develops into a hollow structure ball of cells called the blastocyst, which consists of three layers: trophoblast, blastocoel and inner cell

mass, from which Embryonic stem cells are generated. Adult stem cells are undifferentiated cells found in differentiated multi-cellular organisms. They are said to show plasticity because they may be reprogrammed by growth factors to change their differentiation direction which has important implications for cell therapy and other regenerative treatments. Given suitable conditions, scientists are able to manipulate adult stem cells to introduce new genes or differentiate into specialized cells. For example, in mice, stem cells can be directly generated from adult fibroblast cultures [Takahashi K. et al. 2006].

Research on human adult stem cells has attracted recently a great deal of excitement. In fact, adult blood stem cells from bone marrow have been utilized in human body transplants for many years. Scientists have used adult human neural stem cells to successfully regenerate damaged spinal cord tissue and to improve mobility in mice [UC Irvine Reeve-Irvine Research Center, 2005]. The discovery provided strong evidence to support the plasticity property of human adult stem cells. Unlike embryonic stem cells, which originate from the inner cell mass of blastocyst, it is not very clear how the adult stem cells in mature tissues are created.

Another type of mammalian stem cell, called cord blood stem cell can be extracted from the umbilical cord when it is detached from the newborn baby. Similar to hematopoietic stem cells, cord blood stem cells can also be utilized in stem cell treatment [Raymer E. 2005].

1.2 Time-course Microarray data

Microarrays are arrays of DNA molecules that permit many hybridization experiments to be performed at the same time. They can monitor expression levels of thousands of genes simultaneously.

Genetic information is contained in DNA which is a template for a class of cellular ribonucleic acids (RNAs) called messenger ribonucleic acids (mRNAs). By measuring

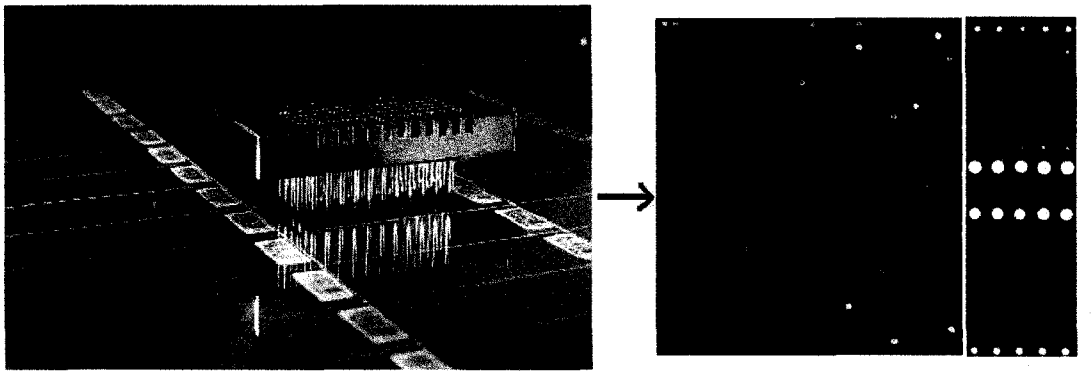

Figure 1.2: Microarray technology and gene expression

levels of mRNAs we can infer which specific proteins the cell is making.

The most popularly used technology for evaluating amount of mRNA being synthesized is through hybridization of nucleic acids with an array of many cDNA sequences or oligonucleotides, printed as spots onto a solid support. The intensity of the label associated with each spot represents the expression level for that particular sequence. This intensity is compared to the intensity of the equivalent spot on an identical gridded array hybridized with the material prepared using mRNA isolated from a different source. This provides a measure of differential mRNA expression specific for that spot DNA sequence (e.g. gene marker). By comparing all equivalent spot intensities between two grids, the changes in expression of many thousands of genes can be evaluated at the same time (see Fig 1.2).

The oligonucleotide differs from cDNA on different material; cDNA uses PCR results created from cDNA clones, while oligonucleotide is produced from synthetic gene-specific oligo-deoxynucleotides.

Microarray chips have attracted much attention recently with the increasing availability of large-scale data and the urgent need to understand gene regulatory pathways. This requires effective identification of co-expressed genes and other co-

herent patterns of gene expression. Data mining on microarray data is an important path to understand gene regulatory networks [Wahde M. et al. 2001].

There are two kinds of microarray data: static and time course expression. Static data are the gene expression levels observed on microarray chips at a fixed time point or under specific experimental conditions or in specific tissues. For instance, the transcriptional response of human cells to ionizing radiation was measured by comparing expression data under different experimental conditions [Tusher VG. et al. 2001]. Time-course gene expression is sequential gene expression recorded over a time course, for example, the response of human fibroblasts to serum [Iyer VR. et al. 1999] and gene expression of ESC over 11 time points [Perez-Iratxeta C. et al. 2005]. We often refer to the time sequential expressions of a gene as a gene profile, for example,

$$(1025.67, 1349.29, 2001.05, 1888.53, 999.12)$$

might be a gene profile across 5 time points.

Time-course transcriptional profiles of stem cells provide us with a dynamical measurement of the transcriptional levels of thousands of genes. Underlying dynamical profile is a biologically continuous process, which can be mathematically represented by a continuous function, i.e. a continuous curve. Such curves can reflect functional forms through expression patterns. For instance, Figure 1.3 shows time course gene expression of mouse stem cells in STAT3/JAK signalling Pathway.

From time-course expression data, we may gain insights into the regulatory network of genomes and understand many fundamental mechanisms of cells at the molecular or cellular mechanism level. For instance, given dynamics of stem cell gene expression over a time course, it might be possible to infer the regulatory mechanisms that drive or repress the differentiation of stem cells. Indeed, by comparing the gene expression levels with different time points, investigators can define a group of "stemness" genes [Ramalho-Santos M. et al. 2002], which control stem cell differentiation.

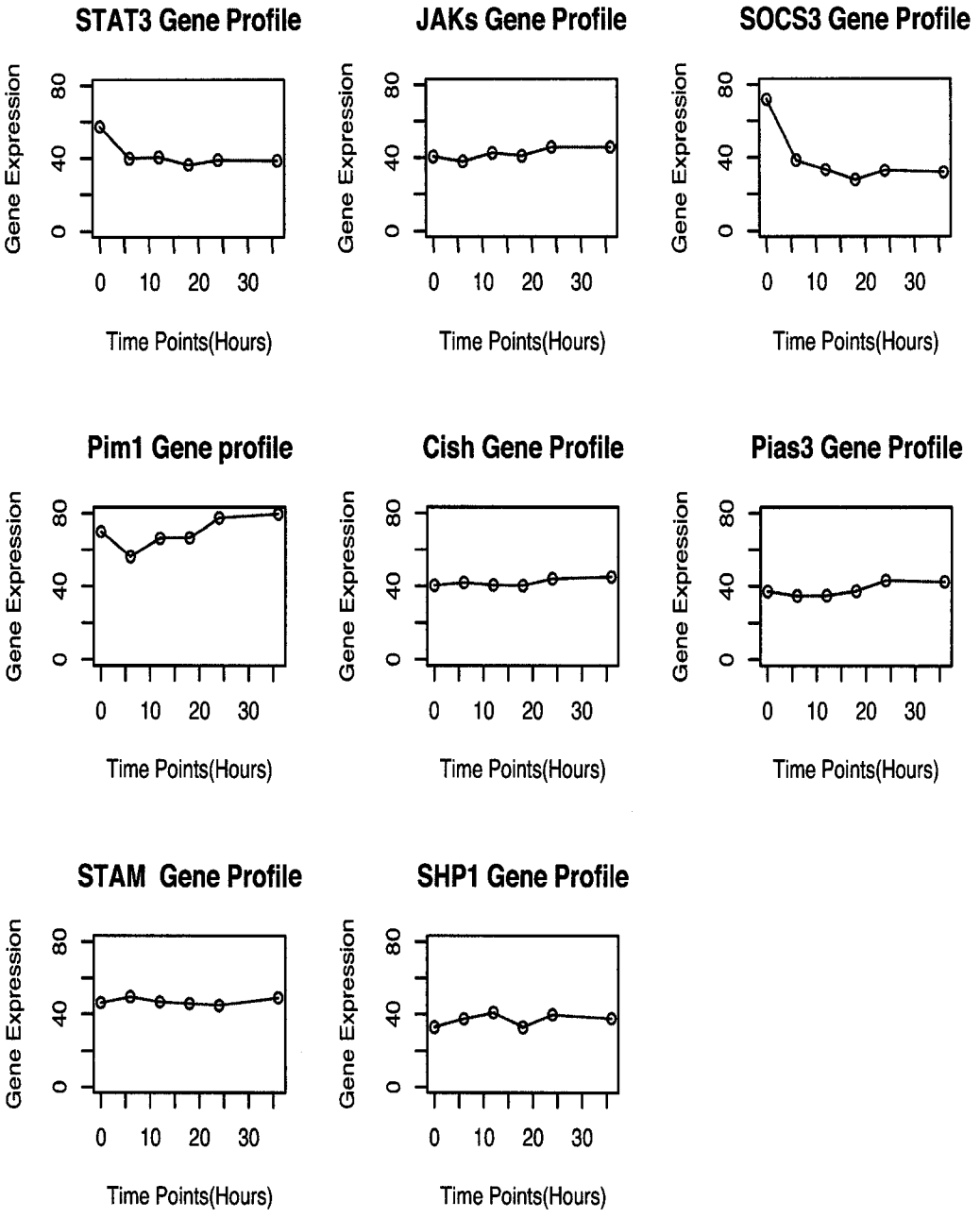

Figure 1.3: STAT3/JAK signalling pathway: Gene profiles within the first 36 hours. Drawn from the data extracted from StemBase [Perez-Iratxeta C. et al. 2005]

Time-course gene expression has the advantage that it includes autocorrelated information across time points. By comparison, static expression data only give a snapshot of information under a fixed condition. This has drawbacks for gene expression research. For example, the existence of coherent trajectory patterns can be interpreted as reflecting related biological functions performing genes in the same cluster, but static expression data could not provide this information. Using static stem cell expression information, [Takahashi K, et al. 2006] summarized a total number of 216 genes that are enriched in stem cells. However, it makes more sense to investigate "stemness" genes differentiation process based on time course expression.

1.3 Data source

In this dissertation, the data for studying stem cell differentiation come from StemBase, a collection of DNA microarray data on stem cells designed for the purpose of scientific research on the topic [Perez-Iratxeta C. et al. 2005]. Up to now, the StemBase has stored many samples including ESC(embryonic stem cells) for mouse and human hematopoietic stem cells. The database was organized by collecting gene expression data from various microarray experiments. All gene expression data were probe set hybridization values generated by the MASS5.0 (a kind of normalization algorithm for gene expression levels) of Affymetrix. The high density of probe sets reflects a huge number of genes and ESTs (expressed sequence tags), as well as predicted exons. For example, E113 is a time course study on V6.5 ESC, where, the probe set expression is measured at 11 time points that span 14 days when LIF (LEUKAEMIA inhibitory factor) is removed. The experimental data are gene or gene product expression levels across the time points: 0 hour, 6 hours, 12 hours, 18 hours, 24 hours, 36 hours, 48 hours, 4 days, 7 days, 9 days, and 14 days. Based on this dynamical information above, we can investigate stem cell properties through advanced statistical analysis and data mining techniques.

1.4 Statistical analysis on stem cells

The thesis contributes three quantitative techniques to explore time course gene expression data:

1. Statistical detection of differential gene expression,
2. Dynamical clustering analysis,
3. Quantitative methods for gene regulatory networks.

Traditional statistical methods study differences via the t statistic defined as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1.4.1}$$

where, $\bar{x}_i$, $(i = 1, 2)$ are the sample means of two groups (with sample sizes n_1 and n_2) and the pooled sample variance

$$s_p^2 = \frac{n_1 s_{x_1}^2 + n_2 s_{x_2}^2}{n_1 + n_2 - 2} \tag{1.4.2}$$

The t test is often employed to deal with sample comparison problems in microarray gene expression data. In these data, the number of genes on a single chip is typically very large (e.g., 20,000) compared with the very small number of replicates or experimental conditions for each gene (3 or 4). This leads to serious inaccuracies in the t test since the t statistic for the gene tends to be higher when its sample variance is smaller.

This effect is especially troubling for the general t or Hotelling T^2 test, because low standard errors often appear purely by chance. To illustrate this, the two plots in Figure 1.4 depict the strong dependence between the t statistic and the pooled standard error (3 replicates) (left) and between the mean difference and the standard error (right), in the comparison of gene expression values in murine J1 Embryonic

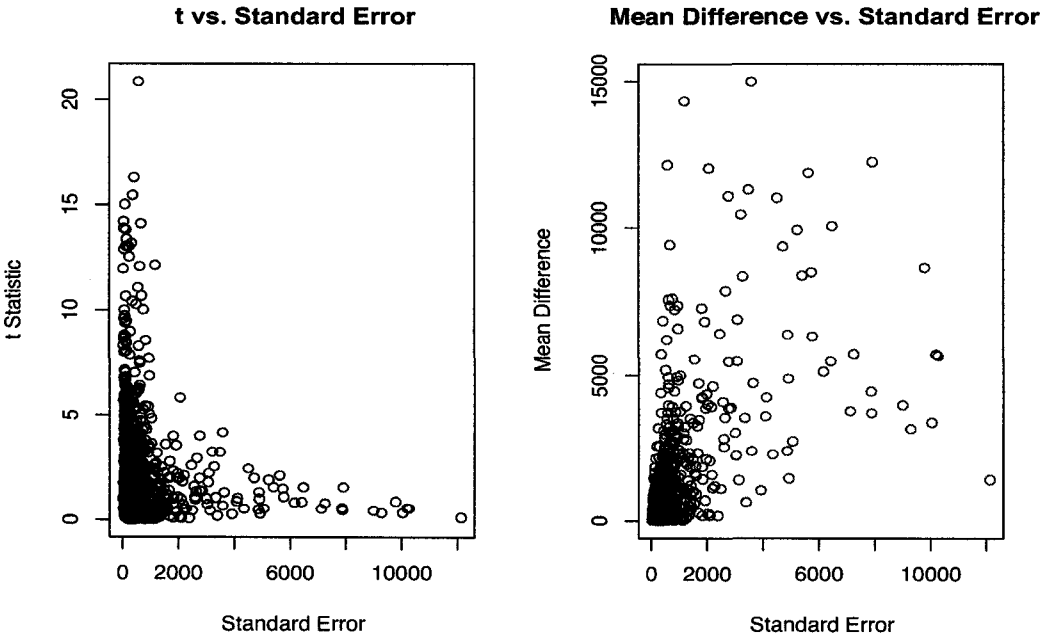

Figure 1.4: Scatter plot of pooled standard error against t-statistic and mean difference

Stem Cells (ESCs) at time point 0 hour and 18 hours respectively of a differentiation experiment [Perez-Iratxeta C. et al. 2005].

The microarray data analysis literature contains several proposals for tackling this problem [Tusher VG. et al. 2001]. The best-known solution is SAM (Significant Analysis of Microarrays) in which the standard error s_g is adjusted by adding a constant fudge factor c. In this way, the dependence of the test statistic on sample variance can be reduced, as long as the fudge factor c is carefully chosen:

$$t_g(c) = \frac{|\bar{x_g} - \bar{y_g}|}{s_g + c} \tag{1.4.3}$$

Where, $\bar{x_g}, \bar{y_g}$ represent the sample means of gene expression levels under different experimental conditions. An advantage of the available implementation of SAM is that it avoids the assumption of sample normality, since the distribution quantiles of the test statistic are calculated via a permutation procedure. This method is widely applied in differential genes investigation. For instance, [Takahashi K. et al. 2006] employed the SAM method to study "stemness" signature by comparing gene expression levels at different experimental conditions.

The analysis of gene expression data presents us with a potential insight into gene function and regulatory mechanisms. A crucial approach is the detection of groups of genes that manifest similar expression patterns. Clustering analysis of gene expression data is often applied to detect co-regulated genetic functions. The classical clustering techniques include optimizing within-between variation methods such as K-means [Hartigan JA. et al. 1979], linkage agglomerative clustering [Ball GH. et al. 1967] and model-based clustering [Chudova D. et al. 2004]. In the K-means method, the number of clusters K is given, and each gene is assigned to one of K clusters through minimizing a measure of dispersion within the clusters. The goal is to divide the sample into a predetermined number K of non-overlapping clusters so that clusters are as homogeneous as possible with respect to the measurements. The K-means algorithm

starts with an initial partition of the observations into K clusters. Subsequent steps modify the partition to reduce the sum of the distance of each data point from its centroid. This leads to a new partition for which the sum of distances is smaller than before. The means of the new clusters are computed and the improvement step is repeated until all the positions of all the centroids become stable. The K-means clustering algorithm:

($k1$) Start with K initial clusters.
($k2$) At every step, each observation is reassigned to the cluster with the "closest" centroid.
($k3$) Recompute the centroids of clusters that have lost or gained a record, and repeat step ($k2$).
($k4$) Stop when moving any more records between clusters increases cluster dispersion.

The idea behind hierarchical linkage clustering is to start with each cluster containing one observation and then progressively agglomerating the two nearest clusters until there is just one cluster left in the end. At final stage, the cluster will include all observations. The hierarchical linkage algorithm:

($h1$) Start with n clusters.
($h2$) The two closest observations are merged into one cluster.
($h3$) At each step, the two clusters with the smallest distance are merged. In other word, single observations are added to existing clusters or two merged clusters.

The hierarchical linkages algorithm can be classified into several sub-algorithms such as single linkage, complete linkage, weighted or unweighted average linkage, median linkage, centroid linkage and Ward linkage. The various clustering methods differ in how the distance between two clusters is computed when merging sub-clusters.

[Johnson RA. et al. 2002]

Both K-means and linkage clustering are pure data-mining methods without statistical background. The common shortcoming is the failure to consider variable correlations and data density. In general, these methods do work well in promptly clustering dynamical gene expressions.

Model-based clustering is a relatively new technique based on sample distribution. Rather than defining a similarity measure such as Euclidean distance or Pearson correlation, model-based clustering methods assume that observed data are generated by an unknown number of probabilistic models. A model with the same parameter set creates a cluster. The clustering proceeds by maximizing the likelihood that observed data are generated by the given models. Cluster membership is determined by posterior probabilities that a gene expression profile is generated by a specific model. Model-based methods handle the gene expression clustering problem based on the distribution of data. This method also takes the correlation structure of observations into account. Model-based clustering can be described by the following procedure: Let the joint probability for a data object x and a cluster y be $P(x;y)$. We maximize the following expected log-likelihood:

$$L = \sum_{(x,y)} P(x,y) log[p(x|\lambda_y)] = \sum_{x} P(x) \sum_{y} P(y|x) log[p(x|\lambda_y)] \tag{1.4.4}$$

where, $P(y|x)$ is the posterior probability of y given x, $p(x|\lambda_y)$ is the incomplete data likelihood representing the distribution of x given cluster y, λ_y) is the membership indicator representing whether an observation belongs to cluster y, and $P(x)$ is the prior of data x. In practice, one typically uses the sample average to calculate L, i.e., $P(x) = 1/N$. As for model-based clustering approaches, the model type is often specified a priori, such as by Gaussian or hidden Markov models (HMMs). The model structure (e.g., the number of hidden states in an HMM) can be determined by model selection techniques and parameters estimated using maximum likelihood algorithms,

such as the EM algorithm. Finally, each gene is assigned into a cluster if its posterior probability being in this cluster is the largest.

These traditional clustering methods are often applied to static data set. Although such procedures are sometimes biologically meaningful and can exhibit gene function under some situations, the dynamic evolution of genes is better represented by gene profiles over time course. Thus, the similarity in shape between two expression curves is a key measurement in clustering analysis. The traditional clustering techniques generally fail to consider the time sequence effect, since one can always exchange the order of observations over a time course without affecting the final clustering result. Because of this, various dynamical clustering methods have been proposed, such as dynamic model-based clustering for time-course data [Wu FX., et al. 2005] and clustering analysis with functional mixture models [Chudova D. et al. 2004]. In general, these new techniques try to tackle the "dynamics" in the data through defining appropriate stochastic error models.

In sum, model clustering techniques treat a data group that satisfies a statistical model as a cluster. Therefore, we can develop new clustering methods by adding various statistical features (such as correlation or trend) to the model.

The detection of gene regulatory network has recently received much interest. There are many modeling techniques for identifying gene regulatory pathways. The major streams include Boolean network [Liang S. et al. 1998], differential equation model [Chen T. et al. 1999] and Bayesian network [Friedman N. et al. 2000]. These models view the genes in a genome as categorical or continuous longitudinal random variables that interact with each other in terms of gene expression levels and genetic regulatory relationships. In differential equation models, the gene expressions are assumed to be determined by the amount of transcription protein and degradation rates of mRNA. i.e., the model is based on the following equation system [D'haeseleer P. et al. 1999] and [Chen T. et al. 1999]:

$$Y(t+\Delta t) = W.Y(t)$$

where the matrix $W = \{w_{ij}\}_{n\times n}$ is regulatory matrix for $Y_i(t+\Delta t)$, a vector representing the expression levels of genes $g_1, g_2, ..g_n$ at time $t+\Delta t$. The model assumes that gene expression is determined by transcriptional protein and mRNA degradation, in other words, the expression level of the current step is determined by expression levels of other gene at previous steps.

The classical Boolean networks were proposed by [Kauffman SA. 1969], as random models of genetic regulatory networks. In a Boolean network model, the genes are in one of two states, either active or inactive, described as completely on or off. The two states are often denoted by the binary values 1 and 0 respectively. According to this model, the gene network is represented as an oriented graph $G = (V, F)$, whose nodes V represent elements of the network, and F defines a topology of edges between the nodes and a set of boolean functions. A node represents either a gene or a biological stimulus, where a stimulus is any relevant physical or chemical factor which influences the network and is itself not a gene or gene product. The binary state varies with respect to time and depends on the states of other genes in the network through a boolean variable equation:

$$Y_i(t+1) = F_i(Y(t)) \tag{1.4.5}$$

where $Y_i(t)$, $i = 1\ldots, n$ are the states of the n genes in the network. The function $F(.)$ is a Boolean function of the states of the network at time t, for updating the state of element i at $1+t$. We also call $F(.)$ the regulator function. We can write the above equation in the matrix form:

$$Y(t+1) = F(Y(t)) \tag{1.4.6}$$

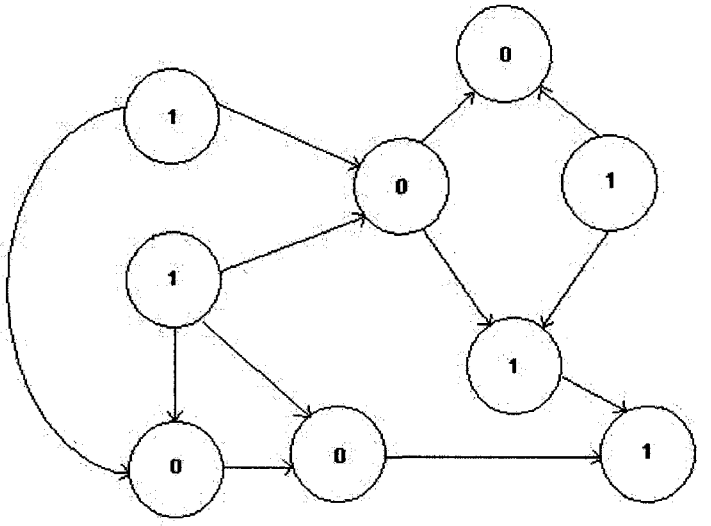

Figure 1.5: Boolean network

where $Y(t) = (Y_1(t), Y_2(t), \cdots, Y_n(t))$ is the gene expression vector at time t for n genes in the network system, and $F(.) = (F_1(.), F_2(.), \cdots, F_n(.))$ is the regulator vector for $Y(t)$ at time $t+1$.

Although Boolean network model only expresses a rough picture of gene regulatory network, it is the first proposed method that is capable of simulating gene interaction relationships.

Bayesian networks (BNs) are based on Bayes theorem, mathematically,

$$P(F|R) = \frac{P(R|F)P(F)}{P(R)} \tag{1.4.7}$$

where we update our belief in factor F given the additional reason R. The left-hand term, $P(F|R)$, is known as the posterior probability or the probability of F after considering the reason R. The term $P(F)$ is called the prior probability of F. The term $P(R|F)$ is called the likelihood and gives the probability of the evidence

assuming the hypothesis F. Finally, the last term $P(R)$ is independent of F and can be regarded as a normalizing or scaling factor.

In mathematical form, a Bayesian network for a gene system is represented by a joint density of $(x_1, x_2, \cdots, x_n)$. This system is composed of a set of local conditional densities combined with a set of conditional independence assertions that allow the construction of a global density from the local density. The chain rule of probability is to be used to ascertain these values:

$$P(x_1, x_2, \cdots, x_n|c) = \prod_i p(x_i|x_1, \cdots, c) \tag{1.4.8}$$

where c is the current state information of the system. BNs provide an efficient representation for expressing joint probability distributions (JPDs). Actually, it is a probabilistic graphical model representing a set of variables and their probabilistic independencies. In a graphical model representation, variables are seen as the nodes connected together by edges marked as relationship among variables, and directed acyclic graph in Bayesian network can be expressed as the joint distribution of the node values and be written as the product of the local distributions of each node and its parents. BNs can provide a parsimonious representation of conditionality among parametric relationships, and the directed edges of the graph are connected to show this dependency.

Bayesian networks have the following advantages over traditional methods of inferring interactive relationships:

- Independent relationships among variables are easy to understand and conditional relationships are identified by directed graph edges,
- The variables are independent if there is no direct path from one to another. By optimizing the graph, we can find that every node has at most k parents. The algorithmic routines run in $O(2^k n)$ instead of $O(2^n)$ time. In fact, the

computational procedure is of linear time (based on the number of edges) instead of exponential time (i.e. based on the number of parameters).

Another kind of gene network inference method is factor analysis, such as pathway analysis [Wright S. 1934], which learns the traits or patterns of independent variables and their correlations with the response variable. The goal of the method is to provide plausible explanations of observed correlations based on a path diagram and the decomposition of observed correlations into path coefficient terms by multiple regression procedure. The method was initially developed to identify causal relations in population genetics. It is a kind of structural equation analysis that does not consider previous time points or time order. However, given enough gene expression samples over a time course, we can model the relations among transcript levels for a set of genes and unravel the strength of casual relations from patterns of variation on transcript levels. The model is formulated by:

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \varepsilon \tag{1.4.9}$$

where y is a random variable that is thought to cause change in other exogenous random variables x_j $(j = 1, 2, \cdots, m)$, which are assumed to be mutually independent. ε is an unmeasured random variable(residual effect). The path diagram of the model can be represented as Figure 1.6(when $m = 3$):
(1.4.9) can be standardized such that y and x_j $(j = 1, 2, \cdots, m)$ have variance 1, i.e.

$$\frac{y - E(y)}{\sigma(y)} = \sum_{j=1}^{m} \frac{\sigma(x_j)}{\sigma(y)} \frac{x_j - E(x_j)}{\sigma(x_j)} + \frac{\sigma(\varepsilon)}{\sigma(y)} \frac{\varepsilon}{\sigma(\varepsilon)} \tag{1.4.10}$$

It is noted that the constant part β_0 disappears in (1.4.10), i.e. the model is only concerned with the proportion of regulatory strength for each independent gene.

$$y^{(s)} = \sum_{j=1}^{m} p_{yj} x_j^{(s)} + p_\varepsilon \varepsilon^{(s)} \tag{1.4.11}$$

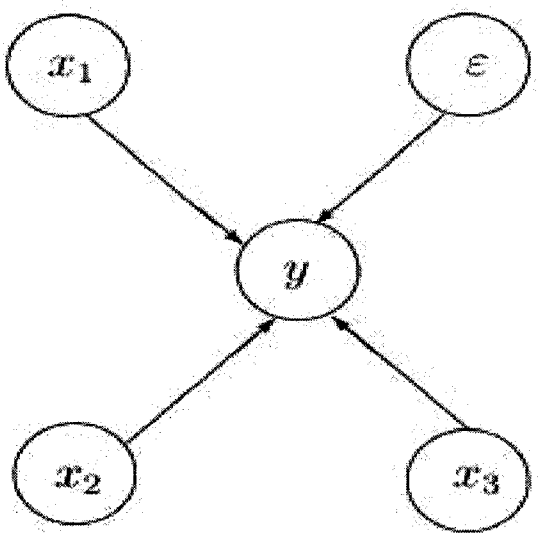

Figure 1.6: Pathway Diagram

where, $x_j^{(s)}$,$y^{(s)}$ $(j = 1, \ldots, m)$ and $\varepsilon^{(s)}$ are standardized versions of x_j,y $(j = 1 \ldots, m)$ and ε respectively. We can estimate the coefficients in the formula from the following procedure:

$$\rho_{yk} = Corr(y^{(s)}, x_k^{(s)}) = Cov(\sum_{j=1}^{m} p_{yj} x_j^{(s)}, x_k^{(s)}) = \sum_{j=1}^{m} p_{yj}\rho_{jk} \tag{1.4.12}$$

where $\rho_{jk} = Cov(x_j^{(s)}, x_k^{(s)})$. Denoting $\rho_{yx} = [\rho_{y1}, \rho_{y2}, \cdots, \rho_{ym}]^T \in \Re^m$, $P_y = [p_{y1}, p_{y2}, ..p_{ym}]^T \in \Re^m$ and $\varrho_{xx} = \{\rho_{ij}\} \in \Re^{m \times m}$. From above, we have

$$P_y = \varrho_{xx}^{-1}\rho_{yx} \tag{1.4.13}$$

To determine the residual coefficient p_ε, we notice that

$$1 = Var(y) = Var(\Sigma_{j=1}^{m} p_{yj} + p_\varepsilon\varepsilon) = \rho_{yx}^T \varrho_{xx}\rho_{yx} + p_\varepsilon\varepsilon \tag{1.4.14}$$

then we have

$$p_\varepsilon^2 = 1 - \rho_{yx}^T \varrho_{zz} \rho_{yx} = 1 - \rho_{yx}^T P_y \tag{1.4.15}$$

The pathway analysis has been applied or extended to study covariation among a set of genes in gene regulatory network [Coffman CJ. et al. 2005].

1.5 Overview of contributions

This dissertation focuses on developing new efficient statistical methods for dynamical gene expression clustering, the detection of differentially expressed a genes over a time course and inferring gene regulatory relationships. Given the computational results in each case, we search on GO(Gene Ontology) terms and public online software FatiGO to interpret specific biological functions of the proposed mESC signatures. The main contributions include:

- The SAM (Significant Analysis of Microarrays) statistic is an effective measurement for comparing small groups of samples of gene expression data. Chapter 2 extends this technique by incorporating an adaptive component to the "fudge factor" used in SAM to achieve a greater shrinkage effect. We also develop a multivariate adaptive statistic for comparing time course samples based on the Hotelling T^2 distribution. Bootstrap methods are applied to determine the adaptive parameter, p-values and false discovery rate (FDR). We carry out computational analysis of time-course gene expression data on embryonic stem cell differentiation to illustrate the advantages of the adaptive statistic,

- In Chapter 3 we address the general problem of using cluster analysis to identify co-regulated genes over a time course. We argue that rather than the Pearson correlation coefficient, similarity measures defined in terms of dynamical correlation are most appropriate for clustering time-course gene expression data.

We back up our argument with simulations and then apply this measure to clustering embryonic stem cells time-course data, verifying the plausibility of the results with gene ontology. These tests show that dynamical correlation is far superior to Pearson correlation for time-course data,

- Based on the trajectories of individual genes, in Chapter 4 we develop a new clustering method for time course gene expression data for ESC differentiation. We propose a class of functions determined by several parameters but flexible enough to model realistic time courses. This serves as a basis for a mixed model clustering method. This method takes into account

  1. Genetic function profile induced or controlled by other regulators,
  2. Unobservable random effects producing heterogeneity within gene clusters, and
  3. Autoregressive components defining the stochastic and autocorrelation structures.

  The EM algorithm is employed to fit the mixture model and clustering follows monitoring via Bayesian posterior probabilities. This method is applied to a mouse ESC line during the first 24 hours of differentiation period. We also assess the biological credibility of the results by detecting significantly associated Gene Ontology terms for each cluster,

- Since the Generalized Linear Model has potential significance in the estimation of gene regulatory networks, in Chapter 5, a BAQL (Bayesian Adjusted Quasi-likelihood) approach is developed to fit Generalized Linear Models (GLMs) by a Bayesian Adjusted Quasi-Likelihood (BAQL) approach, in which the sample mean is treated as a random variable. We define two types of adjusted quasi-likelihood functions using the Newton-Raphson iterative procedure: (1) Adjusted Quasi-Likelihood Function (A-BAQL) and (2) Two-Stage Adjusted

Quasi-Likelihood Function (T-BAQL). Computational tests show the BAQL approach is more competitive in terms of flexibility, estimate bias and prediction accuracy compared to the conventional quasi-likelihood method. We also address related issues such as density estimation, choice of bandwidth and asymptotic property of estimates.

Chapter 2

Adaptive SAM statistic in differential gene expression detection

This chapter is based on the paper submitted to Statistical Applications in Genetics and Molecular Biology. I was responsible for the formulation of all of the mathematical problems and their solutions in this work. I also wrote a complete draft of this chapter. Professor Miguel Andrade verified and expanded upon my interpretations of the gene ontology search results. Professor David Sankoff suggested some of the additional research directions and helped organize and reformulate the exposition.

2.1 Introduction

In chapter 1 we have discussed the problem of t test in differential gene expression detection. The microarray data analysis literature contains several proposals for tackling this problem. For example, [Tusher VG. et al. 2001] developed SAM (Significant Analysis of Microarrays) to study the ionizing radiation response. [Baldi P. et al. 2001] applies a Bayesian probabilistic framework to infer a t statistic using a variance posterior. [Efron B. et al. 2001] proposes a non-parametric empirical Bayes approach for the analysis of factorial data with high density oligonucleotide microarray data. [Lu Y. et al. 2005] considers Hotelling's T2 multivariate profiling for detecting differential expression in microarrays.

In classical SAM's approach the standard error is adjusted by adding a constant fudge factor. In this way, the dependence of the test statistic on sample variance can be reduced, as long as the fudge factor is carefully chosen. An advantage of the available implementation of SAM is that it avoids the assumption of sample normality, since the distribution quantiles of the test statistic are calculated via a permutation procedure. However, SAM has the following disadvantages:

- The shrinkage performance of the fudge factor may not be ideal, because the adjustment of all sample variances is done with a single fudge factor. Shrinkage is performed on all gene comparisons in a unified way without taking each gene's individual variance into consideration. Where, the shrinkage means the pattern change of the test statistic owing to the adjustment of standard deviation.

- Two sample groups may be dependent. For example, gene expression in a time series is highly auto-correlated, and this may introduce additional variation into the test statistic, reducing its power. This can be avoided by treating each group as a multivariate sample using Hotelling's T^2 statistic, which takes correlation effects into account.

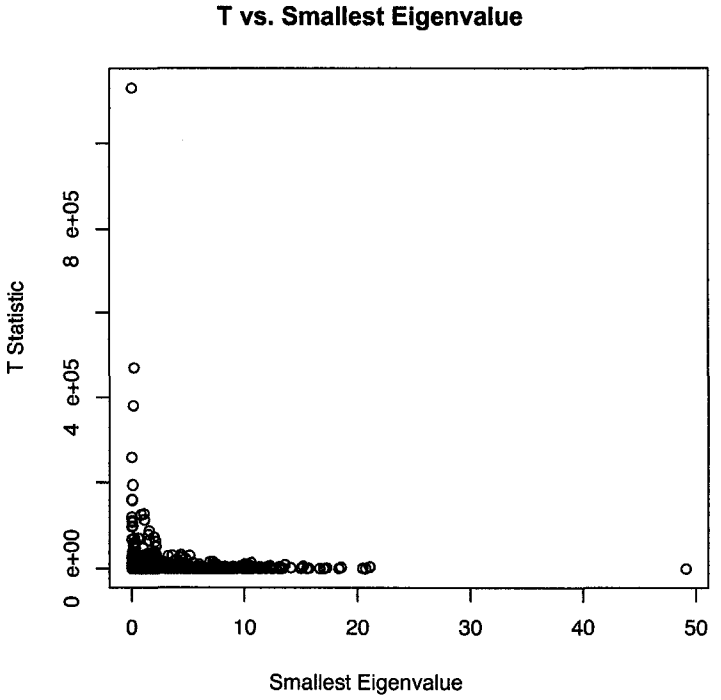

Figure 2.1: Scatter plot of smallest eigenvalue versus Hotelling T statistic

Hotelling's T^2, however, still faces the same problem, since the T^2 statistic becomes large when the eigenvalue of the principal components of the sample variance matrix are small, as shown in Figure 2.1.

This chapter develops a novel adaptive SAM technique to overcome the disadvantages of the test statistics in the classical approach. The objective of this procedure is to identify genes having different transcriptional profiles either in two states or in time course microarray data. The main contributions are:

- An adaptive SAM statistic is proposed, in which the standard error of a gene's expression is adjusted not only by a constant fudge factor, but also according to the standard error itself.
- An algorithm for finding the optimal fudge factor is presented. The first step is to locate an initial point by bootstrap method and the second step is a local search. The criterion being optimized is the concordance between the standard errors of the two sets of replicates being compared.
- We use the same concepts to extend the SAM statistic to the comparison of multivariate samples. This enables us to apply the proposed test statistic to gene expression patterns over a time course, e.g., the pattern of change during the differentiation of stem cells within the first 18 hours.

2.1.1 An adaptive SAM statistic

2.1.2 Univariate sample

We define an adaptive SAM statistic:

$$t_g(c) = \frac{|\bar{x_g} - \bar{y_g}|}{s_g + cf(s_g)} = \frac{d_y}{s_g + cf(s_g)} \tag{2.1.1}$$

where $d_g = |\bar{x_g} - \bar{y_g}|$ is the mean difference between the average level of expression of gene g in two biological states, such as the time of induction of stem cell differentiation

and one day into the differentiation process. $s_g = s_{xy}\sqrt{(1/n_1) + (1/n_2)}$ is the pooled standard error of the two groups (under a common variance assumption), $f(s_g)$ is a function of s_g and c is a constant fudge factor. The test statistic $t_g(c)$ is affected by two factors: the standard error function $f(s_g)$ and the fudge factor. The traditional SAM statistic is obtained by setting $f(s_g) = 1$. The original introduction of the fudge factor was meant to modify the traditional relationship between the distribution of standard error and the test statistic. The definition of adaptive SAM statistic in (2.1.1) is designed for more effective shrinkage performance. By appropriately selecting the function $f(s_g)$ and the fudge factor c, the dependence of test statistic $t_g(c)$ on the standard error s_g is optimally reduced. Even without knowing which c and $f(s_g)$ are optimal, we will later show typical instances where by setting $f(s_g) = 1/s_g$ or $1/\sqrt{s_g}$, shrinkage performance is improved. Intuitively we see in expression (2.1.1) that the adaptive statistic $t_g(c)$ penalizes a smaller value of standard error to a lesser extent than a larger one.

Step 1: Locating a feasible initial point. In searching for the optimal fudge factor, it is desirable to start from a point that is close to the final solution. This can be achieved as follows. Set

$$\frac{d_m}{s_m + c/\sqrt{s_m}} = \frac{d_M}{s_M + c/\sqrt{s_M}}, \tag{2.1.2}$$

where s_m is the median of pooled sample standard errors in gene group G_m whose standard errors are less than the P-th percentile of all genes, and d_m is the median of the mean differences in G_m. s_M is the median of pooled sample standard errors in gene group G_M whose standard errors are between the P-th and $P+\Delta P$-th percentile. d_M is the median of mean difference of G_M. P and ΔP are arbitrarily fixed; 20 % and 15 % would be reasonable values.

To motivate (2.1.2), note that the gene group G_m has smaller average standard errors and will thus tend to have higher t values, while G_M has intermediate average

standard error and t values. In (2.1.2), we are trying to equalize the average of the test statistic of the two groups after standard error adjustment. Solving equation (2.1.2) suggests how to initialize the search for an optimal fudge factor:

$$c_0 = \frac{d_M s_m - d_m s_M}{d_m/\sqrt{s_M} - d_M/\sqrt{s_m}}. \tag{2.1.3}$$

We estimate d_M and d_m by $t_M s_M$ and $t_m s_m$ respectively, where t_M and t_m are the medians of t statistic in gene group G_M and G_m respectively. Then we can replace (2.1.3) by:

$$\hat{c}_0 = \frac{t_m - t_M}{t_M/(s_m\sqrt{s_m}) - t_m/(s_M\sqrt{s_M})}. \tag{2.1.4}$$

Step 2: Search algorithm. In most cases the initializing value of c_0 will not be the best possible, even though it may reduce some of the dependence of the test statistic on the sample variance. To find a better fudge factor $c^* > 0$, we employ a local search algorithm in the neighborhood of c_0. The criterion for c^* is a balance between the number of significantly differential genes with larger sample variance and that of significantly differential genes with smaller sample variance.

Algorithm 3.1

(a) Set an interval length $2L$ $(0 < L < c_0)$, integer $K > 0$, initial value $c_0 > 0$ and iteration index $J = 1$.

(b) Compute the length of the search step: $l = 2L/K$. Let $c_0^* = c_0 - L$, and $c_i^* = c_0^* + i \cdot l$ $(i = 0, 1, 2, \ldots, K)$ be the current set of candidate points.

(c) Set $i = 0$, and for c_i^*, calculate the following ratio on a bootstrap sample of size B:

$$D^b(i) = \frac{\text{Number of genes in group } G_1}{\text{Number of genes in group } G_2} \quad (b = 1, 2, \ldots, B), \tag{2.1.5}$$

where $G_1 = \{g|s^2(g) < p_1^{(b)}$ and $t_g > t^{(b)}\}$ and $G_2 = \{g|s^2(g) > p_2^{(b)}$ and $t_g > t^{(b)}\}$ are two groups of genes. $p_1^{(b)} > 0$, $p_2^{(b)} > 0$ and $t^{(b)} > 0$ are values sampled from three given sets: a "low variance" set of percentiles $P_1 = \{p_1\}$, e.g., $P_1 = \{0.5, 0.10, 0.20, 0.25\}$, a high variance set of percentiles $P_2 = \{p_2\}$ and t-value set of percentiles $T = \{t\}$. For each gene g, t_g is the adaptive test statistic of g using the candidate value of c_i^* and $s^2(g)$ is the sample variance of g. The two groups both have higher values of the adaptive test statistic, but the sample variances in group G_1 are all below a certain percentile level while the variances in group G_2 are all over some percentile level. By averaging these ratios $D^b(i)$, we obtain:

$$D^*(i) = \sum_{b=1}^{B} D^b(i)/B \qquad (i = 1, \ldots, K). \tag{2.1.6}$$

Here, it is necessary to make sure the initial $D^*(0) > 1 + \epsilon_0$. Where, ϵ_0 is a small positive number defined by users.

(d) If $D^*(i) > 1 + \epsilon_0$, then $i = i + 1$, go to (c), otherwise go to (e).

(e) Let $H_J = \min(D^*(i), i = 1, 2, \ldots, K)$. Set $c_0 = c_i^*$, $J = J + 1$, and reset interval length $2L$, integer K if necessary. Go to step (a) unless $|H_J - 1|$ or $|H_J - H_{J-1}|$ is below a predetermined threshold or J is larger than a predetermined positive number.

2.1.3 Multivariate sample

We now discuss an adaptive SAM statistic for comparing multivariate samples. We wish to differentiate the genes having significant variability over the time course from others which do not. Here, a gene profile over a time course can be treated as a sample with a multivariate normal distribution. For example, in stem cell differentiation

studies, we are interested in the following two testing problems:

$$H_0 : \text{mean}(X) = \text{mean}(Y) \tag{2.1.7}$$

or

$$H_0^{(q)} : \mu_1 = \mu_2 = \cdots = \mu_q \tag{2.1.8}$$

where $X = [x_1, x_2 \cdots, x_p]^T$ and $Y = [y_1, y_2 \cdots, y_p]^T$ are two p-dimensional random variables representing a gene expression profile over two different time courses. $\mu_1, \mu_2, \cdots, \mu_q$ are the means of gene expression at q time points respectively. The first test is a general test for comparing mean vectors from two populations, which checks the difference between two functional curves of a gene over two time courses. This is generally done with the F test based on the Hotelling T^2 [Johnson RA. et al. 2002](pages 277–279):

$$T^2 = [\bar{X} - \bar{Y}]^T[(\frac{1}{n_1} + \frac{1}{n_2})S_{\text{pooled}}]^{-1}[\bar{X} - \bar{Y}] \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p,n_1+n_2-p-1}, \tag{2.1.9}$$

where n_1, n_2 are the sample sizes of two groups and

$$S_{\text{pooled}} = \frac{\sum_{j=1}^{n_1}(x_j - \bar{x})(x_j - \bar{x})^T + \sum_{j=1}^{n_2}(y_j - \bar{x})(y_j - \bar{y})^T}{n_1 + n_2 - 2}. \tag{2.1.10}$$

$H^{(q)}$ is a contrast test that checks the difference among expression levels of a gene at different time points over a time course. The test statistic for H is:

$$T^2 = [\bar{X}]^T[CSC^T]^{-1}[\bar{X}] \sim \frac{(n-1)(q-1)}{(n-q+1)n} F_{q-1,n_1+n-q+1}, \tag{2.1.11}$$

where

$$C = \begin{pmatrix} -1 & 1 & 0\ldots & 0 & 0 \\ 0 & -1 & 1\ldots & 0 & 0 \\ \ldots & \ldots & \ldots & 0 & 0 \\ 0 & \ldots & -1 & 1 & 0 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix} \tag{2.1.12}$$

is a contrast matrix. $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ and $S = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})^T$ are the sample mean vector and covariance matrix respectively. The test statistic in (2.1.9) and (2.1.11) can be written as the general form:

$$T^2 = [D]^T[W]^{-1}[D] \tag{2.1.13}$$

where W is a $p \times p$ matrix and D is a p-dimensional vector in the context of H_0 and a W is a $q \times q$ matrix and D is a q-dimensional vector in the context of $H_0^{(q)}$. In the following lemma, we show that the T^2 statistic becomes large as the value of the smallest eigenvalue of the sample variance matrix becomes small.

Lemma 2.1.1 *If W and D are the matrix and the vector defined in (2.1.13), λ_m is the smallest eigenvalue of W and $0 < \lambda_m < M$, then $T^2 > \frac{D^T D}{M}$.*

Proof: By orthogonal decomposition of the positively defined matrix W, we have:

$$T^2 = [QD]^T[\Lambda]^{-1}[QD] \tag{2.1.14}$$

where Q is an orthogonal matrix, $D = (D_1, D_2, \ldots, D_q)^T$ and $E = QD = (e_1, e_2, \ldots, e_q)^T$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \lambda_q \end{pmatrix}, \tag{2.1.15}$$

since $\lambda_m = \min(\lambda_1, \lambda_2, \ldots, \lambda_q)$, then $T^2 = \sum_{j=1}^{q} \frac{1}{\lambda_j} e_j^T e_j > \frac{1}{\lambda_m}\sum_{j=1}^{q} e_j^T e_j > \frac{D^T D}{M}$ □

We now introduce the following adaptive SAM statistic for multivariate sample comparison (following the definitions in (2.1.14) and (2.1.15) and letting $QD = E$):

$$T_g^2(c) = E^T \left(\begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \lambda_q \end{pmatrix} + \begin{pmatrix} \frac{c}{\sqrt{\lambda_1}} & 0 & \ldots & 0 \\ 0 & \frac{c}{\sqrt{\lambda_2}} & \ldots & 0 \\ \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \frac{c}{\sqrt{\lambda_q}} \end{pmatrix} \right)^{-1} E \tag{2.1.16}$$

In adjusting the sample covariance based on the values of all the eigenvalues and the fudge factor c, this approach is similar to the univariate case. In a similar way we have used in univariate case (cf. (2.1.4)) the optimal fudge factor can be determined by a search algorithm starting from an initial fudge factor. The latter can be obtained by solving the following equation:

$$E_m^T \left(\begin{pmatrix} \lambda_m & 0 & \dots & 0 \\ 0 & \lambda_m & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \lambda_m \end{pmatrix} + \begin{pmatrix} \frac{c}{\sqrt{\lambda_m}} & 0 & \dots & 0 \\ 0 & \frac{c}{\sqrt{\lambda_m}} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \frac{c}{\sqrt{\lambda_m}} \end{pmatrix} \right)^{-1} E_m$$

$$= E_M^T \left(\begin{pmatrix} \lambda_M & 0 & \dots & 0 \\ 0 & \lambda_M & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \lambda_M \end{pmatrix} + \begin{pmatrix} \frac{c}{\sqrt{\lambda_M}} & 0 & \dots & 0 \\ 0 & \frac{c}{\sqrt{\lambda_M}} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \frac{c}{\sqrt{\lambda_M}} \end{pmatrix} \right)^{-1} E_M \quad (2.1.17)$$

Similar to equation (2.1.2), but where λ_m is the median of the smallest eigenvalue of the pooled sample covariance matrix in the gene group G_m, whose standard errors are below a certain percentile, and E_m is a vector such that $E_m^T E_m$ is the median of values $D^T D$ in gene group G_m. λ_M is the median of the smallest eigenvalue of the pooled sample covariance of group G_M, whose standard errors are between two given percentiles. E_M is a vector such that $E_M^T E_M$ is the median of values of $D^T D$ for group G_M. Thus, we can get an initial fudge factor c_0 from the solution of (2.1.17):

$$c_0 = \frac{\lambda_m E_M^T E_M - \lambda_M E_m^T E_m}{(E_m^T E_m)/\sqrt{\lambda_M} - (E_M^T E_M)/\sqrt{\lambda_m}}, \quad (2.1.18)$$

where λ_m and λ_M can be estimated by using the median of the unadjusted statistics T_M^2 and T_m^2: i.e.,

$$\hat{\lambda}_m = \frac{E_m^T E_m}{T_m^2} \qquad \hat{\lambda}_M = \frac{E_M^T E_M}{T_M^2} \quad (2.1.19)$$

The search algorithm for improving the fudge factor is similar to Algorithm 3.1

2.1.4 P-value false discovery rate

Since the test statistics (2.1.1) and (2.1.16) are no longer t- or F-distributed, we will undertake a bootstrap procedure similar to the method in the original SAM paper [Tusher, et al, 2001]:

- Suppose the fudge factor c has already been identified. Calculate each gene's adaptive statistic $t_g(c)$ or $T_g^2(c)$.
- For each gene g, with n_1 samples in one state and and n_2 in the other, do B random reassignments of all the samples into two groups of sizes n_1 and n_2. For the i-th such reassignment ($i = 1, 2, \ldots, B$), calculate $t_g^{(i)}(c)$ or $T^{2(i)}_g(c)$.
- Compute $\hat{t}_g(c) = \frac{1}{B}\sum_{i=1}^{B} t_g^{(i)}(c)$ or $\hat{T}_g^2(c) = \frac{1}{B}\sum_{i=1}^{B} T^{2(i)}_g(c)$.
- Any gene such that $\theta_g = |t_g(c) - \hat{t}_g(c)| > \Delta$ or $\Theta_g = |T_g^2(c) - \hat{T}_g^2(c)| > \Delta$ is defined as differentially expressed.

Remark: To conduct the contrast test H_2 would require a small change in the procedure described above. In this case we should shuffle all the samples on each dimension, i.e., a total of $r \times m$ values (where m is the number of dimensions of the sampled variable and r is the number of replicates) are obtained and stored in a vector V. We then assign each of m groups with r values that are randomly selected from V (without replacement). Finally, these m groups are re-defined as an m-dimensional distribution sample group, having r replicates for each dimension of the variable. The calculation of p-values is the same as before. The calculation of the false discovery rate (FDR) for each Δ follows the procedure in [Tusher VG. et al. 2001].

2.2 Numerical experiments and results

This section presents some computational results on the adaptive statistic and associated algorithms as applied to gene expression data on J1 embryonic stem cell (ESC)

lines [Perez-Iratxeta C. et al. 2005]. Our experiments consist of the following test

2.2.1 Experiment I: univariate case

This data set contains microarray expression data on 4000 genes randomly selected from among the 20,000 genes that showed higher-than-median expression levels, out of the 40,000 available on the J1 ESC line. Each gene was sampled at 0 hour and 18 hours, with 3 replicates at each time point.

The purpose is to identify the genes that undergo significant change during this early period, as clues to the mechanisms that trigger the differentiation of ESC. We applied the univariate adaptive SAM statistic with $f(s_g) = 1/s_g$ to this data set. Using expression (2.1.4), we initialized c_0=10045. In Algorithm 1, we used a bootstrap sample of size 100 for the ratio D defined in (2.1.6). After two iterations of the algorithm, we obtained $c = 6900$, while D reached a low value of 111. By comparison, before standard error adjustment, $D = 473$. Figure 2.2 exhibits two graphs of the relationship between the fudge factor c and D values during the two runs of the search algorithm.

When $f(s_g) = 1$, the adaptive SAM statistic reduces to the general SAM statistic. In this case, $c = 105$ and $D = 121$. In order to confirm the superiority of the adaptive statistic, we drew 30 bootstrap samples in the neighbourhood of the fudge factor obtained with and without the adaptive statistic ($c = 6900$ and $c = 105$, respectively). After each bootstrap, we compared the two D values produced with and without the adaptive statistic. In 90 % of the cases, the D value of the general SAM statistic was worse (larger) than that of adaptive SAM statistic. This demonstrates the stronger shrinkage effect of the adaptive statistic.

Figure 2.3 shows the relationship between the pooled standard error and the test statistic (before and after adjustment). The right hand graph shows a greater dispersion of the standard error values and indeed the negative correlation between the test

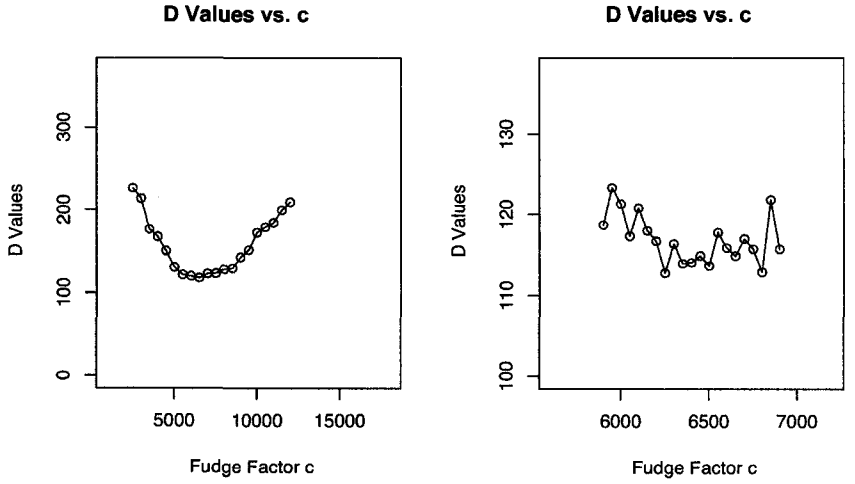

Figure 2.2: Fudge factor c and corresponding D values in two runs of search algorithm

statistic and the standard error that exists in the left hand graph almost disappears in the right hand one. These facts also reflect the greater shrinkage performance of the adaptive statistic.

Table 2.1 contains an evaluation of the adaptive SAM method in testing for differential genes of ESC (0 hour and 18 hours). "Called Number" refers to the genes detected by the method. "False Discovery Rate" is estimated with the help of bootstrapping permutations, sample size 66. The plot shows at initial levels, the increase of cutoff causes larger reduction in Called Number of genes and corresponding FDR. The pattern becomes very flat at higher cutoff levels. Therefore, in differential gene expression study, this plot can help us determine the "wanted" Called Number or FDR. For instance, if we need to capture 250 most differential genes, then cutoff should be chosen as 6 according to Table 2.1.

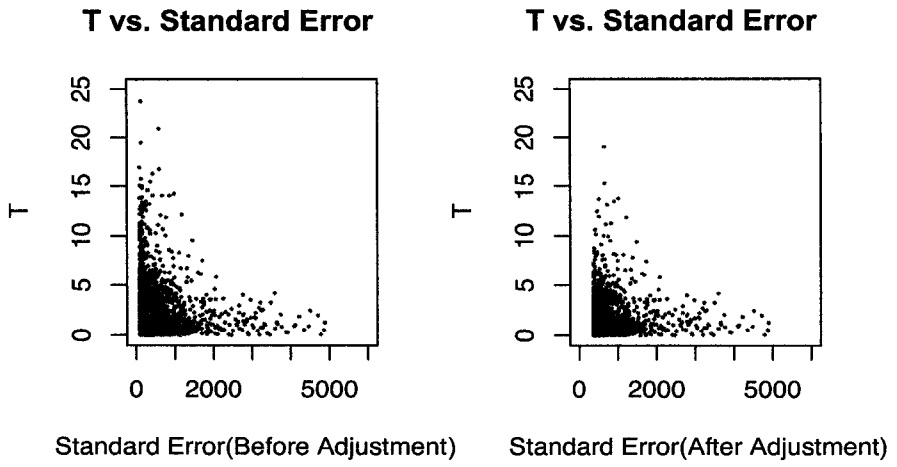

Figure 2.3: Scatter plot of pooled standard error versus test statistic before and after adjustment

Value of Δ	Called Number	False Discovery Rate
0.5	3167	0.801
0.7	2863	0.663
0.9	2583	0.506
1.1	2327	0.403
1.3	2076	0.366
1.5	1822	0.213
1.7	1608	0.172
3.1	686	0.074
3.7	494	0.062
4.3	337	0.061
6.5	113	0.061
7.9	70	0.048
9.9	40	0.071
10.5	34	0.061

Table 2.1: Summary of adaptive SAM method in testing differential genes expression in ESC

2.2.2 Experiment II: multivariate case

This test is based on time series gene expression data on 1200 randomly selected genes from the J1 ESC line. The data were collected at 0, 6,12 and 18 hours, with 3 replicates each. For each gene, we wish to study if the the earlier span of the gene profile, from the initial time point to 12 hours is significantly different from the later span, from the 6 hours to 18 hours. Though these curves share two out of three points, in comparing them, the second curve is offset so that the comparison at all three points involves different values. In this way, we can identify the genes undergoing significant changes over time course while retaining the multivariate nature of the time course data.

Compared with the univariate test, the multivariate one is a more subtle evaluation of the cell differentiation process because it distinguishes among the time points, and is thus sensitive to covariance and other multivariate properties. For our experiment, we used $f(s_g) = 1/\sqrt{s_g}$ as an archetypical way of introducing dependance on the standard error. From (2.1.16) and (2.1.17) we determined an initial fudge factor $c_0 = 0.5$. Employing a search algorithm similar to that in Section 2.2.1, we settled on a final fudge factor $c = 2.5$, with associated $D = 114$. In contrast, the D value for the classical Hotelling T^2 is 1673, indicating that it is seriously biased by small eigenvalue. Figure 2.4 shows the relationship between the fudge factor c and the D values in the final run of the search algorithm. Figure 2.5 describes the connection of smallest eigenvalue to the adaptive SAM statistic (before and after adjustment after trimming some outliers).

Table 2.2 displays the results of the adaptive SAM method in testing the pairs of offset curves for each gene. As before, the False Discovery Rate (FDR) is computed by 66 bootstrapping permutations for the multivariate samples of genes.

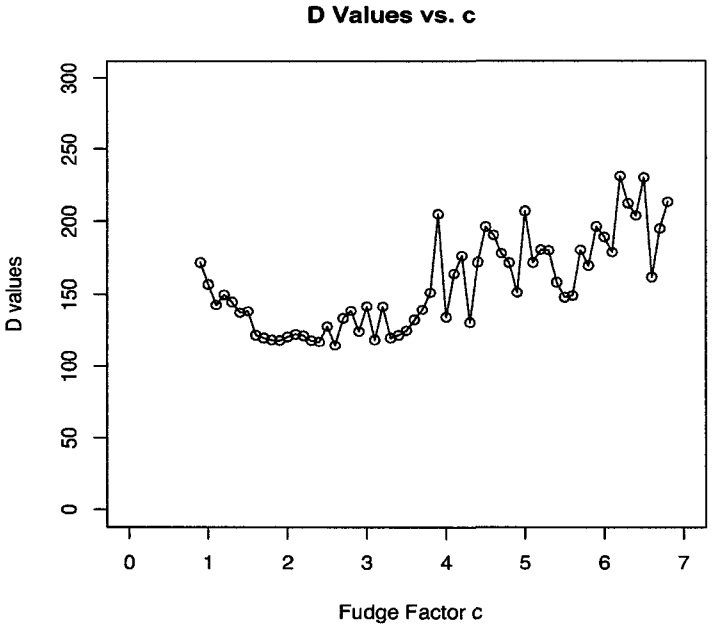

Figure 2.4: Fudge factor c versus D values in the searching algorithm

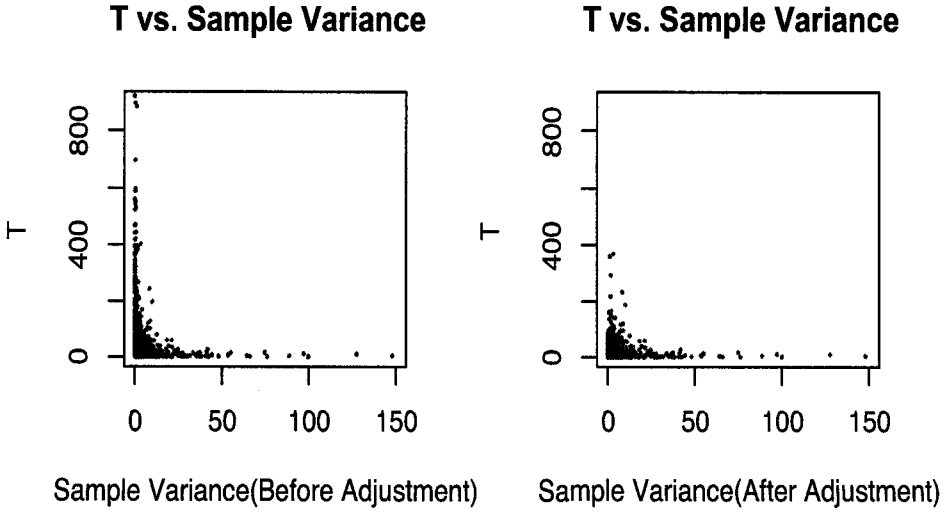

Figure 2.5: The smallest eigenvalue versus test statistic (before and after adjustment

Value of Δ	Called Number	False Discovery Rate
0.8	947	0.606
1.0	928	0.507
2.0	814	0.121
3.0	748	0.078
5.0	607	0.063
10	413	0.121
15	297	0.046
20	230	0.121
25	186	0.181
30	143	0.045
35	115	0.105
40	99	0.030
50	69	0.104
55	57	0.044
60	48	0.133
65	40	0.103
70	34	0.073
75	29	0.102
80	25	0.087

Table 2.2: Test of differential ESC gene profiles over time course

	Differentially Expressed Genes
Down-Regulated	scd2 Nasp Ppm1g Dstn Klf4 Pou5f1 Orc2l Gfpt2 Ywhab Rbm38 Ctbp Clns1a Tcof1 Rest Rrp1 Nup88 Pop1 Slc17a7 Dpp8 Adad2 Nrip1 Nfrkb Jam2 Fbxo46 BC06539 Pigl AU04415 FLJ2071 Mtap6 Ubr5 Eed Hnrpdl Arhgdia Jmjd4 Eftud2 Rbbp7 Chrac1
Up-Regulated	Pfkp Glg1 Cldn4 Itga6 Mad2l1 Rpl14 Chd3 Gfpt1 Atm Prkaa2 Inpp5f Kras Rp2h Ube2n Stx12 Tmem201 Rab11fi Bri3bp Dnpep Smtnl2 Anxa6 Wdr61 Ccng1 Ccnb2 Tpi1 Igsf3 Spire1 Sdf4 Rapgef5 Srp14 AI42929

Table 2.3: Differential genes detected by adaptive SAM but not by classical SAM using 5% cutoff(0 hour-18 hours)

2.2.3 Experiment III: Analyzing the set of genes detected by the adaptive SAM statistic

In this experiment we use 19,000 genes preliminarily selected from a total of 40,000 genes by excluding lowly expressed genes over the time course of J1 ESC differentiation [0h,18h]. Specifically, we use the following approach:

- Apply both adaptive and classical SAM statistics to detect genes with significant differential expression over two time points: 0 hour and 18 hours.
- Use the top 5 percentile cutoff level of statistic θ_g values (see sec 2.3) to identify the candidate genes
- Analyze the biological properties of genes that are identified by adaptive SAM but not by classical SAM (see Table 2.3)
- Analyze the biological properties of genes that are identified by classical SAM but not by adaptive SAM (see Table 2.4)

The list of genes exclusively found by adaptive SAM but not by classical SAM contains a significant number of genes known to have important functions in the

	Differentially Expressed Genes
Down-Regulated	Mrpl52 Tcfe3 Syncrip Krt8 Ndufa9 Nsdhl Sept11 Cbr3 Xpo6 Mkrn1 Atp6v0e Rpl7l1 Csnk1a1 Paics Med19 Ythdf2 Pgs1 Ddx24 Tcl1 Lsg1 Atad3a Cdc42se1 Idi1 Krt18 Iws1 Ppp1r15b Npepps Ppan Myadm Ap3d1
Up-Regulated	Rps15 Ppp2r5c Exoc2 Myo10 Lrwd1 Ube2h Nfe2l2 Rp137 Tmem134 Adfp C030005K06Rik Tmco6 Ppm1m Plcg1 D10Ertd610e Mboat2 Chrna7 Dcxr Plcd1 ENSMUSG00000051554

Table 2.4: Differential genes detected by classical SAM but not by adaptive SAM using 5% cutoff(0 hour-18 hours)

regulation of mESC differentiation. In particular, the group of down-regulated genes is significantly enriched in genes encoding nuclear proteins (18 genes out of 32 with Gene Ontology annotation, P-value 0.0023 using GOstat [Beissbarth T. et al. 2004] including transcription factors with known functions in ESC such as Oct4 (Pou5f1), Rest, and Klf4. For comparison, the list of down-regulated genes detected by SAM but not by adaptive SAM did not include transcription factors, except for Tcfe3.

We note the well-known stem cell marker POU5F1, a transcription factor also known as OCT3/4, which is involved in regulation of pluripotency during ESC differentiation, is listed in the down-regulated part of Table 2.3. It is known that the expression of OCT3/4 is down-regulated in the early period of ESC development and this causes the self-renewing property of OCT3/4 to gradually diminish. The RE1-silencing transcription factor (REST) also appears among the down-regulated genes. REST plays a dual role in maintaining self-renewal and pluripotency. In ESC differentiation it blocks the expression of a microRNA that prevents ESC from reproducing themselves and causes them to differentiate into specific cell types [Singh SK. et al. 2008]. Other genes with ESC related functions found by adaptive SAM but not by classical SAM are Sdf4 (up-regulated) and Clns1a (down-regulated). Sdf4 (Stromal cell-derived factor 4) appears in the up-regulated list. This gene is

abundantly expressed in hematopoietic stem cells (HSC) and plays an essential role in promoting maintenance of self renewal [Park In. et al. 2002]. Clns1a (chloride channel) is in the list of 332 cDNAs proposed to be embryonic stem cell-enriched [Puente LG. et al. 2006]. In studying the proteins detected in ESC and embryonic bodies after phosphoprotein enrichment, Clns1a is found to be over-expressed in both studies. TPI1 (Triosephosphate isomerase) plays a function in energy production and metabolism and is highly expressed in three kinds of stem cell (embryonic, neural and mesenchymal) [Baharvand H. et al. 2007].

In sum, when using a 5 % cut-off level, SAM misses a number of crucial stem cell markers that adaptive SAM finds. The genes found exclusively by classical SAM do not tend to be associated with "stemness".

2.2.4 Numerical experiment IV: Comparing with classical SAM using simulation data

In this section we compare the Correct Discovery Rate (CDR) among adaptive, traditional SAM and ordinary t statistics. The gene expression simulation tool [Albers CJ. et al. 2006] is used to generate expression values for 1500 gene genes, in which 215 are significantly expressed (100 up-regulated, 115 down-regulated).

To compare their CDR performance at different cutoff levels, we first rank (ascending order) 1500 genes according to θ_g (see Section 2.3), t and the traditional SAM statistic respectively, and then separate them into 20 groups of 75 genes. The 20^{th} group represents the most significant genes in terms of the detecting statistic. The means of each statistic within all groups $(1-20)$ are calculated and used as the cutoff levels to determine the differentially expressed genes. The gene lists are then compared with genes actually expressed to compute CDR at different cutoff levels. The results are shown in Table 2.5.

Cutoff	SAM(%)	t(%)	Adaptive SAM(%)	Improvement 1	Improvement 2
1	14.72	14.74	14.73	0.1%	-0.1%
2	15.04	15.20	15.21	1.2%	0.1%
3	15.90	15.98	15.97	0.4%	-0.1%
4	16.75	16.75	16.72	-0.2%	-0.2%
5	17.63	17.68	17.76	0.7%	0.4%
6	18.58	18.95	18.88	1.6%	-0.4%
7	19.78	20.12	19.88	0.5%	-1.2%
8	21.04	21.45	21.26	1.0%	-0.9%
9	22.91	22.98	23.02	0.5%	0.2%
10	24.65	24.81	24.68	0.1%	-0.5%
11	27.25	27.12	27.17	-0.3%	0.2%
12	29.95	30.46	30.49	1.8%	0.1%
13	33.93	34.11	33.87	-0.2%	-0.7%
14	38.51	38.70	38.72	0.5%	0.0%
15	44.53	44.31	44.90	0.8%	1.3%
16	53.57	52.98	53.41	-0.3%	0.8%
17	65.25	64.91	67.18	3.0%	3.5%
18	80.21	79.46	82.98	3.4%	4.4%
19	95.50	94.44	98.15	2.8%	3.9%
20	100.00	100.00	100.00	0.0%	0.0%

Table 2.5: Correct Discovery Rates for Three Methods

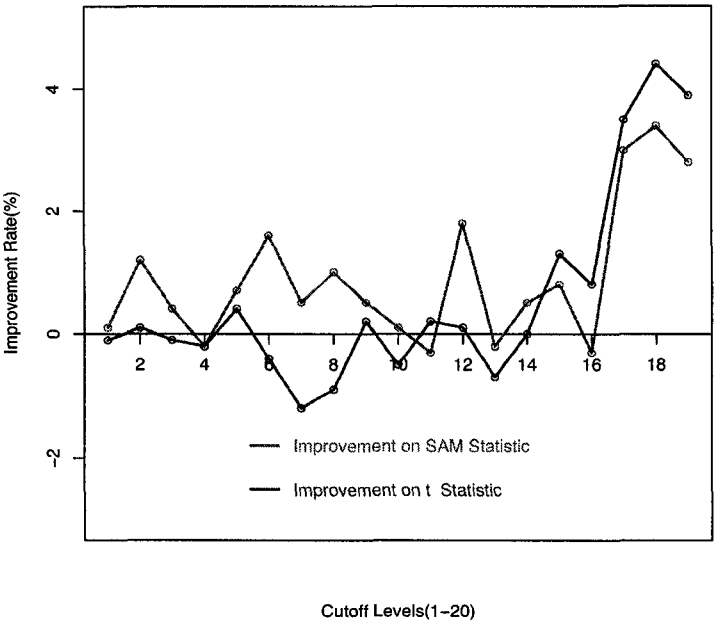

Figure 2.6: Improvement of Adaptive SAM

The $2^{th} - 4^{th}$ columns of Table 2.5 present the CDR for each statistic across different cutoff levels $(1-20)$. The column "Improvement 1" is the CDR improvement rate of the adaptive SAM statistic over the classical SAM statistic, and "Improvement 2" gives the CDR improvement rate over the ordinary t statistic. At first glance, these CDRs are all seen to show an increasing trend from lower to higher cutoff levels. The performance of the three statistics are very similar before 17^{th} groups. For the top ranked groups $17^{\text{th}} - 20^{\text{th}}$ (i.e., top 15 percentile), the SAM statistic and the adaptive SAM statistic perform better than the ordinary t statistic and the adaptive SAM statistic reaches the highest CDR. The trend of CDR improvement rates are plotted in Figure 2.6.

In Figure 2.6, we see that if we adopt the proposed adaptive SAM statistic to detect over-expressed genes, the CDR can significantly increase at higher percentile

levels. This is important since there is only a small fraction of genes that are really significantly differential in many cases.

2.3 Discussion

In this chapter we have proposed an adaptive SAM test for improving differential gene expression detection. We extend the univariate case to the multivariate case, which may be used for comparing time-series gene expression data. There are two features of the proposed techniques that differ from the classical SAM method: (1) The standard error function (SEF) is added to the test statistic as a penalty factor. The fudge factor combined with the SEF results in a more effective shrinkage in gene expression data, as we have shown by comparing their D values. (2) The fudge factor is chosen according to D value. i.e., we try to equalize the number of genes with higher values of the test statistic as well as higher standard error, and the number of genes with higher values of test statistic but lower values of the standard error. In this way, the significant genes are evenly distributed over the lower and upper tails of their standard error. The optimal fudge factor can be obtained by initializing a search at a point obtained from (2.1.4) and (2.1.18) respectively.

The proposed adaptive SAM statistic in univariate and multivariate cases can be regarded as the shrinkage of the general t and F statistics, respectively. There is some work on the SAM F statistic [Amaratunga D. et al. 2003](page:113-123), but the correlation between samples is not taken into account in these studies. [Guo X. et al. 1995] consider sample correlation by using a 2×2 weighted matrix, which adopts a matrix adjustment policy similar to the one in our study. By comparison, our study employs adjusted Hotelling T^2, which can be used to compare multivariate data over a time course.

We next discuss the selection of the SEF. The specific SEF adopted in this chapter can be explained in terms of the t-statistic with a Bayesian adjusted denominator as

proposed by [Baldi P. et al. 2001]. Aspects of the Bayesian probabilistic framework mainly comprises the following approaches:

- Assuming that the gene observed expression levels are normal and independent, the likelihood of the data D is given by:

$$P(D|\mu,\sigma^2) = \prod_{j=1}^{n} Normal(x_j;\mu,\sigma^2) \tag{2.3.1}$$

- Introducing a prior distribution $P(\mu,\sigma^2)$ using Bayesian theory and a conjugate prior $P(\sigma^2)$:

$$P(\mu|\sigma^2) = Normal(\mu;\mu_0,\sigma^2/\lambda_0) \tag{2.3.2}$$

$$P(\sigma^2) = I(\sigma^2;v_0,\sigma_0^2) \tag{2.3.3}$$

$$P(\mu,\sigma^2) = P(\mu|\sigma^2)P(\sigma^2), \tag{2.3.4}$$

  where $I(\sigma^2;v_0,\sigma_0^2)$ is a scaled inverse gamma density, and (2.3.4) is still a scaled inverse gamma density with the parameter set $\alpha = (\mu_0,\lambda_0,v_0,\sigma_0^2)$

- The posterior of (μ,σ^2) is given by

$$P(\mu,\sigma^2|D,\alpha) = Normal(\mu;\mu_n,\sigma^2,\sigma^2/\lambda_n)I(\sigma^2;v_n,\sigma_n^2) \tag{2.3.5}$$

  with $\mu_n = \frac{\lambda_0}{\lambda_0+n}\mu_0 + \frac{n}{\lambda_0+n}m$, $\lambda_n = \lambda_0 + n$, $v_n = v_0 + n$ and $v_n\sigma_n^2 = v_0\sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0+n}(m-\mu_0)^2$ where, v_n is degree of freedom of the posterior, n is the sample size, and m is the sample mean. The posterior sum of squares $v_n\sigma_n^2$ is the sum of prior sum of squares $v_0\sigma_0^2$, the sample sum of squares $(n-1)s^2$, and the residual created by the prior mean and sample mean.

According to (2.3.5), if we use $\mu_0 = m$, then the variance of the posterior can be estimated by

$$\sigma^2 = \frac{v_n \sigma_n^2}{v_n - 1} \doteq \sigma_n^2 = \frac{v_0}{v_0 + n}\sigma_0^2 + \frac{(n-1)s^2}{v_0 + n} \tag{2.3.6}$$

The sample sum square $(n-1)s^2$ can further be expressed as $n(\sigma_s^2 + m^2)$, where σ_s^2 is the sample variance. Therefore, (2.3.6) can be written as a generic form:

$$\sigma^2 \doteq \frac{v_0}{v_0 + 1}\sigma_0^2 + \frac{n}{v_0 + n}\sigma_s^2 + \frac{n}{v_0 + n}m^2 = C_1 + C_2\sigma_s^2 + C_3 m^2, \tag{2.3.7}$$

where C_j $(j = 1, 2, 3)$ are constants. If we compare the gene expression levels between two groups, the pooled variance of the posterior is give by

$$\sigma_p^2 \doteq C_1 + C_2\sigma_{sp}^2 + C_{31}m_1^2 + C_{32}m_2^2 = C_1 + C_2\sigma_{sp}^2 + C_3 m_{sp}^2, \tag{2.3.8}$$

where m_j $(j = 1, 2)$ are the sample means of two gene groups, σ_{sp}^2 is the pooled sample variance, and $C_3 m_{sp}^2 = C_{31}m_1^2 + C_{32}m_2^2$ is the pooled sample mean square. Note, the offset $C_1 + C_3 m_{sp}^2$ is similar to the constant fudge factor in classical SAM, except that the term here is associated with the sample mean. If we treat $C_3 m_{sp}^2$ equally for all genes, then the offset turns to be a constant, the case in classical SAM, which means the sample mean has no relationship with sample variance. However, if such a relationship (often negative) exists, then we need a stronger shrinkage effect in addition to the constant fudge factor. Thus, we can think of the pooled sample mean square as having negative relationship (could be nonlinear), with sample variance, for instance:

$$m_{sp}^2 = C_0/\sigma_{sp}^2 \tag{2.3.9}$$

or

$$m_{sp}^2 = C_0/\sqrt{\sigma_{sp}^2} \tag{2.3.10}$$

This justifies the SEF used in the previous sections. There are certainly many other possibilities, and it would be desirable to improve the proposed method using more efficient SEFs for describing gene expression levels. One way to do this, for example, is to check the relationship between the sample mean and variance in term of the data patterns.

The other issue is the convergence of the fudge factor. Note in 2.1.6, because $D^*(i)$ is a decreasing sequence with a lower bound 1, it then converges to some point larger than 1.

Chapter 3

Dynamical correlation for clustering expression trajectories in differentiating stem cells

This chapter is based on the paper in Computational Methodologies in Gene Regulatory Networks (S. Das, D. Caragea, W. H. Hsu & S.M. Welch, eds.) IGI Global, in press. I wrote a complete draft of this chapter. Professor Miguel Andrade verified and expanded upon my interpretations of the gene ontology search results. Professor David Sankoff suggested some of the additional research directions and helped organize and reformulate the exposition.

3.1 Background

A quantitative indicator of gene expression is provided by the amount of mRNA produced by DNA transcription of genetic information. Taking advantage of this, DNA microarrays represent a high throughput source of gene expression data that can be applied to analyze gene function and networks of gene regulation. To investigate the changing genetic function of cells, we derive a dynamical gene expression profile (or curve) over a time course. In this chapter, for example, we study embryonic stem cell differentiation, where the expression of different genes may follow different trajectories, as measured at a number of sampling times.

The set of measurements pertaining to a gene at discrete points is often treated as an observation of a multivariate probability distribution. These data are generally submitted to a cluster analysis, based on the assumption that in clustering time-course expression data, genes in the same cluster are directly or indirectly involved in the same biological functions. Gene expression levels at a certain time point may be impacted by other mechanisms such as the expression levels of other regulators at previous or current time points, or by experimental errors. Rather than treating a gene profile as a static multivariate sample, then, it is more appropriate to treat it as a random curve across a time interval.

Cluster analysis is used to identify groups or clusters of objects on the basis of a series of measurements made on these objects. It is crucial to be able to characterize the clusters in ways pertinent to the provenance of the data. In regulatory networks, for instance,a cluster of genes could be identified as up-regulated or down-regulated over time. Many well-known clustering methods have been presented in chapter 1. We stress here, however, that gene profiles are derived from the trajectories of gene expression over a time course and that the similarity in "shape" between two expression curves should be the key input to the clustering analysis [Ramoni MF. et al. 2002]. For example, the Pearson correlation coefficient partially reflects the similarity and

differences of two curves, but it fails to adequately consider the time effect since one can always exchange all the observations at one time point with all those at another point, without affecting the clustering result.

In this chapter we address microarray data on embryonic stem cell lines (ESC). The data set records the differentiation information in the early period of ESC through screening gene expression levels over a 14-day observation period [Perez-Iratxeta C. et al. 2005]. As the number of genes is very large (about 26,000), the data is pre-processed by selecting those genes that manifest the most significant differential over the 0 - 24 hour time course, using the statistical procedure SAM [Tusher VG. et al. 2001]. The smaller data set thus obtained contains 524 genes, representing the "stemness" signature in the early period of ESC differentiation.

First, we define the dynamical correlation coefficient is in terms of random curves, to be employed as a similarity measure in the comparison of gene profiles prior to cluster analysis. For clustering real time-course expression data, we propose an estimation method to handle the discrete-time nature of the observations. We carry out computational experiments on simulated data before applying our methods to a real ESC data set. We check the performance of this procedure by comparing Gene Ontology(GO) terms within and between clusters.

3.2 Dynamical Correlation

3.2.1 Definition

We define a shape similarity measure in terms of dynamical correlation and integrate it into a general agglomerative hierarchical clustering scheme. To define dynamical correlation, we visualize a time-course gene profile as a random curve:

$$f(t) = a + \mu(t) + \eta(t) \tag{3.2.1}$$

where $\mu(t)$ is the mean function of time t. a is a constant item representing the mean of the random curve. $\eta(t)$ is the dynamical random variable such that $E(\eta(t)) = 0$. The functions $\mu(t)$ and $\eta(t)$ are assumed to be smooth (twice continuously differentiable) over the time course.

A key feature of a random curve is that it contains both static and dynamical random variables. Thus, it is capable of accurately describing both the deterministic and stochastic aspects of the time-course gene expression observations. If two expression profiles are expressed in the form (3.2.1), then the similarity between the shapes of two gene curves will mainly depend on the function. Before defining dynamical correlation, we should define the shape of a random curve. For doing so, we first approximate (3.2.1) by a Taylor expansion in the neighborhood of t^*:

$$f(t) \doteq a + \mu(t^*) + \mu'(t^*)(t - t^*) + \eta(t) \tag{3.2.2}$$

then we can use the following derivative to describe the random curve shape around t^*:

$$\frac{d(f(t))}{dt} \mid_{t=t^*} = f'(t) \mid_{t=t^*} = \mu'(t^*) \tag{3.2.3}$$

If $f(t)$ and $g(t)$ are two random curves for gene f and g respectively, then under the assumption that $f'(t)$ and $g'(t)$ are jointly measurable over the time interval $[T_0, T_1]$,we define the following dynamical correlation coefficient:

$$\varrho = \frac{1}{T_1 - T_0} \int_{T_0}^{T_1} E|f'(t) - g'(t)|dt \tag{3.2.4}$$

Since $f'(t)$ and $g'(t)$ are related to the expression change rate and direction at time t, and have nothing to do with the constant items, $|f'(t) - g'(t)|$ can be regarded as the 'shape difference' between two curves. Thus (3.2.4) reflects the average shape difference' between $g(t)$ and $f(t)$. In the next subsection we will estimate the dynamical correlation coefficient ϱ using observations at discrete time points. Then we will

use it as a similarity measure to produce input for the cluster analysis of time-course gene expression.

Remark: The definition (3.2.4) does remove the static random effects of the random curves $f(t)$ and $g(t)$ through derivative operation on the curves of gene profiles.

3.2.2 Discretization

Assume we observe expression levels at K points over a time course, denoted by $g(t_k)$ and $f(t_k)(k = 1, 2, , K)$. To apply ([?]), we adopt the following approximation procedure [Dubin JA. et al. 2005]:

$$\hat{\varrho} = \sum_{s=1}^{K-1} \frac{1}{2}(\delta_s + \delta_{s+1}) E\{|f^{'}(t_s) - g^{'}(t_s)|\} \tag{3.2.5}$$

where $\delta_s = t_s - t_{s-1}$ $(s = 1, 2, \cdots, K)$. $\hat{\varrho}$ is an estimate of ϱ on the basis of discretized time-course expression data. Here we use the Riemann sum to approximate the integral. In a similar way, we can approximate $f^{'}(t_s)$ and $g^{'}(t_s)$ with the following formulae:

$$f^{'}(t_s) = \frac{1}{2}\{\frac{f(t_{s+1}) - f(t_s)}{t_{s+1} - t_s} + \frac{f(t_s) - f(t_{s-1})}{t_s - t_{s-1}}\} \qquad (s = 2, 3, \cdots, K-1) \tag{3.2.6}$$

$$\hat{g}^{'}(t_s) = \frac{1}{2}\{\frac{g(t_{s+1}) - g(t_s)}{t_{s+1} - t_s} + \frac{g(t_s) - g(t_{s-1})}{t_s - t_{s-1}}\} \qquad (s = 2, 3, \cdots, K-1) \tag{3.2.7}$$

For t_1 and t_K we have:

$$\hat{f}^{'}(t_1) = \frac{f(t_2) - f(t_1)}{t_2 - t_1} \quad \hat{f}^{'}(t_K) = \frac{f(t_K) - f(t_{K-1})}{t_K - t_{K-1}} \tag{3.2.8}$$

$$\hat{g}^{'}(t_1) = \frac{g(t_2)) - g(t_1)}{t_2 - t_1} \quad \hat{g}^{'}(t_K) = \frac{g(t_K) - g(t_{K-1})}{t_K - t_{K-1}} \tag{3.2.9}$$

Substituting the expressions (3.2.6)-(3.2.9) into (3.2.5), we obtain a similarity measure that approximates the dynamic correlation between two random gene curves. Note that if there are replicate observations over the time course, a bootstrap approach can be used in estimating the expectation in (3.2.6); we will discuss this in section 3.5.

3.3 Simulation study

In this section we will compare the performance of dynamic correlation with classical Pearson correlation when applied into the cluster analysis of two simulated time-course gene expression data sets. The simulation study is composed of three parts. First, we compare the within/between cluster dispersion effect between the two similarity measures using bootstrap method. Second, the performances of the similarity measures are compared when input to the K-means clustering method. Finally, we will compare them when used in a hierarchical agglomerative approach.

3.3.1 Within-Between-Cluster dispersion effect

This example consists of simulated time-course expression data from two groups of genes, with trajectories as follows:

$$G(t) = R_1|\sin(t)| + Q_1|t| + C + \eta(t) + \varepsilon \tag{3.3.1}$$

$$F(t) = R_2|\sin(t)| + Q_2|t| + C + \eta(t) + \varepsilon \tag{3.3.2}$$

where R_1,Q_1,R_2,Q_2,$\eta(t)$ and ε for each gene are sampled from the following probability distributions:

$$R_1 \sim Normal(16, 0.1) \quad R_2 \sim Normal(12, 0.2) \tag{3.3.3}$$

$$Q_1 \sim Normal(1, 0.1) \quad Q_2 \sim Normal(2.5, 0.1) \tag{3.3.4}$$

$$\eta(t) \sim Normal(0, 0.05) \qquad \varepsilon \sim Normal(\mu = 0, \sigma = 0.08) \tag{3.3.5}$$

Thirty successive time step sizes were fixed at the outset by randomly selecting from the set $1, 2, 3, 4$. It is obvious that both clusters are up-regulated over the time course $[0, 142]$. To study the within/between cluster dispersion performance of the two similarity measures, we use the following bootstrap procedure:
Calculate the two similarity coefficients as follows:

- For each group, bootstrap 200 times to compute the within cluster similarity between two curves that are randomly selected from that group. Then store the 200 similarities in vector $V1$,
- Randomly select one curve from each group and compute the between cluster similarities between two curves. Bootstrap 200 times and store these similarities in vector $V2$,
- Do pairwise comparison between the elements of $V1$ and $V2$, and calculate the number of times the between-cluster similarity is less than the "within cluster" similarity. That is:

$$R = \frac{\#(comparisons | within \geq between)}{40000} \tag{3.3.6}$$

We find that $R = 0.54$ for the Pearson correlation coefficient, and $R = 0.79$ for the dynamical correlation coefficient. The latter measure is more effective in clustering the simulated time-course gene expression data.

Cluster Found	Dynamical correlation: members recovered from original clusters (1, 2, 3, 4)	Pearson correlation: members recovered from original clusters (1, 2, 3, 4)
I	(36, 0, 0, 0)	(13, 0, 0, 14)
II	(0, 33, 0,0)	(11, 0, 0,7)
III	(0, 1, 52, 0)	(0, 34, 52, 0)
IV	(0, 0, 0, 38)	(12, 0, 0, 17)

Table 3.1: Clustering Result of K-Means Method Using Two Similarity Measures

3.3.2 Comparison of two similarity measures in K-means clustering approach

In this subsection we cast two similarity measures into K-means clustering method. Four clusters of time-course gene expression data are designed as follows:

$$A_i(t) = r_i|\sin(t)| + q_i|t| + C + \eta(t) \qquad (i = 1, 2, 3, 4) \tag{3.3.7}$$

As in (3.3.1) and (3.3.2), the r_i and q_i are all random variables with different normal distributions (different means). The time step sizes and the lengths of the time course are also selected in a similar way. It should be noted that clusters 1 and 4 are basically down-regulated, while clusters 2 and 3 are up-regulated. The trajectories also manifest a periodic tendency due to the sine term. The clustering results are shown in Table 3.1.

This analysis shows that the K-means method using dynamical correlation can effectively separate four groups. By comparison, Pearson correlation fails to distinguish between clusters 2 and 3 or between clusters 1 and 4. This test demonstrates that dynamical correlation can, and Pearson correlation cannot, accurately reflect the simulated gene expression changes.

Cluster Found	Dynamical correlation: members recovered from original clusters (1, 2, 3, 4)	Pearson correlation: members recovered from original clusters (1, 2, 3)
I	(36, 0, 0)	(35, 0, 11)
II	(0, 38, 0)	(0, 39, 7)
III	(0, 1, 25)	(1, 0, 6)

Table 3.2: Clustering Result of Agglomerative Method Using Two Similarity Measures

3.3.3 Comparison of the Similarity measures with Agglomerative Clustering

The agglomerative clustering method starts with n clusters (the original n random curves) and then groups the two closest curves, according to the similarity measure, into one cluster. The new cluster is assigned a similarity to the remaining objects equal to the mean of the corresponding scores of the members of the cluster. The process continues until all curves merge into one cluster. Here we compare the clustering performance of Pearson's correlation coefficient, often used in gene expression clustering analyzes, and the dynamical correlation coefficient.

The data set is generated as in the preceding subsection. These include 100 random curves sampled from three different models. Each random curve passes 30 time points. The first and second groups of profiles consist of down-regulated and up-regulated genes respectively over the time course [0, 100]. The gene profiles in the third group are periodic curves without upward or downward trend. Table 3.2 lists the final clusters resulting from the two similarity measures.

Clustering quality with Pearson's correlation coefficient is much lower than that with dynamical correlation. Pearson's correlation does not separate clusters 1 and 3 and also fails to distinguish between clusters 2 and 3, even though it is capable of distinguishing cluster 1 and 2. By comparison, dynamical correlation correctly clusters the first group except for one gene, which is misclassified from the second to

third cluster.

Summarizing, these simulation studies provide a clear insight into the way dynamical correlation takes account of the time course and thus improves the performance of a variety of cluster analyzes of gene expression profiles.

3.4 Clustering gene expression profiles in embryonic stem cells

3.4.1 Data

As mentioned in Chapter 1, the gene expression data for mouse ESC we used is a part of data collection in StemBase designed for investigating stem cell differentiation. We selected three ESC cell lines: J1, R1 and V6.5. These cell lines were analyzed at the following 11 time points: 0 hour, 6 hours, 12 hours, 18 hours, 24 hours, 36 hours, 2 days, 4 days, 7 days, 9 days and 14 days. To reduce the variability between microarray platforms, 3 replicates were analyzed for each time point.

The goal is to cluster the genes that manifest the most significant differential over the time course as apparently driving, repressing or responding to stem cell differentiation in the early development period. Using the statistical procedure SAM, the first 5 time points (from 0 hours to 24 hours) can be seen to best reflect the early period of ESC differentiation. We thus select these time points for cluster analysis. A superficial characterization of the transcriptional function of genes is quite intuitive: either up-regulation or down-regulation. However, in a time-varying system, a typical gene profile often reflects a more complicated pattern. Rather than assigning pre-defined patterns to gene profiles, we first cluster them over the time course and then describe the pattern shared by the genes in a cluster.

The number of genes whose expression will register on a microarray is very large, but most of them are not expressed over the time course we are interested in. More-

over, since only a small group of genes participate in early cell differentiation, we need to select only those genes that undergo significant changes in expression over the pertinent time course. Thus we first normalize the data through a logarithmic transformation. Second, we detect genes whose expression is significantly differently expressed over the time course. In this chapter, using 524 genes detected in chapter 2 as the clustering data source.

3.4.2 Clustering results

We use the proposed dynamical correlation based distance measures to cluster these 524 genes. The agglomerative clustering analysis results in final 4 clusters according to the pseudo t^2 statistics. In order to further validate the clustering performance, we check the percentages of molecular functional genes in original and post-clustered groups based on Gene Ontology (GO) [Ashburner M. et al. 2000], which gathers comprehensive information on classified molecular functions. To do GO term analysis, we used the public online software FatiGO, which allows for quick and convenient interactive querying. According to clustering theory, an efficient clustering method should create well separated gene groups according to a measure. Molecular function of gene is one of such measures. Based on this, the percentages of some molecular functional genes should be significantly different across different clusters.

To compare and validate the clusters biologically, we checked the proportion of genes with various molecular functions in the original and clustered groups based on Gene Ontology (GO), which gathers comprehensive information on various classes of molecular function. For this, we consulted the public online software FatiGO, which allows for quick and convenient interactive querying. For our clusters to be biologically interpretable, the percentages of molecular functions should be significantly different across different clusters. Obviously, a better clustering method would result to more distinct groups with more significantly different biological processes or

Functions	All	Cluster1	Cluster2	Cluster3	Cluster4
Metal ion binding	23.98	26.47	18.75	25.77	29.17
DNA binding	23.98	23.53	12.50	24.74	35.42
RNA binding	18.37	8.82	28.12	21.65	16.67
purine necleotide binding	16.84	11.76	18.75	18.56	12.50
cation binding	14.80	17.65	6.25	13.40	22.92
transferase activity	10.71	14.71	6.25	9.28	10.42
N	524	92	93	228	111

Table 3.3: Distribution (%) of Genes with Various Molecular Functions in Four Clusters

Index	Term	p value
GO biological process: level 5	Macromolecule biosynthetic process	0.023
GO biological process: level 6	Translation	0.012
GO molecular function: level 3	Structural constituent of ribosome	0.001

Table 3.4: K-means Method Using Pearson Correlation:Significant GO Item Using Fisher's Exact Test to Compare Cluster 2 with Rest of Genes

molecular functions. We ran FatiGO over four clusters as well as the entire set of genes, as the baseline for comparison purpose. The table 3.3 lists the FatiGO output we obtain.

To further investigate significant biological patterns and compare clustering performance, we use Fisher's exact test in FatiGO, to search all significant different biological functions between one cluster with the rest genes. Tables 3.4-3.6 list all the significant items for Pearson and dynamical correlation distance using k-means method.

The result shows the performance of dynamical correlation is superior to that of Pearson Correlation, mainly supported by the following evidences:

1. The clusters based on Pearson Correlation only show significant different items at cluster 2. Relatively, the K-means method using dynamical correlation has created more distinct clusters, shown in Tables 3.5 and 3.6, cluster 2 and 1

Index	Term	p value
GO molecular function: level 3	Nucleic acid binding	0.002
GO molecular function: level 4	RNA binding	0.001
GO molecular function: level 3	Structural constituent of ribosome	0.001
GO biological process: level 6	Translation	0.029

Table 3.5: K-means Method Using Dynamical Correlation:Significant GO Item Using Fisher's Exact Test to Compare Cluster 2 with Rest of Genes

Index	Term	p value
GO biological process: level 4	Alcohol metabolic process	0.0132
GO biological process: level 4	Lipid metabolic process	0.018

Table 3.6: K-means Method Using Dynamical Correlation:Significant GO Item Using Fisher's Exact Test to Compare Cluster 1 with Rest of Genes

both have significant items. Cluster 2, in particular, consists many genes which are significantly distinct from others in many molecular function and biological process components.

2. The two significant items: structural constituent of ribosome and translation in cluster 2 showing in table 3.4 have been contained in cluster 2 resulted from dynamical correlation based method. By comparison, the important items nucleic acid binding and RNA binding are not found in any cluster created by Pearson Correlation based method.

3. Table 3.6 shows that dynamical correlation based method can create cluster 1, which is highlighted by alcohol and lipid metabolic process, but we cannot find such items from Pearson Correlation based clusters.

3.4.3 Biological interpretation

Genes belonging to the same cluster may play parallel or interacting functional roles in cells. The distribution patterns in Table 3.3 may suggest the common features

within each cluster, even though the percentage of any one functional type does not predominate. Cluster 1 has the lowest percentage of RNA binding genes (only 8.8%). From Figure 3.1, we see that the expression pattern of most of this cluster follows a gradual up-regulated trend before the fifth time point (24 hours) and then exhibits an intense down-regulated pattern. Since RNA binding is closely related to the regulation of translation, it is reasonable to think that the up-regulated gene group mainly participates in other genetic activities, such as nucleotide binding and transferase activity.

For example, cluster member Csnk1e is a nucleotide binding gene that participates typical signaling pathways such as TGF-beta signaling, Wnt Signaling and Hedgehog signaling, all known to be associated with stem cell differentiation.

The feature that most distinguishes cluster 2 is its highest percentage of RNA binding and lowest percentage of DNA binding genes. Further exploration of this cluster reveals that many of its genes are RNA polymerase II transcription factors. Another feature is that the percentage of cation binding is very low compared with other clusters. We may conclude that the second gene cluster is largely concerned with the initiation of transcription by RNA polymerase II and thus with the synthesis of proteins. For example, Utf1 (undifferentiated embryonic cell transcription) is classified in cluster 2, and this gene is a stem cell marker with RNA polymerase II transcription factor activity [Nishimoto M. et al. 1999]. It plays an important role as transcriptional co-activator and is mainly expressed in pluripotent embryonic stem cells.

The third cluster includes the highest number (228) of genes, and its GO terms pattern similarly to the random sample in the first column of Table 3.3, indicating a higher percentage of nucleic binding capabilities (25% DNA binding and 22% RNA binding). Figure 3.1 shows that most group members appear follow a down-regulated pattern during early differentiation. Of the genes involved in early stem cell differentiation, this cluster represents the majority trend.

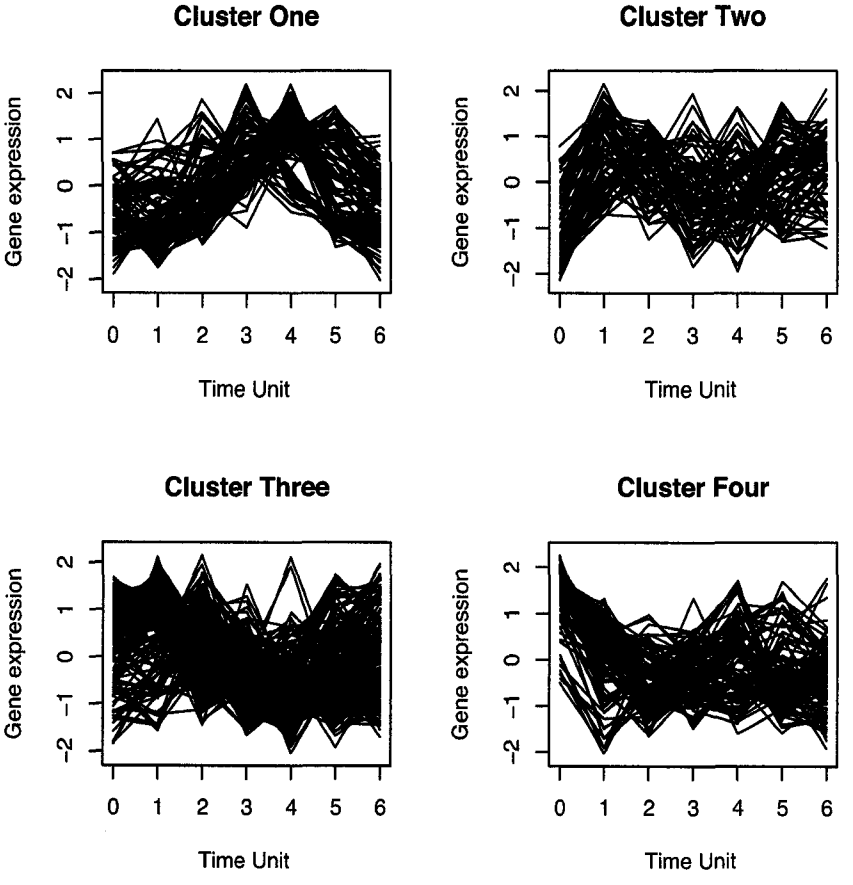

Figure 3.1: Clusters of Stem Cell Gene Expression Time-Course Profiles. Time Units are 0h, 6h, 12h, 18h, 24h, 36h and 48h, though only the first five (0 - 24h) were used for clustering

The fourth cluster expresses a pattern opposite to that of cluster two, as shown in Figure 3.1. In particular, the steep down-regulated pattern occurs within the first six hours. By checking the GO term percentage levels in Table 3.3, we verify that the genes with DNA binding capabilities follow a very different pattern from that of cluster 2. For example, Klf4 (The zinc finger transcription factor) is classified in cluster 4.

It is known that the differentiation of murine embryonic stem cells is promoted by SOCS-3 and inhibited by Klf4 [Li Y. et al. 2005]. Zfp42 (Rex1), which is highly expressed in embryonic carcinoma cells and embryonic stem cells [Rosfjord E. et al. 1994], is also in cluster 4. Upon differentiation with retinoic acid, the expression level of the Rex-1 gene down-regulates rapidly. In general, several of the genes, such as Nanog, that maintain pluripotency in ES cells, are found in cluster 4.

3.5 Discussion

This chapter proposes a dynamical correlation coefficient as the basis for cluster analysis, specifically applicable to gene expression profiles expressed as random curves. An advantage of this similarity measure is that it is sensitive to the time sequence of the observations and step length. Dynamical correlation can serve as input to many classical clustering methods such as K-mean and linkage clustering techniques. In addition, the presented approach is capable of handling the case of uneven time step size in time course clustering analysis and making use of repeated measures to minimize random errors at each time point.

We have mentioned that the method can make use of repeated measures to minimize the random errors at each time point. Suppose $g^{(j)}(t_s)$, $f^{(j)}(t_s)$ $(j = 1, 2..r)$ are replicates for gene f and g respectively at the time point t_s, the bootstrap approach can be described as following:

- Bootstrap B times to select $f^{(p_k)}(t_s)$ and $g^{(q_k)}(t_s)$, where, $p_k \in \{1, 2, ..r\}$ and $q_k \in \{1, 2, ..r\}$ are the gene expression levels (for gene f and g) at the time point t_s. $(k = 1, 2, ..B)$.
- For each $k = 1, 2, ..B$, compute $\mid f_k'(t_s) - g_k'(t_s) \mid$ using formula (3.2.6)-(3.2.9)
- Estimate $E \mid f'(t_s) - g'(t_s) \mid$ by $\frac{1}{B}\sum_{k=1}^{B} \mid f_k'(t_s) - g_k'(t_s) \mid$

Chapter 4

A customized class of functions for modeling and clustering gene expression profiles in embryonic stem cells

This chapter is based on the paper accepted in Brazilian Symposium on Bioinformatics BSB 2008, Lecture Notes in Bioinformatics, Springer. I wrote a complete draft of this chapter. Professor Miguel Andrade verified and expanded upon my interpretations of the gene ontology search results. Professor David Sankoff suggested some of the additional research directions and helped organize and reformulate the exposition.

4.1 Introduction

Cluster analysis of gene expression profiles over a time course often treats each sampling time as one of the dimensions of a multivariate variable, reducing the problem of clustering gene trajectories to one of clustering multivariate observations. Since the object of gene expression clustering is to detect common trajectories, however, this "static" multivariate characterization loses power and precision, since the temporal order plays no role in the analysis; indeed the times can be permuted in any order without changing the results of the cluster analysis.

Gene expression profiles during cell growth, differentiation, tissue activation of various kinds, and during cyclical changes tend to follow well-defined patterns and are often highly autocorrelated. These lead to the incorporation of more dynamic" considerations into clustering procedures. Model-based clustering techniques [Fraley C. et al. 1998] incorporate dynamical features by implementing formal statistical modeling and parametric models in preference to pure data mining.

Some model-based clustering methods for dynamical data have been proposed [Aach J. et al. 2001], [Brumback BA. et al. 1998], [Ramoni MF. et al. 2002]. Here a group of genes following the same probability model fall into the same cluster, taking into account autoregressive and other random effects. For example, [Wu FX., et al. 2005] discussed autoregressive models for clustering time course gene expression data. Spline trend based models [Bar-Joseph Z. et al. 2002], [Luan Y. et al. 2003] apply cubic spline and B-spline mixture models to analyze gene expression time series data.

Purely autoregressive models, however, only consider autocorrelative structure without explicitly taking account other properties of functional interactions among genes. Although a p-order autoregressive model may explain a p-order polynomial time trend effect, it fails to match exponential time growth effects such as $exp(\lambda t)$ or periodic effects such as $sin(\omega t)$. Spline techniques can represent data trends but still remain essentially black boxes [Chudova D. et al. 2004] with respect to genetic

functions. [Chudova D. et al. 2004] suggests a novel clustering technique based on gene functional curves, making use of plausible biological models for gene expression dynamics.

4.1.1 The proposed model

Here we treat a gene profile over a time course as composed of the following components:

1. impact of related genes or other regulation mechanisms (modeled as a dynamical system of differential equations),
2. autoregressive factors representing autocorrelation structure and feedback or loop effects,
3. random components allowing for heterogeneity among the individual genes in each cluster.

This method is flexible in assigning both dynamic and mixed functional components into an integrated time series model with random mixtures. We note that by appropriately incorporating randomness into the model we can avoid the tendency to generate clusters based purely on chance properties of the data.

4.2 The mixture time series model for functional curve clustering

We denote an expression profile (or observation vector over time course) of an individual gene g_i as $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$. In this notation T refers to the total number of time points sampled over a time course. Our objective is to cluster target genes into distinct clusters in terms of the similarities of gene functional behavior. Genes are

grouped together on the basis of the form of their expression profiles (curve shapes) rather than similarity in Euclidean distance. The model-based clustering analysis assumes that the genes in every cluster will perfectly fit an underlying mixture time series model, where any gene profile Y_i is considered as an observation of a functional curve following the probability model:

$$\begin{aligned} Y_{it} &= \text{Gene Network Interactions} + \text{Auto Effects} \\ &+ \text{Time–dependent Random Effect} + \text{Noise} \end{aligned} \tag{4.2.1}$$

If all target genes $g_i, i = 1, 2, \ldots, N$ satisfy the model above, then the $Y_{it}, i = 1, 2 \ldots, N$ are generated by the following model:

$$Y_{it} = f^{(c)}(t|\Phi^{(c)}) + \sum_{j=t-p}^{t-1} \alpha_{jc} Y_{ij} + \gamma_c + \varepsilon \quad (t = 1, 2 \ldots, T \text{ and } i = 1, 2, \ldots, N), \tag{4.2.2}$$

where the "within" random effect $\gamma_c \sim \text{normal}(0, \tau_c^2)$ $(c = 1, 2, \ldots, C)$ are independent of each other, indicating that genes in the same cluster have a unified correlation structure, ε is the i.i.d. white noise $\sim \text{normal}(0, \sigma^2)$, the trend curve $f^{(c)}(t|\Phi^{(c)})$ contains the internal function effect of g_i, combined with the other network interactions (or regulatory effect on g_i), and $\Phi^{(c)}$ is the parameter set for the fixed mean curve in cluster c $(c = 1, 2, \ldots, C)$. We have included autoregressive items in the model because we wish to capture feedback or loop effects in genetic pathways.

The key problem here is to determine the form of the function $f^{(c)}(t|\Phi^{(c)})$. We will estimate it with the trend deduced from a genetic transcription model. The B-spline techniques previously employed to deal with this problem [Luan Y. et al. 2003] have no direct biological functional interpretation. Instead we are inspired by Chen's linear transcription model [Chen T. et al. 1999] to elaborate a function $f^{(c)}(t|\Phi^{(c)})$. Chen's model is a nonlinear dynamic system with the following form:

$$\frac{dr}{dt} = Cp - Vr \quad \frac{dp}{dt} = Lr - Up \tag{4.2.3}$$

where r is the mRNA concentration, p is the protein concentration, L contains the translational constants, V is the degradation rate of mRNA and U is the degradation rate of proteins. According to Theorem 1 in [Chen T. et al. 1999], the solution to model (4.2.3) has the form:

$$x(t) = Q(t)e^{t\lambda} \tag{4.2.4}$$

where $Q(t) = \{q_{ij}\}$ is a $2n \times 2n$ matrix whose elements are polynomial functions of t. In [Chen T. et al. 1999] it is stated that a gene system should be stable system and its expression should not have an exponential or a polynomial growth rate, which implies that $Q(t)$ is a constant and $x(t)$ is actually an ordinary exponential function. However, an exponential function is monotonic and cannot fully reflect the wave-like shapes that characterize many gene functional curves. Consequently we do not adopt (4.2.4) directly, though we will utilize features of (4.2.3) and (4.2.4) to develop a more effective functional curve.

To do so, we have studied the functional gene curves of mouse embryonic stem cells (mESC) in the first 24 hours of differentiation (for experimental details and the data bank, see [Sene KH. et al. 2007] and found that gene expression of these genes can be well described by the following hyperbolic function:

$$f(t) = \frac{1}{\exp[b(t-a) + \sqrt{1+(t-a)^2}]} \tag{4.2.5}$$

To see the connection between (4.2.5) on one hand and (4.2.3) and (4.2.4) on the other, we note that the gene's growth rate satisfies:

$$\frac{df(t)}{dt} = -[b + \frac{t-a}{\sqrt{1+(t-a)^2}}]f(t), \tag{4.2.6}$$

where the expression (4.2.6) indicates that this gene system also remains stable, but is more flexible than the one in (4.2.3). Here $\frac{t-a}{\sqrt{1+(t-a)^2}}$ is restricted to the range [-1,1], but can model behaviours more varied than a the constant growth rate.

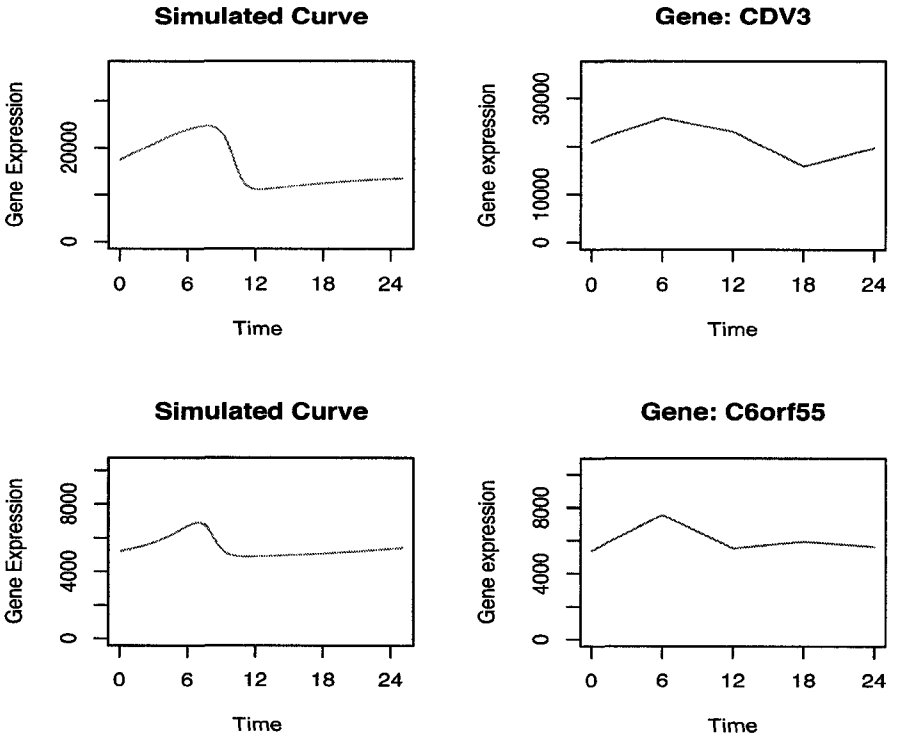

Figure 4.1: Simulated curve *vs.* actual gene curve

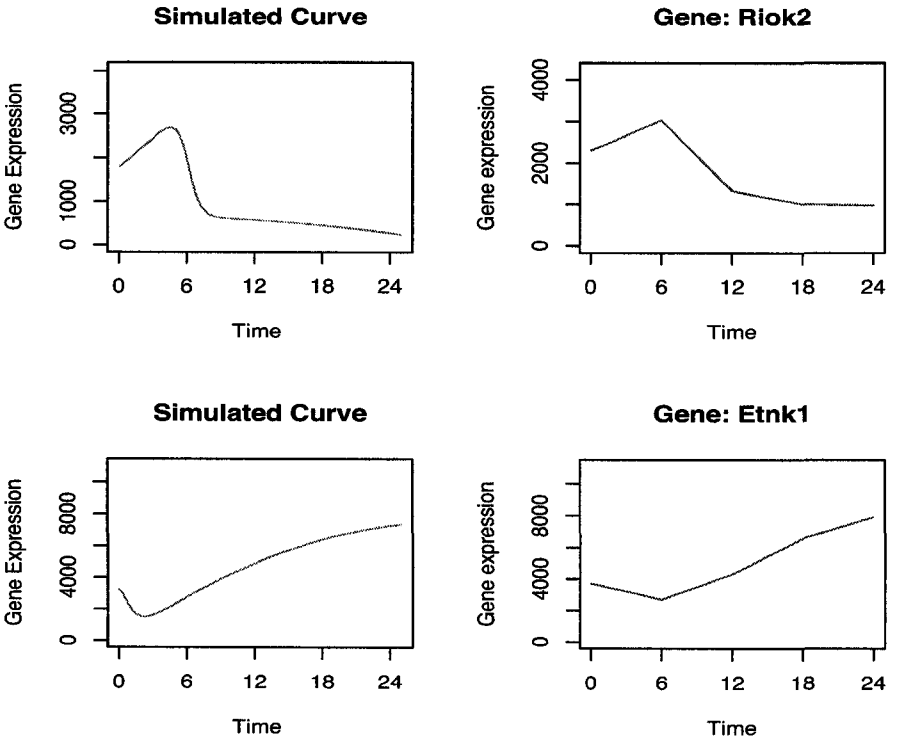

Figure 4.2: Simulated curve *vs.* actual gene curve

In Figure 4.1, the gene profiles (over the time course 0 to 24 hours) of Cdv3 (CDV3 homolog (mouse)) and C6orf55 (chromosome 6 open reading frame 55) are plotted against the corresponding simulated hyperbolic functional curves. In the same way, Figure 4.2 presents the plots for genes Riok2 (RIO kinase 2) and Etnk1(Ethanolamine kinase 1). These genes are highly active over the time course in the early period of stem cell differentiation. These figures illustrate how parameter settings are available to produce functional curve shapes similar to those of actual gene profiles.

For curve fitting purposes, we can include linear and quadratic parts without affecting system stability. Thus, we define gene functional curves as follows:

$$f(t|\beta_l) = \beta^{(1)} + \beta^{(2)} \exp\{-\beta^{(3)}(t - \beta^{(4)}) - \sqrt{1 + (t - \beta^{(4)})^2}\} + \beta^{(5)}t + \beta^{(6)}t^2 \quad (4.2.7)$$

From (4.2.2) and (4.2.7), by setting different curve parameters $\beta_c = (\beta_c^{(1)}, \beta_c^{(2)}, \beta_c^{(3)}, \beta_c^{(4)}, \beta_c^{(5)})$, random effects γ_c and auto regression coefficients α_{jc} $(c = 1, 2, \ldots, C)$ we can determine different gene clusters.

To implement cluster recognition, we assume the following mixture-density model:

$$P(Y|\Theta, \Omega) = \sum_{c=1}^{C} \omega_c p_c(Y|\theta_c), \quad (4.2.8)$$

where the parameter set $\Theta = (\beta_1, \ldots, \beta_C, \alpha_{j1}, \ldots, \alpha_{jC}, \gamma_1, \ldots, \gamma_C)$ $(j = 1, 2 \ldots, p)$ and $\Omega = (\omega_1, \ldots, \omega_C)$ such that $\sum_{c=1}^{C} \omega_c = 1$ and $p_c(Y|\theta_c)$ is the density function generated by the model (4.2.2) defined previously. Given observed samples $Y = (Y_1, Y_2, \ldots, Y_N)$, the log-likelihood expression for the above mixture density is:

$$\log(P(Y|\Theta, \Omega)) = \log(\sum_{c=1}^{C} \omega_c p_c(Y|\theta_c)) = \sum_{i=1}^{N} \log(\sum_{c=1}^{C} \omega_c p_c(Y_i|\theta_c)), \quad (4.2.9)$$

where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ $(i = 1, 2, \ldots, N)$ represents the profile of gene g_i over the time course $t = 1, 2, \ldots, T$. Since it is hard to optimize (4.2.9) by ordinary analytical methods, we will apply the well known expectation maximization (EM)

algorithm [McLachlan GJ. et. 1997] to estimate these parameters. The details of the EM algorithm approach and the corresponding inference technique is given in the Appendix.

4.3 Numerical test

4.3.1 Clustering Results

In this section we illustrate the result of clustering gene profiles for V6.5 mouse embryonic stem cells over the time course [0h, 6h, 12h, 18h, 24h],which represents the early period of mESC differentiation. At the first stage, we select as target genes 419 with high differential expression as detected by the SAM method [Tusher VG. et al. 2001]. A standardization procedure is first applied to all gene profiles before running the clustering program. We arbitrarily choose a small number of clusters, namely $C = 3$, in order to facilitate the evaluation of the method, and set the initial cluster proportion to be $\frac{1}{3}$. Thus, we will optimize (4.2.9) with respect to $\Theta = (\theta_1, \theta_2, \theta_3)$, $\Omega = (\omega_1, \omega_2, \omega_3)$ and σ^2.

Figure 4.3 presents the results of clustering the 419 genes. The three clusters clearly have different patterns over the time course. Briefly, cluster one is basically up-regulated compared to cluster three which has a distinct down-regulated pattern. Cluster two contains the largest number of genes. Note that that these clusters have different "compactness" levels. For instance, cluster three is characterized by the fact that the gene functional curves are more closely intertwined while cluster two is more loosely arranged, indicating a larger intra-cluster variation. This effect is captured by the different random gene effect within the three clusters, a feature of the mixture model. In visually assessing the clusters, we must recall that the curves are not clustered by the overall proximity to each other, but by the similarities of their patterns.

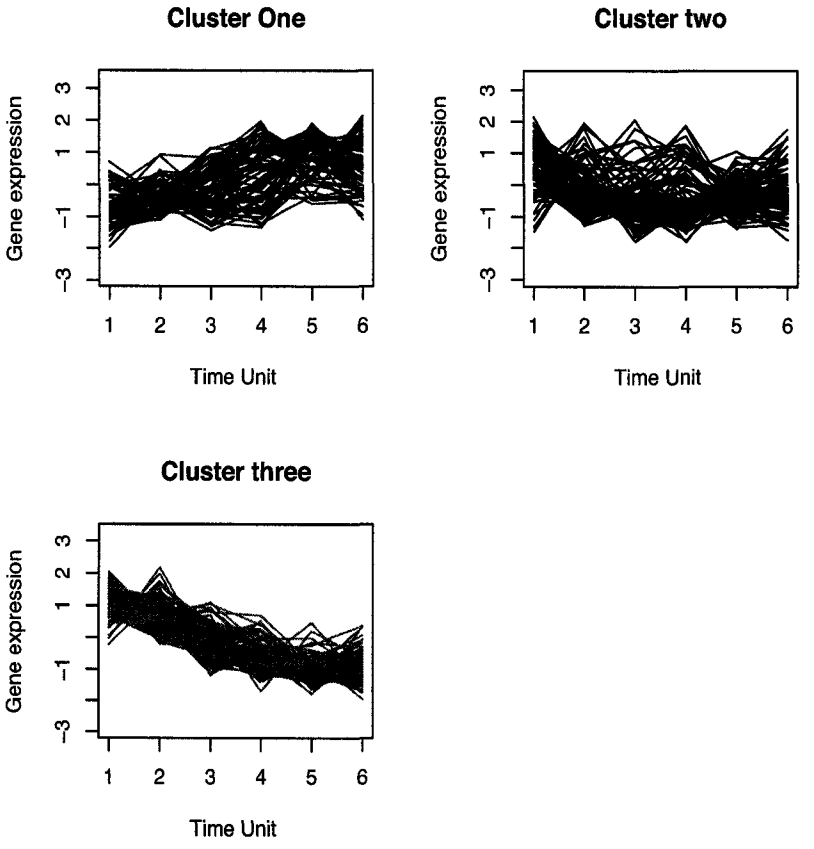

Figure 4.3: Clustering result for gene profiles of V6.5 embryonic stem cells

Biological Functions	Cluster one	Cluster two	Cluster three
RNA metabolic	17.14	40.57	42.65
Cellular lipid metabolic	13.33	2.83	1.47
Cellular protein metabolic	45.71	29.25	35.29
Nucleoside, nucleic metabolic metabolic	25.64	50.91	52.94
Regulation of cellular process	23.29	43.64	27.42
Cellular macromolecule metabolic	43.84	28.18	35.59
Protein metabolic	43.84	30.02	38.24
Transcription	18.57	33.33	20.97
Cellular biosynthetic process	21.92	13.76	27.42
Biosynthetic process	24.66	14.16	28.97
N	94	153	85

Table 4.1: Distribution (%) of genes according to biological process

We choose the order of autoregressive component to be 1 (i.e. $p = 1$) to avoid the increased noise effect associated with larger order components demonstrated in numerical experiments [Wu FX., et al. 2005].

To validate the clustering, we characterized the genes in each cluster by compiling Gene Ontology (GO) terms. We also searched the FatiGO server to compare the three clusters. Table 4.1 displays the distribution of genes featured by different biological processes.

4.3.2 Biological analysis

Cluster one includes 119 genes, of which 94 had associated GO terms, largely with protein metabolism, processing of macromolecules and catalytic functions, such as Sc4mol, the Rpl and Rps family (Rpl4 Rpl13 Rpl14 Rpl23 Rpl41 Rps2 Rps6 Rps12 Rps17 Rpl22 Rps27 Rps28 Rps27l), Uba52, Csnk1e, Otx2, Igfbp2 and Gpi1. These genes follow a generally up-regulated trend during ESC differentiation and participate in ESC metabolism or protein synthesis processes. For example, Otx2 is an early stage murine ESC marker playing a cental role in gastrulation, essential for the early specifi-

cation of the neuroectoderm destined to become fore midbrain [Morsli H. et al. 1999]. Csnk1e, which undergoes a persistent up-regulation pattern over the time course, performs a catalytic function for serine/threonine kinase activity. The Rpl/Rps family is involved in ribosomal structure and hence protein biosynthesis. Note that there was some evidence, discounted by the authors as well as our results here, that Rpl4 and Rps24 are among the genes constituting a unique molecular signature in human ESC [Bhattacharya B. et al. 2004].

There are 198 genes in cluster two, 153 with GO annotations. These genes are involved in RNA metabolism, regulation of cellular process and DNA-dependent transcription. Many key ESC markers such as Nanog, Sox21, Zfp57, Cbx7, Vcl, Dnajb6, Msi2h and Cggbp1 fall into this cluster. They generally usually down-regulated for the first 18 or 24 hours and are up-regulated later. By checking the correlation distance, we found these gene profiles are remarkably similar to the expression patterns of undifferentiated ESC markers such as OCT4, Sox2 and Foxd3. From the point of view of stem cell differentiation, these genes are crucial for stem cell self-renewal and maintenance of the inner cell mass (ICM) of the blastocyst. With the loss of cell proliferation and pluripotency, they undergo a significant and persistent down-regulation pattern [Niwa H. et al. 1998],[Chambers I. et al. 2003].

Many members of cluster two are also related to the cell cycle, cell proliferation or growth. For example, C2ORF29, Cdv3 and Rbbp7 are involved in cell proliferation. Socs3 and Ltbp4 function in the regulation of cell growth. Nmyc1, Cdk2ap1, Nipbl, Cdc34, Mcm5, Nipbl, BC068171 and Ccne1 are cell-cycle related genes. They mediate the progression through the cell cycle, particularly the G1/S transition of the mitotic cell. That these genes are in cluster two agrees with theories [Bilmes JA. 1997], whereby the cells initially derived from a population of stem cells undergo rapid cell division during early differentiation; throughout this period, expression of genes keeping control of the cell cycle or proliferation should be attenuated.

Another 102 genes are found in cluster three, 85 with GO annotations. Com-

pared with cluster one, cluster three is enriched for genes in charge of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, such as Ctbp2, Cdk2ap1, Apex1, Hmgb2, Mybbp1a Ran, Gars, H2afz, Sod2, Nola2, Mki67ip, Etv4, Atp5l, Bzw1, Nr0b1, Klf5, Rbm14, Polr2f, Ankrd17, Lsm3, Zfp36l1, Ddx5, Sfrs1, Rbm3, Cars and Psmc5. This cluster is also associated with microtubule-based movement, signal transduction and biosynthetic processes. These include Arpc4, Lefty1, Ptprf, Stmn1, Cfl1, Trh and Eif4ebp1. Lefty1 is an important "stemness" marker expressed in the left half of gastrulating mouse embryos and involved in the TGF-beta signaling pathway. Ptprf has an intrinsic protein tyrosine phosphatase activity (PTPase) and plays a role in cell adhesion receptor and the insulin signaling pathway. Ptprf has also been identified in human ESC studies. Stmn1 is involved in signal transducing, participates in the regulation of the microtubule (MT) filament and promotes disassembly of microtubules. Trh plays a role in cell-cell signaling and neuroactive ligand-receptor interaction. Arpc4 is related to cytoskeleton and actin filament polymerization.

Since clusters two and three have a similar down-regulated pattern, especially in the first 12 hour period, it is not surprising that some of their members have similar gene functions in signalling, ATP/GTP binding, actin binding and microtubules. For example, the genes Pfn1, Wdr1, Efna2, Lefty2, Ak7, Ptch1, Riok2, Arf6 and Actg1 are classified into cluster two. This is not surprising; maintenance of the pluripotent state of ESC requires intrinsic signalling as well as extrinsic environmental elements, involving microtubules. It is known that environmental factors such as cell-surface receptors and cytokines play an important role in the maintenance of stem cell functions [Arias AM. et al. 2002]. Transcription profiles of this type generally take a persistent down-regulated pattern during differentiation. Examples include the actin binding related gene Ptprf and receptor activity-associated gene Tagln, which decrease sharply throughout the first 24 hours in ESC differentiation.

Table 4.2 uses a t-test to evaluate the over-representation of various kinds of GO terms in each cluster. The analysis in Table 4.2 confirms and deepens that of Table

Cluster	Biological Functions	In cluster	In genome	P-value
I	primary metabolic process	60	5694	0.000107
	cellular protein metabolic process	34	2421	7.44E-05
	structural constituent of ribosome	14	147	4.41E-10
Total		94	14456	
II	gene expression	57	2452	4.21E-08
	RNA metabolic process	46	2082	1.59E-05
	transcription, DNA-dependent	35	1704	0.00215
	ribonucleoprotein complex biogenesis	11	174	0.000302
	translational initiation	5	40	0.00266
	glycolysis	5	42	0.00326
Total		153	14456	
III	gene expression	33	2452	2.02E-05
	translation	13	369	2.50E-05
	amino acid,derivative metabolic process	9	264	0.00111
	ribonucleoprotein complex	10	382	0.00267
Total		85	14456	

Table 4.2: Over-represented GO terms for each cluster from GOstat

4.1. For example, the high value for RNA metabolism in cluster 2 possibly results from the selection of genes involved in transcriptional regulation. The higher proportion of cellular protein metabolic process in cluster 1, amino acid derivative metabolic processes and ribonucleoprotein complex, in cluster 3, are all closely related to protein biosynthesis. This similarity possibly originates from genes involved in translation process(cluster 3). We also found interesting to see five glycolitic enzymes in cluster 2, and a number of genes involved in the formation of RNA-protein complexes in clusters 2 and 3. In the case of cluster 2, this includes five genes involved in translation initiation.

4.4 EM algorithm of mixture model

4.4.1 Inference of the log-likelihood function for the mixture model

For convenience,we only consider the first order autoregressive item (i.e. $p = 1$ in model (4.2.2)). According to model (4.2.2), we infer that if the gene g_i belongs to cluster c then its expression profile $Y_i = (y_{i1}, \ldots, y_{iT})$ at time point t satisfies the following conditional probability model:

$$P(y_{it}|\theta_c, Y_{i(t-1)}) = \text{Normal}((f^{(c)}(t|\Phi^{(c)}) + \alpha_c y_{i(t-1)}), {\gamma_c}^2 + \varepsilon^2) \tag{4.4.1}$$

Thus, the log-likelihood for Y_i over time course $t = 1, 2, \ldots, T$ is

$$L_c(Y_i|\theta_c) = \log[P(y_{i1}|\theta_c) \prod_{j=2}^{T} P(y_{ij}|\theta_c, y_{i(j-1)})] \tag{4.4.2}$$

From expression (4.4.1), (4.4.2) can be finally expressed as

$$L_c(Y_i|\theta_c) = C_0 - TLog(\sigma_i) - \frac{1}{2\sigma_i}[(y_{i1} - f(1|\beta_c) - \mu_0)^2 + \sum_{j=2}^{T}(y_{ij} - f(j|\beta_c) - \alpha_{(c)} y_{i(j-1)})^2] \tag{4.4.3}$$

where the standard deviation of Y_i, $\sigma_i = \sqrt{\gamma_c^2 + \varepsilon^2}$, C_0 is a constant and μ_0 represents the mean at the initial time point. The log-likelihood of the mixture model of total observed samples can be expressed as

$$L(Y|\Theta, \Omega) = \sum_{i=1}^{N} \log(\sum_{j=1}^{C} \omega_j P_j(Y_i|\theta_j)) \tag{4.4.4}$$

where $P_j(Y_i|\theta_j)$ is identified by $\exp(L_j(Y_i|\theta_j))$, and ω_j, $(j = 1, \ldots, C)$ are the unknown cluster membership parameters such that $\sum_{j=1}^{C} \omega_j = 1$.

4.4.2 EM Algorithm Approach

The optimization of (4.4.3) can be carried out by using an EM algorithm. To see the specific computational steps, the reader can refer to [Bilmes JA. 1997].

1. Set initial parameters $\Theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_C^{(0)})$ for the given number of clusters C and cluster proportion $\Omega^{(0)} = (\omega_1^{(0)}, \ldots, \omega_C^{(0)})$.

2. Define cluster membership (a random variable) l, where $l \in \{1, 2, \ldots, C\}$ and $l = c$ if gene g_i belongs to cluster c. For $k = 0, 1, 2, \ldots$ repeat the following iterative steps until a given threshold is reached.

3. E-step: Calculate the posterior of the cluster membership l given Y_i, $\Theta^{(k)}$ and $\Omega^{(k)}$ at the k^{th} iterative step from the following procedure:

$$P(l|Y_i, \Theta^{(k)}, \Omega^{(k)}) = \frac{\omega_l^{(k)} P_l(Y_i|\theta_l^{(k)})}{\sum_{j=1}^{C} \omega_j^{(k)} P_j(Y_i|\theta_j^{(k)})} \tag{4.4.5}$$

4. M-step: Maximize the expected posterior log-likelihood with respect to cluster membership parameters Ω and model parameters Θ given observed Y and current parameter estimates:

$$\omega_l^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} P(l|Y_i, \Theta^{(k)}, \Omega^{(k)}) \tag{4.4.6}$$

and

$$\text{Maximize } F = \sum_{l=1}^{C} \sum_{i=1}^{N} [L_l(Y_i|\theta_l^{(k+1)}) P(l|Y_i, \Theta^{(k)}, \Omega^{(k)})] \tag{4.4.7}$$

where (4.4.7) can be maximized by non-linear optimization techniques such as BFGS or the conjugate gradients method. Note that solving (4.4.7) is much easier than directly optimizing (4.2.9).

4.5 Discussion

In this chapter we use cluster analysis to infer gene expression patterns and related biological characteristics during stem cell differentiation.

Since gene expression data of stem cell differentiation are time dependent and have different correlation structure, the traditional clustering methods in which dynamics of gene expression are not accounted for cannot perform as well for time-course

gene expression data. The model based clustering method proposed in the chapter is trying to overcome the shortcomings above.

The most important feature of the presented method is that it take many dynamics factors into consideration: genetic biological function induced by other regulators, random effects producing heterogeneity within gene clusters, and autoregressive components defining the loop or feedback impact. In addition, the proposed methods are flexible in the sense that it can incorporate a priori information into the models as they have a solid probabilistic foundation.

Accordingly, many dynamics factors will make the approach parameter-rich (when cluster number is large), and genetic biological function is nonlinear, which might cause higher computational cost. Nevertheless, the numerical test shows that EM algorithm and BFGS method are capable of handling these problems very well.

Following some procedures of clustering validation, we can select 'optimal' cluster number. Since the proposed model is a type of partition clustering approach, one can evaluate the quality of a clustering performance by means of silhouette index [Rousseeuw P.J. et al. 1987]. The silhouette index assesses the ratio of inter-cluster separation and intra-cluster similarity. For each object i in a data set containing n objects, its silhouette width $s(i)$ is defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{4.5.1}$$

Where $a(i)$ is the average distance of object i to other objects in the same cluster. $b(i)$ is the average distance of object i to other objects in its nearest neighbor cluster and can be seen as the dissimilarity between object i and the nearest neighbor cluster to which it does not belong. Objects with a large $s(i)$ (almost 1) are very well clustered, a small $s(i)$ (around 0) means that the object lies between two clusters, and objects with a negative $s(i)$ are probably placed in the wrong cluster. The average $s(i)$ across all objects:

$$S = \sum_{i=1}^{n} s(i) \tag{4.5.2}$$

is called the silhouette index of the clustering. The value of S reflects the overall quality of the clustering and has range $[-1, 1]$. A larger value of S indicates a better overall clustering quality. Silhouette index is often used to validate K-means method in which the similarity distance can be Pearson correlation or Euclidean distance. For model based clustering, it can be defined as $1/p_{ij}$, where p_{ij} is the cluster membership parameters indicating the probability of gene i belongs to cluster j. One could also use External Indices, which define a measure of the agreement between two partitions of the same data set. In the clustering literature, measures of agreement between two partitions are referred to as external indices. Several such indices have been proposed [Dudoit S. et al. 2002].

For model based clustering method, various information criterion can be applied to deal with parameter-rich and cluster number problem. For instance, Akaike's information criterion (AIC) [Burnham K.P. et al. 1998], Hurvich and Tsai (AICC) [HURVICH CM. et al. 1989], Schwarz criterion (BIC)[Schwarz G. 2008], and Spiegelhalter Deviance Information Criterion (DIC)[Spiegelhalter, DJ. et al. 2002]. These Criteria work as the different measures in selecting the best subset of predictors:

$$AIC(\Theta) = -2\log(-L(\Theta|D)) + 2p \tag{4.5.3}$$

$$AICC(\Theta) = -2\log(-L(\Theta|D)) + \frac{2(p+q+1)n}{n-p-q-2} \tag{4.5.4}$$

$$BIC(\Theta) = -2\log(-L(\Theta|D)) + 2p\log(n) \tag{4.5.5}$$

$$DIC(\Theta) = 2E_{\Theta|Y}[-2\log(-L(\Theta|D))] - 2\log[-L(E_{\Theta|Y}(\Theta)|D] \tag{4.5.6}$$

Where, n is the sample size, p is the number of parameters,q is the number of dimension of the observation, $L(\Theta|D)$ is the likelihood function based on observed data D, and Θ is the parameter set. Theoretically, a better model has smaller values of the index above. Since our clustering approach is based on fitting model, the information criterion above could be use to select the best cluster number and parameter set.

Chapter 5

Bayesian Adjusted Quasi-likelihood for generalized linear models

This chapter is based on the paper submitted to Computational Statistics & Data Analysis. I was responsible for the formulation of all of the mathematical problems and their solutions in this work. I also wrote a complete draft of this chapter. Professor David Sankoff suggested some of the additional research directions and helped organize and reformulate the exposition.

5.1 Introduction

The Quasi-Likelihood Function (QLF) [Wedderburn RWM. et al. 1974] for solving the generalized linear model (GLM) only requires knowledge of the relationship between the mean and variance of the observations rather than the specific form of the density function. The QLF retains many properties of the log-likelihood function (e.g., consistency and asymptotic properties [McCullagh P. 1983], while generating workable estimators. In this chapter, we suggest an extensions to QLF to the situation where the mean is treated as a random variable. The sample mean can then provide us with useful information to adjust parameter estimates in the traditional QLF approach.

In this chapter we propose two extensions to QLF analyses of GLMs, through what we call the Bayesian Adjusted Quasi-Likelihood (BAQL) function. The innovation here is the borrowing of Bayesian techniques on the sample mean, as follows:

1. The sample mean is treated as a random variable with a prior distribution.

2. Based on the original QLF and Newton-Raphson iterative steps, we adjust the QLF and the estimated model parameters. Specifically, the posterior function (or density) is updated at each iterative step, and conversely QLF is adjusted by learning from the prior density.

3. The parameter estimates are obtained from maximizing BAQL.

In GLM, the observed samples are assumed to come from an exponential family with means $\mu_i = E(y_i), (i = 1, 2, \cdots, n)$. In the classical quasi-likelihood method, the μ_i are considered to be fixed and to satisfy $g(\mu_i) = X_i'\beta$, where $g(\cdot)$ is a link function, $X_i', (i = 1, 2, \cdots, n)$ is a co-variate vector and β is an unknown parameter vector to be estimated. By comparison, the BAQL approach views $\mu_i, (i = 1, 2, \cdots, n)$ as the observed values of a random variable μ, where μ follows some prior probability distribution.

We can illustrate the advantages of this approach through a concrete example. Suppose we are studying the relationship between the gender ratio and salary offered among the companies in a city. We collect samples from each company and fit the sample data (observations of gender and salary) with a GLM (such as the logistic model). These samples, however, may not be drawn evenly, for instance, 70% of the samples could be drawn from companies with the female:male ratio 3 : 1, and only 30% of the samples collected from companies with the ratio 2 : 5 (or less). The traditional quasi-likelihood method would ignore this bias, since it allocates an equal weight to each of the samples y_i, and thus fails to take information about the distribution of μ_i into account.

The implementation of the BAQL approach for solving GLMs problem is straightforward and is introduced in the following section.

The organization of the chapter is as follows: in Section 5.2 we present the Adjusted Quasi-Likelihood Function (A-BAQL). Under the random variable assumption, the densities of $\mu_i, (i = 1, 2, \cdots, n)$ are approximated, and are to be utilized in turn for the adjustment of original quasi-likelihood function. Therefore, rather than treating all observations as equal, we adaptively update the weighted quasi-likelihood at each iterative step. The parameters are then estimated by maximizing Adjusted Quasi-Likelihood until convergence or until some threshold is reached. In Section 5.3 we propose a second BAQL approach, namely Two-Stage Adjusted Quasi-Likelihood (T-BAQL). Here both the traditional QLF and the prior density μ_i are maximized at the same iterative step, i.e., two estimated outcomes are generated at each step. The final shrinkage parameter estimate is obtained from the inverse variance weighted average of the two estimates. We study the density estimate required in our methods through a locally weighted density procedure and bandwidth choice in Section 5.4. We present a simulation study to illustrate the proposed technique in Section 5.6, where the BAQL approach is compared to the classical QLF method in terms of flexibility, bias of estimates and prediction accuracy. The asymptotic normality

properties of parameter estimates and bandwidth selection are discussed in Section 5.7. Even though the methods presented here are not directly applied to study stem cell differentiation, they could be potentially utilized to infer gene regulatory network when gene expression data are not normally distributed. For instance, in boolean network and differential/difference equation system, the gene expression levels could be binary or categorical. The issue in detail will be discussed in section 5.8.

5.2 Adjusted Iterative Weighted Quasi-likelihood Function (A-BAQL) for GLMs

5.2.1 Rationale of the A-BAQL approach

A GLM can be formulated [Lindstrom MJ. et al. 1988]:

$$g(E(y_i)) = X_i'\beta \quad or \quad g(\mu_i) = X_i'\beta \tag{5.2.1}$$

where $\mu_i = E(y_i)$ $(i = 1, 2 \ldots, n)$ and $g(.)$ is the link function. X_i' $(i = 1, 2 \ldots, n)$ are the fixed m-dimensional predictor co-variates and y_i $(i = 1, 2 \ldots, n)$ is the i^{th} observation, and β is the p-dimensional parameter vector to be estimated. To avoid confusion in the later discussion, we denote the constant vector $X_i'\beta$ as c_i $(i = 1, 2 \ldots, n)$. The conventional Quasi-Likelihood technique for GLMs problem is to tackle the following optimization problem with respect to β:

$$\text{Maximize} \quad Q_y(y|\beta) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2(t)} dt = \sum_{i=1}^{n} \int_{y_i}^{g^{-1}(c_i)} \frac{y_i - t}{\sigma^2(t)} dt = \sum_{i=1}^{n} q_{y_i} \tag{5.2.2}$$

where $\sigma^2(.)$, the variance of observations is a function of mean and $q_{y_i} = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2(t)} dt$ is defined as the "Quasi-density"of y_i. In order to implement the BAQL approach, we make the following assumptions:

(S1) The y_i $(i = 1, 2 \ldots, n)$ follow the same probability distribution (a mixture distribution of exponential family) with the mean μ and variance σ^2, i.e., y_i are

the i.i.d. samples of a random variable $y \sim F_y(y|\mu, \sigma^2)$.

(S2) The μ_i $(i = 1, 2 \ldots, n)$ are n observations of a random variable μ, i.e., μ_i are treated as the i.i.d. samples from the prior distribution $f_\mu(\mu)$ of μ, where $E(\mu_i) = \mu$ $(i = 1, 2 \ldots, n)$.

(S3) There exists a random variable u such that $E(y_i|u) = \mu_i$, where y_i is the observed value of y given $\mu = \mu_i$ $(i = 1, 2 \ldots, n)$. Here, it is obvious that $E(E(y_i|u)) = E(\mu_i) = \mu$.

(S4) The independent random variables $v_i = g(\mu_i)$ $(i = 1, 2 \ldots, n)$ follow a probability distribution $f_v(v)$.

(S5) The pairs (y_i, μ_i) $(i = 1, 2 \ldots, n)$ are independent each other.

Under the assumptions above, we can use the following 'Quasi' function:

$$q_y(y|v) = \int_y^{g^{-1}(v)} \frac{y - t}{\sigma^2(t)} dt \tag{5.2.3}$$

to approximate the log-density of y, given random variable $v = g(\mu)$. If we have obtained the prior density of v , i.e., $f_v(v)$, then the following:

$$q_y^*(y, v) = \left(\int_y^{g^{-1}(v)} \frac{y - t}{\sigma^2(t)} dt \right) f_v(v) \tag{5.2.4}$$

can be regarded as the weighted log-density approximation of y at v. Thus, the weighted likelihood of $y = (y_1, y_2, \ldots, y_n)$ at $v = (v_1, v_2, \ldots, v_n)$ can be expressed as:

$$L(y, v) = \sum_{i=1}^{n} \left(f_v(v_i) \int_{y_i}^{g^{-1}(v_i)} \frac{y_i - t}{\sigma^2(t)} dt \right) = \sum_{i=1}^{n} f_v(v_i) q_{y_i} \tag{5.2.5}$$

We further standardize $f_v(v_i)$ by $w_i = f_v(v_i) / \sum_{i=1}^{n} f_v(v_i)$. Then β is estimated from

$$\text{Maximize} \quad Q_1(\beta) = \sum_{i=1}^{n} w_i q_{y_i} \tag{5.2.6}$$

5.2.2 Statistical interpretation

We have mentioned that in A-BAQL approach, rather than maximizing classical QLF: $\sum_{i=1}^{n} q_{y_i}$, we maximize the weighted QLF: $\sum_{i=1}^{n} w_i q_{y_i}$ at $v = (v_1, v_2, \ldots, v_n)$. Note that 5.2.5 can also be treated as the estimate of the expectation of the likelihood of $y = (y_1, y_2, \ldots, y_n)$ over v, i.e.

$$L(y, v) = \sum_{i=1}^{n} f_v(v_i) q_{y_i} \doteq E_v(L^*(y|v)) \tag{5.2.7}$$

where $L^*(y|v)$ is the log-likelihood of y at v. Since 5.2.7 contains additional information on the density of μ, we can expect it to give a better estimate than that of the classical QLF method.

Comparing expressions (5.2.2) and (5.2.6), we find the following differences: (i) the conventional QLF method tries to maximize the average q_{y_i} with respect to β whereas (5.2.6) optimizes the weighted average of q_{y_i}; (ii) the conventional QLF maximizes the approximation of log-likelihood of y when $\mu_i = c_i = X_i^T \beta$ $(i = 1, 2 \ldots, n)$ is treated as a constant but the A-BAQL approach maximizes the approximation of the expectation of log-likelihood of y over $v = g(\mu)$ when μ is regarded as a random variable.

5.2.3 Computational procedure for A-BAQL

Note that we have no idea of v_i (or μ_i) before β is known, so we cannot initiate the A-BAQL approach by itself. However, if β is estimated at each iterative step (such as in the Newton-Raphson algorithm [Lindstrom MJ. et al. 1988] [Press, W. et al. 1992] for the quasi-likelihood procedure), then v_i can be estimated and used to update the quasi-likelihood at the same time. The advantage of this approach can be appreciated if we consider that in the classical QLF method each iterative step uses an estimate of β (i.e., $\hat{\beta}$) from the previous step. If one believes $\hat{\beta}$ is subject to further improvement,

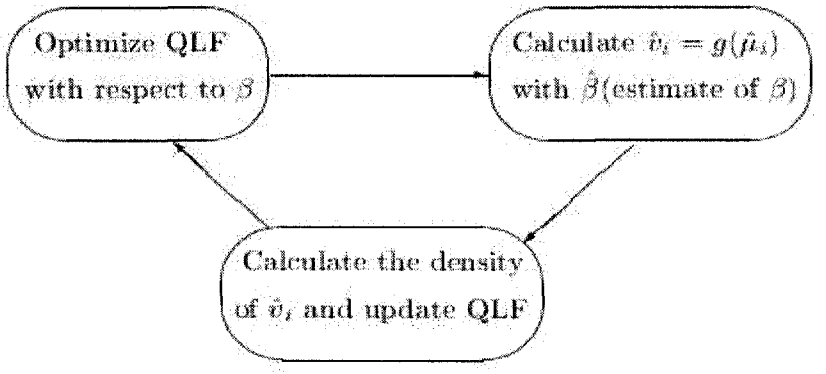

Figure 5.1: The procedure of A-BAQL method for GLMs

this is sufficient reason to believe that $\hat{v}$ or $\hat{\mu}$ (the estimate of v or μ) can provide further help at each iterative step since, as we have stressed, A-BAQL considers further information on the density of random variable v. The procedure of the A-BAQL method for GLMs is schematized in Figure 5.1 and detailed as:

A-BAQL Algorithm

(A1) Set initial index $j = 1$, choose some $\hat{\beta}^{(0)}$ and define the initial A-BAQL function to be the conventional quasi-likelihood defined in (5.2.2).

(A2) Apply the Newton-Raphson iterative algorithm to the A-BAQL function and obtain $\hat{\beta}^{(j+1)}$ from $\hat{\beta}^{(j)}$.

(A3) Calculate $\hat{v}_i = g(\hat{\mu}_i)$ $(i = 1, 2 \ldots, n)$ based on $\hat{\beta}^{(j+1)}$ and estimate the probability density $w_i(\hat{v}_i)$ $(i = 1, 2 \ldots, n)$.

(A4) Adjust the A-BAQL function using w_i $(i = 1, 2 \ldots, n)$ (see expression (5.2.6)).

(A5) Set $j = j + 1$ and go to (A2) until some given threshold is reached.

5.3 Two-Stage Adjusted Quasi-Likelihood (T-BAQL) Approach

5.3.1 Inference of T-BAQL approach.

The main idea of the T-BAQL approach is that β is iteratively estimated through maximizing of the QLF of y_i (see (5.2.2)) as well as the QLF of v_i, whose observed values are $c_i = X_i'\beta$ $(i = 1, 2 \ldots, n)$:

$$Q_v(v_1, v_2, ..v_n|\beta) = \sum_{i=1}^{n} \int_{v_i}^{\theta_i} \frac{v_i - t}{\sigma_1^2(t)} dt = \sum_{i=1}^{n} \int_{g(\hat{\mu}_i)}^{\theta_i} \frac{g(\hat{\mu}_i) - t}{\sigma_1^2(t)} dt, \tag{5.3.1}$$

where (5.3.1) serves to approximate the log-likelihood of c_i $(i = 1, 2 \ldots, n)$, σ_1^2 is the variance function and θ_i is the mean of the random variable v_i $(i = 1, 2 \ldots, n)$. The T-BAQL approach consists of two stages: (i) maximize the classical QLF with respect to β and set up the QLF of $v = g(\mu)$ based on $\hat{\beta}$; (ii) maximize the QLF of v with respect to β again.

To understand the T-BAQL method, recall that in the A-BAQL approach the weighted quasi-likelihood of y over $v = g(\mu)$ is maximized with respect to parameter β. T-BAQL, however, the conditional quasi-likelihood of y given v (see the classical QLF method discussed in Section 5.2) is maximized. However, rather than using the QLF conditional on $\mu = X_i^T\beta = c_i$ mentioned previously, we choose the QLF conditional on the improved c_i. The so-called "improved" c_i means that c_i is updated at the second stage through adjusting β estimate. The shrinkage estimate of β is finally obtained from the inverse-variance weighted average of the β estimates of the two stages. Therefore, T-BAQL method will estimate β in a crisscrossing way during iterative procedure. The T-BAQL method is depicted in Figure 5.2; the details follow.

The T-BAQL Algorithm

(B1) Set initial index $j = 0$ and choose $\hat{\beta}^{(0)}$.

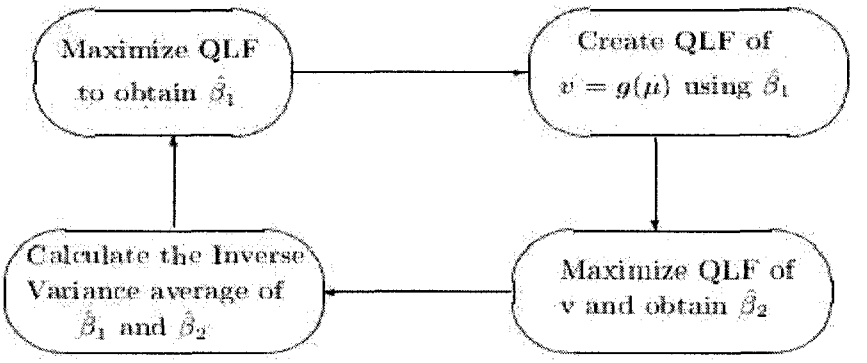

Figure 5.2: The procedure of A-BAQL method for GLMs

(B2) Maximize the classical QLF $Q(y|\beta)$ of y using the Newton-Raphson iterative algorithm and obtain $\hat{\beta}_1^{(j)}$ from $\hat{\beta}^{(j)}$.

(B3) Calculate $\hat{v}_i = g(\hat{\beta})$ $(i = 1, 2 \ldots, n)$ based on $\hat{\beta}^{(j+1)}$, and construct the QLF of $\hat{v} = (\hat{v}_1, \hat{v}_2, ..\hat{v}_n) = (g(\hat{\mu}_1), g(\hat{\mu}_2), ..g(\hat{\mu}_n))$: $Q_v(\hat{v}_1 ... \hat{v}_n|\beta)$.

(B4) Maximize $Q_v(\hat{v}_1 ... \hat{v}_n|\beta)$ with respect to β and get $\hat{\beta}_2^{(j)}$.

(B5) Set $\hat{\beta}^{(j+1)} = w_1\hat{\beta}_1^{(j)} + w_2\hat{\beta}_2^{(j)}$, i.e., $\hat{\beta}^{(j+1)}$ is the inverse-variance weighted average of $\hat{\beta}_1^{(j)}$ and $\hat{\beta}_2^{(j)}$.

(B6) Set $j = j + 1$ and go to (B2) until some given threshold is reached.

5.4 Approximation of Density

Both A-BAQL and T-BAQL approaches require prior density (or log-density) information of the random variable v. In the A-BAQL algorithm it is necessary to estimate the density of $\hat{v}_i$ $(i = 1, 2 \ldots, n)$ which can be used to update the QLF in the following step. Generally, we can use semi-parametric or non-parametric methods for

density estimation since these techniques let the data reflect the distribution or the error term or indicate the shape of relationship between observations and probability density without imposing too many distributional assumptions. In this chapter we considered the following two approaches:

(1) Kernel Estimation [Boente, G. et al. 2000]: $f^* = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$, where K is the kernel function, which is always positive and integrates to 1. h is the bandwidth, which can adjust the tradeoff between variance and bias. The most popularly used kernels include the Gaussian (normal), Epanechnikov and bi-weight kernels.

(2) Locally Weighted Distribution (LWD): If we only estimate the density of a specific sample x, then the observations in the local area of x can be thought as the samples from a normal distribution. Thus, the density $d(x)$ of x is estimated by

$$d(x) \doteq N(x|\hat{\mu}_x, \hat{\sigma}_x^2)(\frac{n_i}{n}), \tag{5.4.1}$$

where $N(.)$ is the normal density function, $\hat{\mu}_x$ and σ_x^2 are the estimated mean and variance via local area samples respectively. n_i is the number of observations in the local area of x and n is the total count of samples. Therefore, $\frac{n_i}{n}$ represents the proportion of the sample in the local area.

According to Bayesian theory, (5.4.1) can be applied for the purposes of density estimation. It is easy to see that (5.4.1) is relatively computationally convenient compared with the kernel estimation.

In the T-BAQL approach we use the QLF of v_i, i.e., (5.3.1), to approximate the log-likelihood. However, (5.4.1) can still be simplified through the LWD approach introduced above. This result is based on the following lemma:

Lemma 5.1 Denote $F(v_i) = \int_{v_i}^{\theta_i} \frac{v_i - t}{\sigma_1^2(t)} dt$, $P(v_i) = \frac{-1}{2}\frac{(v_i-\theta_i)^2}{\sigma_1^2(\theta_i)}$, $\delta_i = v_i - \theta_i$, then $|F(v_i) - P(v_i)| = O[(v_i - \theta_i)^3] = O(\delta_i^3)$.

Proof. See Appendix A.

Lemma 4.1 reminds us that we may adopt the following procedure to estimate β at step (B4) in the T-BAQL algorithm:

$$\text{Maximize} \quad \sum_{i=1}^{n} \frac{1}{2} \frac{(\hat{\xi}_i - c_i)^2}{\hat{\sigma}_i^2} \tag{5.4.2}$$

where $c_i = X_i^T \beta$. $\hat{\xi}_i$ and $\hat{\sigma}^2$ are local area sample mean and variance respectively. It is noted that we still need to select a suitable bandwidth to define a good LWD function.

5.5 Another potential approach

Another possible BALQ approach for GLMs is the combination of A-BAQL and T-BAQL. As stated before, the classical Quasi-likelihood function:

$$Q_y(y|\beta) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2(t)} dt = \sum_{i=1}^{n} \int_{y_i}^{g^{-1}(c_i)} \frac{y_i - t}{\sigma^2(t)} dt = \sum_{i=1}^{n} q_{y_i} \tag{5.5.1}$$

can be treated as the approximation of conditional likelihood given $v = g(\mu)$. Thus, the following function:

$$\Omega(y, v) = \sum_{i=1}^{n} q_{y_i} + \sum_{i=1}^{n} \int_{g(\hat{\mu}_i)}^{\theta_i} \frac{g(\hat{\mu}_i) - t}{\sigma_1^2(t)} dt \tag{5.5.2}$$

can be regarded as the approximation of joint log likelihood of y and v. Where the Quasi-likelihood of v: $\sum_{i=1}^{n} \int_{g(\hat{\mu}_i)}^{\theta_i} \frac{g(\hat{\mu}_i) - t}{\sigma_1^2(t)} dt$ is given in expression (5.3.1) in section 5.3. We can estimate β by maximizing (5.5.2).

The logic of this method is that maximizing 5.5.2 is equivalent to maximizing the posterior of v given y with respect to β. This method is similar to A-BAQL approach since they all optimize updated Quasi-likelihood, but rather than the weighted Quasi-likelihood in A-BAQL, here we maximize the posterior of v given y. This method

is also similar to T-BAQL because it is also involved the Quasi-likelihood function of v in (5.3.1), but instead of maximizing two Quasi-likelihood functions separately at the same iterative step in T-BAQL procedure, here we maximize the sum of two Quasi-likelihood functions.

5.6 Numerical Study

In this section we work out the computational test to illustrate the performance of the proposed BALQ approach for GLMs. The simulated data represent three gene expression profiles over a time course. They are considered random variables and denoted as y,x_1 and x_2. We will model regulatory relations among these genes, where y receives a binary recoding as "high" (1) or "low" (0) because the distribution of gene expression is often over-dispersed over the time course. The samples were simulated using S-Plus. Here, we only study the relation between y (as dependent variable) and x_1 and x_2 (as independent variables) based on the following two logistic regression models:

$$\text{logit}(y) = \log(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1 \qquad \text{(Example I)}$$

$$\text{logit}(y) = \log(\frac{P}{1-P}) = \beta_1 x_1 + \beta_2 x_2 \qquad \text{(Example II)}$$

The underlying data for independent variables x come from a normal distribution and response y (1 or 0) are generated from the Bernoulli data with the probability determined by the covariates x and the given β.

5.6.1 Example 1

The first model involves a sample taken at 120 points. The simulated regression model is

$$\text{logit}(y) = \beta_0 + \beta_1 x_1 = 1 - 0.5 x_1 \qquad (5.6.1)$$

Example 1	QLF	A-BAQL	T-BAQL
β_0 Estimate	1.114156	1.089796	1.113768
Discrepancy	0.114156	0.089796	0.113768
Std Err	0.183975	0.387115	0.271031
β_1 Estimate	-0.5152006	-0.5000058	-0.5149535
Discrepancy	0.0152006	0.0000058	0.0149535
Std Err	0.101752	0.221032	0.146436
SSE	0.054209	0.052471	0.054172

Table 5.1: Example 1. Simulation results for comparing methods

We estimate the parameter $\beta = (\beta_0, \beta_1)$ with classical QLF, A-BAQL and T-BAQL respectively. Through adjusting the bandwidth, we can get the A-BAQL and T-BAQL algorithms to converge to the results in Table 5.1.

These results indicate that both the A-BAQL and the T-BAQL approaches have less bias than the QLF method. Even though the QLF method has the smallest standard error, we can always decrease the standard error of A-BAQL or T-BAQL by increasing bandwidth. For instance, we can have T-BAQL reach the same standard error as QLF by increasing the bandwidth by 30%, but this will increase the bias of the β estimate. There will always be a tradeoff between variance and bias when applying the BAQL approach. However, the new methods provide us with more flexibility compared with QLF, as we have more space to choose the final estimate in terms of the balance between variance and bias through adjusting bandwidth. The SSE (sum square of error) is calculated by

$$SSE = \sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} (g(v_i) - g(\hat{v}_i))^2, \tag{5.6.2}$$

where μ_i $(i = 1, 2, \dots)$ are the actual mean of the y_i and $\hat{\mu}_i = g(\hat{v}_i(\hat{\beta}))$ are the predicted means from above algorithms. SSE can be considered to be a measurement for comparing the prediction accuracy of three methods. Table 5.1 shows that A-

Example 2	QLF	A-BAQL	T-BAQL
β_1 Estimate	1.0948908	1.498714	1.351676
Discrepancy	0.4051092	0.001286	0.148324
Std Err	0.115139	0.574631	0.073584
β_2 Estimate	-0.8546583	-1.125661	-1.086034
Discrepancy	0.1453417	0.125661	0.086034
Std Err	0.093953	0.464593	0.061722
SSE	0.517844	0.244715	0.598713

Table 5.2: Example 2. Simulation results for comparing methods

BAQL has the best prediction accuracy while T-BAQL and QLF are quite similar. This makes sense since the β inferred from A-BAQL has the smallest bias.

5.6.2 Example 2

The second example involves another logistic regression model with two parameters but without a constant term.

$$\text{logit}(y) = \beta_1 x_1 + \beta_2 x_2 = 1.5x_1 - x_2. \tag{5.6.3}$$

We generate a set of 100 pairs (x_1, x_2) and corresponding responses y_i $(i = 1, 2 \ldots, n)$, using (5.6.3). The parameter estimates obtained with three approaches are shown in Table 5.2.

We see from Table 5.2 that QLF is the most biased in the estimate of β, especially for the parameter $\beta_1 = 1.5$ (bias=0.4). It seems appealing that A-BAQL attains a fairly close estimate for β_1 ($\hat{\beta}_1 = 1.499$), and that T-BAQL is also less biased (bias=0.148) than OLF. On the other hand, T-BAQL enjoys the smallest standard error among three, while A-BAQL has the highest one. As for prediction accuracy, the smallest SSE is attained by A-BAQL.

To further compare these methods we also validate the resulting β estimates according to the response y_i $(i = 1, 2, \ldots, 100)$. We justify this procedure by the following argument: the higher the probability (or mean) of y_i, the more likely is

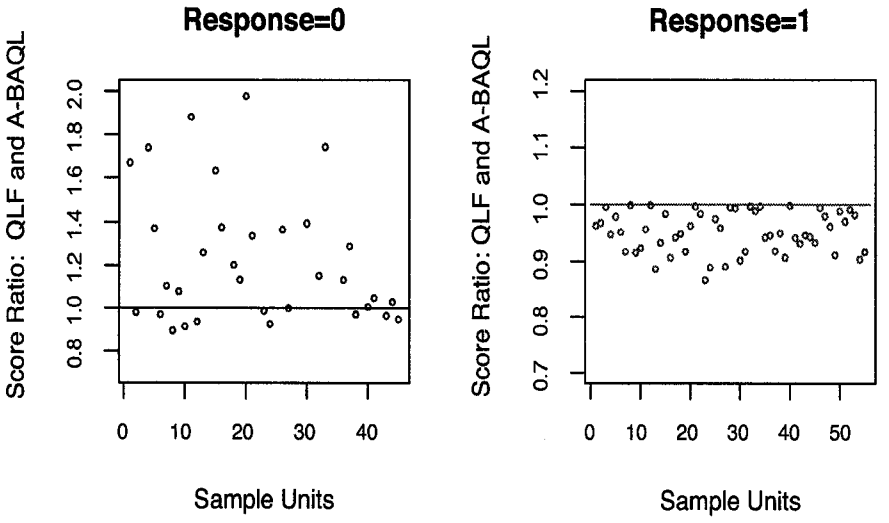

Figure 5.3: Comparison of A-BAQL and Classical QLF Methods

$y_i = 1$. Conversely y_i tends to be 0 with a lower probability (or mean). Therefore, for comparing QLF with A-BAQL and T-BAQL, the y_i ($i = 1, 2, \ldots, 100$) are first separated into two groups: (G0): response=0 and (G1): response=1. Then, within each group for each sampling point we calculate the ratio of predicted means (QLF:A-BAQL and QLF:T-BAQL respectively). Figure 5.3 and Figure 5.4 provide the plots for sample units and resulting ratio. It can be seen in these plots that both A-BAQL and T-BAQL perform better than QLF, since for group G0 both ratios (the ratio of predicted mean for QLF:A-BAQL and QLF:T-BAQL) are higher (generally higher than 1), and for group G1 are lower (less than 1). This trend is particularly clear for group G0, since the ratio of many samples exceeds 1.5, which indicates that the A-BAQL and T-BAQL methods are capable of avoiding large type II error. For group G1, though the ratios are not very low, the pattern is also very clear because the the ratios of almost all observations are beneath the base line (ratio=1).

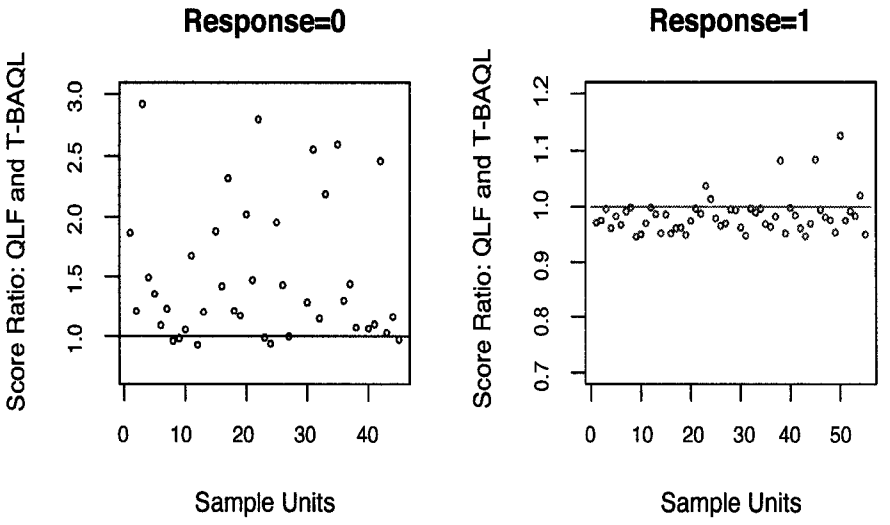

Figure 5.4: Comparison of T-BAQL and Classical QLF Methods

5.7 Other issues

5.7.1 Asymptotic properties

At this point, we mention some issues related to the proposed algorithms, namely their asymptotic properties and bandwidth selection, though these are not the main focus of this chapter.

Before discussing asymptotic properties of A-BAQL and T-BAQL algorithms, we need to check their Newton-Raphson procedures based on the first and second derivative of the estimation equation. The details of the iterative step for the A-BAQL approach is expressed in

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + [X'RWX]^{-1}[X'RW\Delta(y - \hat{\mu})] = \hat{\beta}^{(m)} + \Lambda^{-1}b, \tag{5.7.1}$$

where $R = \text{Diag}\{r_i\}$, $\Delta = \text{Diag}\{g_{\hat{\mu}}(\hat{\mu}_i)\}$, $r_i = (\sigma^2(\mu_i)g^2_{\hat{\mu}}(\hat{\mu}_i))^{-1}$ $(i = 1, 2 \ldots, n)$, and $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)^T$. The diagonal matrix $W = \text{Diag}(w_1, \ldots., w_n)$ stores the weight

w_i $(i = 1, 2 \ldots, n)$ for the QLF of $y_i (i = 1, 2 \ldots, n)$. The corresponding part of the T-BAQL approach consists of the following two steps:

$$\begin{cases} \hat{\beta}_1^{(m+1)} = \hat{\beta}_1^{(m)} + [X' R X]^{-1}[X' R \Delta (y - \hat{\mu})] = \hat{\beta}_1^{(m)} + \Lambda_1^{-1} b_1 \\ \hat{\beta}_2^{(m+1)} = [X' \Phi X]^{-1}[X' \Phi \hat{\xi}] = \Lambda_2^{-1} b_2, \end{cases}$$

where the first equation is the same as the iterative step in classical QLF, and the second one is inferred from the optimal solution of problem (5.4.2). $\Phi = \text{Diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots \hat{\sigma}_n^2)$ is a diagonal matrix which stores the estimated variance of $v = (v_1, \ldots, v_n)$ and $\hat{\xi} = (\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_n)$ is the estimated v mean vector. It is worth mentioning that both $\hat{\sigma}_i^2$ and $\hat{\xi}_i$ $(i = 1, 2, \ldots, n)$ are estimated by $\beta_1^{(m+1)}$ obtained in the first iterative step. Therefore, the right side of the equation in step two potentially contains β information obtained from the previous step.

Similar to the asymptotic property study in the classical QLF approach, the following assumption is needed for the asymptotic properties of the A-BAQL approach

(M1) There exists a positive definite matrix Σ, such that, when $n \to \infty$

$$\frac{1}{n}\{[X' R W X]^{-1}[X' W R W X][X' R W X]^{-1}\} \to \Sigma \tag{5.7.2}$$

We then have the following asymptotic theorem for A-BAQL method.

Theorem 5.1 Under the assumption (M1), the estimated β in A-BAQL algorithm is asymptotically normally distributed such that as $n \to \infty$

$$\sqrt{n}(\hat{\beta} - \beta) \to N_p(0, \Sigma) \tag{5.7.3}$$

The proof can be found in Appendix B.

The case of T-BAQL is more complicated, since the estimated β at one stage is the base of resulting estimate at another stage. However, if we assume that the β estimate has converged at the second stage, then the relationship between two neighboring estimates can be checked, from which the asymptotic property for T-BAQL algorithm can be studied. We do not pursue this further in this chapter.

5.7.2 Bandwidth choice

Both the A-BAQL and T-BAQL procedures require estimation of the density based on the estimated β from previous iterative step. The smoothing approach is associated with bandwidth selection, since this will affect the balance between bias and variance, i.e., larger bandwidth can result in higher bias and vice versa. Indeed from computational tests we find the estimates produced by A-BAQL and T-BAQL are sensitive to bandwidth choice.

We have traditional schemes such as cross-validation and goodness-fitting statistics in numerical tests but the result are not consistent. We also find that using a kernel function, as bandwidth$\rightarrow\infty$, the estimate from A-BAQL converges to the one from classical QLF method. Too small a bandwidth may lead to the failure of locally weighted distribution method.

Discussion

This chapter develops two BAQL approaches to cope with GLM problem, during which the sample means are treated as random variables.

Compared with the classical QLF method, the presented techniques can incorporate the probability distribution of parameter estimate from previous iterative steps. This is important, for example, when we study logistic or Log-linear regression model in which the response variable is binary or categorical, traditional approaches such as QLF or MLE will treat the same responses equal when Quasi-likelihood or Log-likelihood functions are created(even though these responses have different probability densities). The proposed methods will assign different weights to the likelihood functions of these responses in terms of probability densities using bandwidth strategy. These features make BAQL approaches more efficient and flexible. The numerical study has also confirm our conclusion.

BAQL approaches can be applied to describe or compare gene expression levels in gene differential studies. For instance, [Newton MA. et al. 2001] has considered empirical Bayes models for gene expression levels based on gamma and log-normal distributions.

Consider two-group gene expression data where we have obtained expression levels of m genes for n_1 samples from the first group and n_2 samples from another group. Denote the measured gene expression data as the following matrix x_{ij} , $i = 1, \ldots, m$, $j = 1, \ldots, n$, $n = n_1 + n_2$. In differential gene expression detection, we can test whether gene i is differentially expressed by the following GLM

$$g(\mu_i) = a_0 + a_1 y_j \quad (j = 1, \ldots n), \tag{5.7.4}$$

where $\mu_i = E(x_{ij})$, y_j is the group indicator, i.e. $y_i = 1$ if the $j - th$ sample comes from the first group, and $y_i = 0$ if the $j - th$ sample belongs to the other group. We can formally test the difference between two groups by testing whether $a_1 = 0$. Because the x_{ij} are not normally distributed, it is not adaptable to apply classical linear regression model (using t or F test). Instead, a GLM using reciprocal (or inverse) link function (for the gamma distribution) would be a better choice.

BAQL approaches can also be used to infer gene regulatory networks when gene expression data are not normally distributed. For instance, the GLM technique can be employed to tackle the boolean network model:

$$Y_i(t+1) = F_i(Y(t)) \tag{5.7.5}$$

where $Y_i(t)$, $i = 1 \ldots, n$ are the states of the n genes in the network. The function $F(.)$ is a boolean function of the states of the network at time t, for updating the state of element i at $1 + t$. Also, one can choose the following log-linear regression (three-way contingency table) to investigate the relationship among three genes:

$$\ln \pi_{ijk} = \lambda + \lambda_i + \lambda_j + \lambda_k \tag{5.7.6}$$

where $\ln \pi_{ijk}$ is the *log* of the expected cell frequency of the cases for cell ijk in contingency table. Finally, we have used the *logit* models in Section 5.6 to express a differential equation system (when gene expression levels are defined by binary variables). All these models assume the response variable to be non-normally distributed and the link function to be non-identity, which makes linear regression model inappropriate for the gene regulatory network problem. The other concern is the high dispersion levels in gene expression data, i.e. the higher means of expression levels over the time course often come with the higher variances. Therefore, GLM is more suitable to cope with such problems.

In GLMs the variance is often a function of mean. However, if the specific function form is unknown, we have to use estimation methods, such as the non-parametric variance function approach [Chiou, JM. et al. 1999]. The idea behind this is that the variance function at the current iterative step is estimated (or updated) through a non-parametric approach using the estimated mean obtained from the previous iterative step.

Appendix A: The proof of lemma 5.1

[Proof] Using the Taylor expansion at $v_i = \theta_i$, since $F(\theta_i) = 0$, we have

$$\frac{dF}{dv_i} \mid_{v_i=\theta_i} = \frac{d}{dv_i}[v_i \int_{v_i}^{\theta_i} \frac{1}{\sigma_1^2(t)} dt - \int_{v_i}^{\theta_i} \frac{t}{\sigma_1^2(t)} dt] \mid_{v_i=\theta_i} \tag{5.7.7}$$

$$= [\int_{v_i}^{\theta_i} \frac{dt}{\sigma_1^2(t)} + v_i \frac{d}{dv_i}(\int_{v_i}^{\theta_i} \frac{dt}{\sigma_1^2(t)}) - \frac{d}{dv_i}(\int_{v_i}^{\theta_i} \frac{t}{\sigma_1^2(t)})] \mid_{v_i=\theta_i} \tag{5.7.8}$$

$$= -\frac{\theta_i}{\sigma_1^2(\theta_i)} + \frac{\theta_i}{\sigma_1^2(\theta_i)} \tag{5.7.9}$$

$$= 0 \tag{5.7.10}$$

For the second derivative:

$$\frac{d^2F}{dv_i^2}\mid_{v_i=\theta_i}=\frac{d}{dv_i}[\frac{-v_i}{\sigma_1^2(v_i)}+\frac{v_i}{\sigma_1^2(v_i)}+\int_{v_i}^{\theta_i}\frac{dt}{\sigma_1^2(t)}]\mid_{v_i=\theta_i}=\frac{1}{\sigma_1^2(\theta_i)} \tag{5.7.11}$$

thus we have

$$F(v_i)=\frac{-1}{2\sigma_1^2(\theta_i)}(v_i-\theta_i)^2+O((v_i-\theta_i)^3) \tag{5.7.12}$$

$$=P(v_i)+O((v_i-\theta_i)^3) \tag{5.7.13}$$

$$=P(v_i)+O(\delta_i^3) \tag{5.7.14}$$

□

Appendix B: The proof of Theorem 5.1

[Proof] From (5.7.1) we adjust $\beta^{(m)}$ by

$$\hat{\beta}^{(m+1)}-\hat{\beta}^{(m)}=[X'RWX]^{-1}[X'RW\Delta(y-\mu)] \tag{5.7.15}$$

Then, following the proof procedure of McCullagh [McCullagh P. 1983] on the asymptotic property of the classical QLF approach, we first show:

$(H1)\mathrm{E}(y-\mu)=0$

$(H2)\mathrm{Var}([X'RWX]^{-1}[X'RW\Delta(y-\mu)])=[X'RWX]^{-1}[X'WRWX][X'RWX]^{-1}=i_\beta$

where $-i_\beta$ is the expected second derivative matrix of quasi-likelihood (or quasi-information matrix). To do this, notice that here $\mu=\mathrm{E}(y|u)$ is a random variable vector (see assumption (S2) in Section 5.2),we have

$$\mathrm{E}(y-\mu)=\mathrm{E}(y-\mathrm{E}(y|u))=\mathrm{E}(y)-\mathrm{E}(\mathrm{E}(y|u))=\mathrm{E}(y)-\mathrm{E}(y)=0 \tag{5.7.16}$$

so (H1) is true. Also,

$$\mathrm{Var}(y-\mu)=\mathrm{Var}(y-\mathrm{E}(y|u))$$

$$
\begin{aligned}
&= \mathrm{Var}[\mathrm{E}((y - \mathrm{E}(y|u))|u)] + E[\mathrm{Var}((y - \mathrm{E}(y|u))|u)] \\
&= \mathrm{Var}[\mathrm{E}(y|u) - \mathrm{E}(y)] + \mathrm{E}(\mathrm{Var}(y|u)) \\
&= \mathrm{Var}(\mu) + \mathrm{Var}(y) - \mathrm{Var}(\mathrm{E}(y|u)) \\
&= \mathrm{Var}(y) = \mathrm{Diag}\{\sigma^2\}
\end{aligned}
$$

hence

$$
\mathrm{Var}([X'RWX]^{-1}[X'RW\Delta(y-\mu)]) = [X'RWX]^{-1}X'RW\Delta\Delta WRX[X'RWX]^{-1}\sigma^2. \tag{5.7.17}
$$

Since matrix R and W are both diagonal and $R\Delta\Delta R\sigma^2 = R$ (see Section 5.7), then (H2) holds.

The "observed" quasi-likelihood information matrix of β I_β has the following relationship with i_β

$$
|I_\beta - i_\beta| = O_p(n^{\frac{1}{2}}) \tag{5.7.18}
$$

and since I_β is evaluated at a point β^* lying on the line segment joining "real" β and $\hat{\beta}$, the asymptotic property for A-BAQL follows immediately. □

Chapter 6

Conclusion and future work

6.1 Conclusion

Microarray data are rich in information that allows us to gain insight into biological processes. In this thesis we are especially interested in exploring time-course experiments during which gene expression is monitored dynamically, namely "time-course" gene expression. Such information is very useful since the dynamics of the data can lead to a deeper understanding of biological processes such as stem cell differentiation. Many existing computational and statistical methods can be applied to the study of time-course data if each time point is treated as one dimension in the algebraic space, such as in traditional linkage clustering analysis and the ANOVA F test. However, the time dependency property of time-course data means that much information is lost by these classical methods. For instance, failure to consider auto-correlation over the time course may create extra and unnecessary clusters. In this thesis we have developed several statistical and data mining methods specifically for analyzing time-course gene expression data on embryonic stem cells.

To compare gene expression levels under different conditions, we propose a novel adaptive SAM technique to overcome the disadvantages of the classical SAM method.

Rather than using a universal fudge factor, our procedure chooses a flexible fudge factor strategy to adjust the sample standard error. The optimal fudge factor is obtained by using a direct searching approach combined with a bootstrap method. Employing the same technique, we extend the SAM statistic to the comparison of multivariate samples. Compared with the ANOVA F test and Hotelling T square test, the new multivariate SAM statistic is capable of tackling time dependency effect as well as the variance-mean dependency problem. The computational analysis of time-course gene expression data on embryonic stem cell differentiation, combined with a gene ontology study confirms that the proposed method is more than competitive with the classical method.

In order to capture common gene regulatory patterns over a time-course, we have developed two dynamic clustering methods: dynamical correlation based clustering and gene functional trajectories clustering. The first method defines a shape similarity measure in terms of the concept of random curves and integrates it into a general agglomerative hierarchical clustering scheme and into k-means. The second method belongs to model-based clustering, during which several important components are taken into account: gene network interactions, autocorrelation effects, time-dependent random effects and noise. We use a genetic transcription model to deduce a family of functional curves which is capable of fitting many complicated gene profiles. Both methods are applied to a mouse ESC line during the first 24 hours of differentiation. The biological credibility of the results, verified by FatiGO, shows that these methods outperform classical approaches in accurately classifying gene profiles.

Motivated by the problem of inferring gene regulatory networks from time-course gene expression data, we develop two Bayesian Adjusted Quasi-likelihood (BAQL) approaches to fit Generalized Linear Models (GLMs): Adjusted Quasi-Likelihood Function (A-BAQL) and Two-Stage Adjusted Quasi-Likelihood Function (T-BAQL). These methods should have application beyond the context of microarray analysis. Computational tests demonstrate that the presented approach is more powerful in

adjusting estimate bias and prediction accuracy than the classical quasi-likelihood method. Other numerical tests confirm that, by setting suitable bandwidth, both A-BAQL and T-BAQL can avoid large type II error.

6.2 Future work

Though the genes in the same cluster are believed to reflect related biological processes or to be co-regulated, this may only be true in a partial range of the time course. For example, some genes (such as Pouf51 and Sox2) are strongly co-regulated only in the early period of stem cell differentiation. Therefore, rather than using clustering analysis over entire time course, biclustering analysis would be more desirable. There are some biclustering methods that work on multivariate data sets [Chen Y. et al. 2000, Madeira SC. et al. 2004], but little work on time-course gene expression data. Further improvements in this sense could be made to both proposed dynamic clustering methods.

In the adaptive SAM method, we have improved on the standard error function for the SAM test statistic, but there may well be even better ways of constructing the fudge factor, a goal for further research. Furthermore, when using the classical or adaptive SAM methods, we note that the final results in both cases are highly dependent on the initial transformation. For instance, applying these methods to raw data, log transformed data and square-root transformed data will generate quite different results, such as in the number of significant genes and the FDR. This is because the log or square root transformations also have shrinkage effects. How to combine SAM techniques with data transformation is an important direction for future work related to differential gene expression.

Bibliography

[Aach J. et al. 2001] Aach J., Church G. M. 2001. Aligning gene expression time series with time warping algorithms. Bioinformatics 17:495–508.

[Albers CJ. et al. 2006] Albers CJ, Jansen RC, Kok J, Kuipers OP, van Hijum SA. 2006. SIMAGE: simulation of DNA-microarray gene expression data. BMC Bioinformatics 205(7): 1471–2105.

[Amaratunga D. et al. 2003] Dhammika Amaratunga, Javier Cabrera. 2003. Exploration and Analysis of DNA Microarray and Protein Array Data. Hoboken(NJ): John Wiley & Sons.

[Arias AM. et al. 2002] Alfonso Martinez Arias, Alison Stewart, and Alfonso Martinez Arias. 2002. Molecular Principles of Animal Development. Oxford University Press, NY.

[Ashburner M. et al. 2000] Ashburner M., Ball CA., Blake, JA., Botstein D., Butler H., Cherry JM., Davis AP., Dolinski K., Dwight SS., Eppig JT., Harris MA., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese J.C., Richardson JE., Ringwald M., Rubin, GM., Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 25: 25–29.

[Baldi P. et al. 2001] Baldi P., and Long, A. D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17, 509-519.

[Baharvand H. et al. 2007] Hossein Baharvand, Ali Fathi, Dennis van Hoof, Ghasem Hosseini Salekdeh. 2007. Trends in stem cell proteomics. Stem Cells Express, published online doi:10.1634/stemcells.2007–0107.

[Ball GH. et al. 1967] Ball GH., Hall, DJ. 1967. A clustering technique for summarizing multivariate data. Behavioral Science. 12: 153–155.

[Barker JN. et al. 2005] Barker JN, Weisdorf DJ, DeFor TE, Blazar BR, McGlave PB, Miller JS, Verfaillie CM, Wagner JE. 2005. Transplantation of 2 partially HLA-matched umbilical cord blood units to enhance engraftment in adults with hematologic malignancy. Blood 105(3): 1343–7.

[Bar-Joseph Z. et al. 2002] Bar-Joseph Z., Gerber G., Gifford D., Jaakkola T., Simon I. 2002. A new approach to analyzing gene expression time series data. Annual Conference on Research in Computational Molecular Biology archive Proceedings of the sixth annual international conference on Computational biology Washington page:39 - 48.

[Bar-Joseph Z. et al. 2003] Bar-Joseph, Z., Gerber G., Gifford D., Jaakkola T., Simon, I. 2003. Continuous representations of time-series gene expression data. Journal of Computational Biology 10: 341–356.

[Beissbarth T. et al. 2004] Beissbarth T, Speed TP. 2004. GOstat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 20(9): 1464–1465.

[Bhattacharya B. et al. 2004] Bhattacharya B, Miura T, Brandenberger R, Mejido J, Luo Y, Yang AX, Joshi BH, Ginis I, Thies RS, Amit M, Lyons I, Condie BG,

Itskovitz-Eldor J, Rao MS, Puri RK. 2004. Gene expression in human embryonic stem cell lines: unique molecular signature. Blood 103: 2956–2964.

[Bilmes JA. 1997] JEBilmes JA. 1997. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Techical Report 97-021. Berkeley, CA: International Computer Science Institute.

[Boente, G. et al. 2000] Boente, G. and Fraiman, R. 2000 Kernel-based functional principle components, Statistics and Probability Letters, 48, 335-345

[Brumback BA. et al. 1998] Brumback BA., Rice J. 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. Journal of the American Statististical Association. 93: 961–976.

[Burnham K.P. et al. 1998] Burnham K.P. and Anderson D.R. 1998. Model Selection and Inference:a Practical Information-Theoretic Approach. New York: Springer.

[Chambers I. et al. 2003] Chambers Ian, Douglas Colby, Morag Robertson, Jennifer Nichols, Sonia Lee, Susan Tweedie, and Austin Smith. 2003. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113: 643–655.

[Chiou, JM. et al. 1999] Chiou, JM. and Muler, H.-G. 1999. Nonparametric quasi-likelihood,Annals of Statistics 27, 34-64.

[Chen Y. et al. 2000] Cheng Y, Church GM 2000. Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology: 93103.

[Chen T. et al. 1999] Chen T., He, HL., Church GM. 1999. Modeling gene expression with differential equations. In Pacific Symposium on Biocomputing 1999: 29–40.

[Coffman CJ. et al. 2005] Coffman CJ, Wayne ML, Nuzhdin SV, Higgins LA, McIntyre LM. 2005. Identification of co-regulated transcripts affecting male body size in Drosophila Genome Biology. 6(6): R53

[Chudova D. et al. 2004] Chudova D., Hart C., Mjolsness E., Smyth P. 2004. Gene expression clustering with functional mixture models. Advances in Neural Information Processing 16, MIT Press.

[D'haeseleer P. et al. 1999] D'haeseleer P, Wen X, Fuhrman S, Somogyi R. 1999. Linear modeling of mRNA expression levels during CNS develoment and injury. Pacific Symposium on Biocomputing 4: 41–52.

[Dubin JA. et al. 2005] Dubin JA. , Mller HG. 2005. Dynamical correlation for multivariate longitudinal data. Journal of the American Statistical Association. 100(471): 872–88.

[Dudoit S. et al. 2002] Dudoit S. and Fridlyland J. 2002 A prediction-based resampling method for estimating the number of clustering in a dataset. Genome Biology 3: research 0036.1-0036.21.

[Efron B. et al. 2001] Efron B., Tibshirani, R., Storey J. D., and Tusher V. 2001 Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96: 11511160.

[Fraley C. et al. 1998] Fraley C., Raftery AE. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41(8):578-588 41: 578–588.

[Friedman N. et al. 2000] Friedman N., Linial M., Nachman I., D. Pe'er J. 2000. Using Bayesian Network to Analyze Expression Data Computational Biology 7: 601–620.

[Greenway DJ. et al. 2006] Deborah J. Greenway, Miyoko Street, Aaron Jeffries, Noel J. Buckle. 2006. REST maintains a repressive chromatin environment in embryonic hippocampal neural stem cells. Stem Cells Express, published online doi:10.1634/stemcells.2006–0207.

[Guo X. et al. 1995] X. Guo, W. Pan. 2005. Using weighted permutation score to detect differential gene expression with microarray data. Journal of Bioinformatics and Computational Biology 3(4): 989–1006.

[Hartigan JA. et al. 1979] Hartigan JA, Wong MA. 1979. A K-Means Clustering Algorithm. Applied Statistics 28(1): 100-108.

[HURVICH CM. et al. 1989] HURVICH, Clifford M., and Chih-Ling TSAI, 1989. Regression and Time Series Model Selection in Small Samples. Biometrika, 76(2): 297-307.

[Iyer VR. et al. 1999] Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. 1999. The transcript-ional program in the response of human fibroblasts to serum. Science 283: 83–87.

[Johnson RA. et al. 2002] Richard A Johnson, Dean W Wichern. Applied Multivariate Statistical Analysis. 2002. 5th edition. New Jersey: Prentice Hall.

[Kauffman SA. 1969] Kauffman SA. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology. 22: 437–467.

[Kauffman SA. 1993] Kauffman SA. 1993. Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press. Technical monograph. ISBN 0-19-507951-5

[Li Y. et al. 2005] Li Y., McClintick J., Zhong L., Edenberg HJ., Yoder MC. , Chan RJ. 2005. Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. Blood, 105(2): 635–637.

[Liang S. et al. 1998] 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.

[Lindstrom MJ. et al. 1988] Mary J. Lindstrom, Douglas M. Bates 1988. Newton-Raphson and EM algorithm for linear mixed-effects model for repeated-measures data. Journal of the American Statistical Association, 83(404): 1014–1022.

[Lu Y. et al. 2005] Lu Y, Liu PY, Xiao P, Deng HW. 2005. Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics 21(14): 3105–3113.

[Luan Y. et al. 2003] Luan Y., Li H. 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19: 474–482.

[McCullagh P. 1983] McCullagh P. 1983. Quasi-likelihood functions Annals of statistics. 11: 59–67.

[McLachlan GJ. et. 1997] McLachlan GJ., Krishnan T. 1997. The EM Algorithm and Extensions. John Wiley and Sons, New York.

[Morsli H. et al. 1999] Morsli H., Tuorto F., Choo D., Postiglione MP., Simeone A., Wu DK. 1999. Otx1 and Otx2 activities are required for the normal development of the mouse inner ear. Development 126: 2335–2343.

[Madeira SC. et al. 2004] Madeira SC, Oliveira AL 2004. "Biclustering Algorithms for Biological Data Analysis: A Survey". IEEE Transactions on Computational Biology and Bioinformatics 1 (1): 2445.

[Nakatake Y. et al. 2006] Yuhki Nakatake, Nobutaka Fukui, Yuko Iwamatsu, Shinji Masui, Kadue Takahashi, Rika Yagi,Kiyohito Yagi, Jun-ichi Miyazaki, Ryo Matoba, Minoru S. H. Ko, Hitoshi Niwa1. 2006. Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. Molecular and Cellular Biology 26(20): 7772–7782.

[Newton MA. et al. 2001] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. 2001 On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology 8(1): 37-52.

[Nishimoto M. et al. 1999] Nishimoto M., Fukushima A., Okuda A., Muramatsu M. 1999. The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. Molecular and Cellular Biology, 19(8): 5453–5465.

[Niwa H. et al. 1998] Niwa H., Burdon T., Chambers I., Smith A. 1998. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes and Development 12: 2048–2060.

[Park In. et al. 2002] In-Kyung Park, Yaqin He, Fangming Lin, Ole D. Laerum, Qiang Tian, Roger Bumgarner, Christopher A. Klug, Kaijun Li, Christian Kuhr, Michelle J. Doyle, Tao Xie, Michl Schummer, Yu Sun, Adam Goldsmith, Michael F. Clarke, Irving L. Weissman, Leroy Hood, Linheng Li. 2002. Differential gene expression profiling of adult murine hematopoietic stem cells. Blood 99: 488–498.

[Perez-Iratxeta C. et al. 2005] Perez-Iratxeta C, Palidwor G, Porter CJ, Sanche NA, Huska MR, Suomela BP, Muro EM, Krzyzanowski PM, Hughes E, Campbell PA, Rudnicki MA, Andrade MA. 2005. Study of stem cell function using microarray experiments. FEBS Letters 579(8): 1795–801.

[Press, W. et al. 1992] Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; and Vetterling, W. T. 1992. Newton-Raphson Method Using Derivatives and Newton-Raphson Methods for Nonlinear Systems of Equations. 9.4 and 9.6 in Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed. Cambridge, England: Cambridge University Press, pp 355-362 and 372-375.

[Puente LG. et al. 2006] Lawrence G. Puente, Douglas J. Borris, Jean-Franois Carrire, John F. Kelly, Lynn A. Megeney. 2006. Identification of candidate regulators of embryonic stem cell differentiation by comparative phosphoprotein affinity profiling. Molecular and Cellular Proteomics 5: 57–67.

[Ramalho-Santos M. et al. 2002] Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. 2002. 'Stemness': transcriptional profiling of embryonic and adult stem cells. Science 298(5593): 597–599.

[Ramoni MF. et al. 2002] Marco F. Ramoni, Paola Sebastiani, Isaac S. Kohane. 2002. Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences of the United States of America. 99(14): 9121–9126

[Raymer E. 2005] Raymer Elizabeth 2005. New strategy will boost cord blood stem cells. University of Toronto. Retrieved on September 20, 2006.

[Rosfjord E. et al. 1994] Rosfjord E., Rizzino A. 1994. The octamer motif present in the Rex-1 promoter binds Oct-1 and Oct-3 expressed by EC cells and ES cells. Biochemical and Biophysical Research Communications, 203(3): 1795–802.

[Rousseeuw P.J. et al. 1987] Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53–65.

[Sene KH. et al. 2007] Kagnew Hailesellasse Sene, Christopher J Porter, Gareth Palidwor, Carolina Perez-Iratxeta, Enrique M Muro, Pearl A Campbell, Michael

A Rudnicki, Miguel A Andrade-Navarro. 2007. Gene function in mouse embryonic stem cell differentiation. BMC Genomics 8: 85

[Singh SK. et al. 2008] Singh SK, Kagalwala MN, Parker-Thornburg J, Adams H, Majumder S. 2008. REST maintains self-renewal and pluripotency of embryonic stem cells. Nature 453(7192): 223–7.

[Schwarz G. 2008] Schwarz G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461–464.

[Spiegelhalter, DJ. et al. 2002] Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64(4): 583-639.

[Takahashi K. et al. 2006] Takahashi K. Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126(4): 663–76.

[Tusher VG. et al. 2001] Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences USA 98: 5116–5121.

[UC Irvine Reeve-Irvine Research Center, 2005] UC Irvine Reeve-Irvine Research Center. 2005. Adult human neural stem cells used to successfully regenerate spinal cord tissue. Medical Research News.

[Wahde M. et al. 2001] Wahde M., Hertz J., 2001. Modeling genetic regulatory dynamics in neural development. Journal of Computational Biology. 8: 429–442.

[Wedderburn RWM. et al. 1974] Wedderburn RWM. 1974. Quasi-likelihood functions, generalised linear models, and the gauss-newton method. Biometrika 61: 439–447.

[Wright S. 1934] Wright S. 1934. The Method of Path Coefficients. Annals of Mathematical Statistics. 5: 161–215.

[Wu FX., et al. 2005] Wu F., Zhang WJ, Kusalik AJ. 2005. Dynamic model-based clustering for time-course gene expression data. Journal of Bioinformatics and Computational Biology 3: 821–836.