2006

New aspects of statistical methods for missing data problems, with applications in bioinformatics and genetics

Dong Wang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1312

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
New aspects of statistical methods for missing data problems,
with applications in bioinformatics and genetics

by

Dong Wang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dan Nettleton, Co-major Professor
Song Xi Chen, Co-major Professor
    Jack C. M. Dekkers
    Kenneth J. Koehler
    Soumendra N. Lahiri
    Mervyn G. Marasinghe

Iowa State University
Ames, Iowa
2006

Copyright © Dong Wang, 2006. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Dong Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major-Professor
Signature was redacted for privacy.

Co-major Professor
Signature was redacted for privacy.

For the Major Program
DEDICATION

To my family
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Missing Data Problems</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Estimating Equations and Empirical Likelihood</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Statistical Issues in Microarray Data Analysis</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Quantitative Trait Locus Mapping</td>
<td>8</td>
</tr>
<tr>
<td>1.5 Thesis Organization</td>
<td>9</td>
</tr>
<tr>
<td>1.6 References</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 2. Nonparametric Imputation of Missing Values for Estimating Equation Based Inference</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Nonparametric Imputation</td>
<td>20</td>
</tr>
<tr>
<td>2.3 Empirical Likelihood</td>
<td>22</td>
</tr>
<tr>
<td>2.4 Main Results</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Simulation Results</td>
<td>28</td>
</tr>
<tr>
<td>2.5.1 Correlation Coefficient</td>
<td>29</td>
</tr>
<tr>
<td>2.5.2 Generalized Linear Models with Missing Covariates</td>
<td>33</td>
</tr>
<tr>
<td>2.6 Empirical Example</td>
<td>34</td>
</tr>
<tr>
<td>2.7 References</td>
<td>38</td>
</tr>
</tbody>
</table>
4.6 Conclusion and Discussion ........................................ 111
4.7 References ......................................................... 112
4.8 Appendix .......................................................... 115

CHAPTER 5. General Discussion and Future Research ............ 120
5.1 References .......................................................... 121

ACKNOWLEDGEMENTS .................................................. 122
LIST OF TABLES

Table 2.1  Inference for the correlation coefficient .......................... 30
Table 2.2  Inference for parameters in a logistic regression model .......... 32
Table 2.3  Mouse eye weight and gene expression ............................. 36
Table 3.1  The power of the Wald test, LRT, and score test for the random
            selection strategy ............................................. 71
Table 3.2  The power of the Wald test, LRT and score test when selecting
            individuals with extreme trait values .......................... 72
Table 3.3  The type I error rate for the Wald test ............................ 73
Table 3.4  The type I error rate and power in testing the significance of the
            correlation coefficient .......................................... 75
Table 3.5  Mean and median misclassification error rates and power for QTL
            detection for varying QTL effect sizes. ......................... 77
Table 3.6  Power of the Wald test with misclassification ..................... 78
Table 3.7  Power of the test for within-QTL-genotype correlation with mis-
            classification .................................................... 79
Table 4.1  Power properties of the EL based test with extra trait data under
            bivariate normality and equal covariance ........................ 104
Table 4.2  Power of the EL based test with extra trait data under bivariate
            normality but with unequal covariance .......................... 106
Table 4.3  Power properties of the EL based test with extra trait data under different distributions and covariance matrices .......................... 107

Table 4.4  Power properties of the EL based test using extra trait data, with preferential selection ....................................................... 110
LIST OF FIGURES

Figure 3.1 An idealized depiction of the joint distributions of trait and expression ........................................ 66
Figure 3.2 The power of the Wald test, LRT, and score test for the random selection strategy ........................................ 70
Figure 3.3 The power of the Wald test with regard to the correlation coefficient 74
Figure 4.1 Power of the EL based test with extra trait data under bivariate normality and equal covariance 105
Figure 4.2 Power properties of different tests when the distributions are different across genotypes 109
CHAPTER 1. Introduction

This thesis discusses some new aspects of statistical methodology in dealing with missing data problems and the application of missing data methods to biological research. As missing data problems become more commonplace in complex biological research and other areas, a method with relaxed assumptions while flexible enough to accommodate a wide range of situations is highly desired. The nonparametric imputation method proposed in this thesis has great potential for addressing many problems involving missing data. Missing data methods are known to be indispensable when some observations are not available due to reasons beyond the researcher's control, such as instrument malfunction, human error or nonresponse. We will show that missing data methods are also highly valuable from an experimental design point of view. Researchers may choose to obtain measurements for some variables on only a subset of objects out of financial or time constraints. The selective transcriptional profiling approach considered in this thesis is a good example of using missing data methodology to improve the efficiency of statistical inference in studying gene regulation networks. First we shall review several topics important to later discussion in the thesis. More literature reviews can be found in the introduction sections of Chapter 2 to Chapter 4. The organization of this thesis is outlined at the end of this chapter.
1.1 Missing Data Problems

Missing data problems can occur in many statistical settings. It is thus unsurprising that researchers have shown great interest in methods dealing with missing data since the early half of the twentieth century. Wilks (1932) studied maximum likelihood estimation for multivariate normal models with fragmentary data, and this topic has since been studied intensively by Anderson (1957) and other researchers. Yates (1933), Bartlett (1937), and Healy and Westmacott (1956) introduced the idea of filling in missing values with least square estimates, which is a special form of imputation. Missing data methods have flourished since the early 1970's with the advance of computation technology, which makes it possible to carry out previously laborious numerical calculations. Important developments include the EM algorithm (Dempster, Laird, and Rubin, 1977), multiple imputation (Rubin, 1987), weighted estimating equations (Robins, Rotnitzky, and Zhao, 1994), and the widespread adoption of bootstrap and Bayesian methods, among others. Commonly used methods for missing data problems are reviewed in Chapter 2, and an excellent source of general information is Little and Rubin (2002). In this section, we give more details on the mechanisms that lead to missing data.

An important issue for missing data mechanisms is whether the missingness of a certain variable is related to the unobserved value of that variable in the data set. This question is crucial since methods dealing with missing data depend on assumptions made on the dependency in these mechanisms. For illustration, consider a random sample \( Z_i, i = 1, \ldots, n \). Suppose \( Z_i = (X_i, Y_i)^T \), where \( X_i \) is always observed but \( Y_i \) is subject to missingness. This is called a multivariate two pattern in the nomenclature of Little and Rubin (2002), and it is the pattern created by double (or two stage) sampling designs. Readers are referred to Rubin (1987) for discussion on general missing patterns. Denote \( \delta_i \) as the indicator of observation, i.e., \( \delta_i = 1 \) if \( Y_i \) is observed and \( = 0 \) if \( Y_i \) is missing. Let \( \Delta \) and \( X \) represent \( \{\delta_i\}_{i=1}^n \) and \( \{X_i\}_{i=1}^n \) respectively. Let \( Y_{obs} \) represent all the observed
Y_i's and let Y_{mis} represent all the missing Y_i's. Also let Z be the full data, including the missing values. The missing data mechanism is characterized by the conditional distribution of \Delta given Z, say \( f(\Delta|Z, \phi) \), where \( \phi \) denotes unknown parameters.

The data are called missing completely at random (MCAR), if the missingness does not depend on the values in Z, i.e.,

\[
f(\Delta|Z, \phi) = f(\Delta|\phi).
\]

The MCAR condition is very strong and not always satisfied. It is often achieved by the design of the experiment. A less restrictive assumption for missingness is that the missingness depends on \( \mathbf{X} \) and \( \mathbf{Y}_{obs} \), but not on \( \mathbf{Y}_{mis} \), that is

\[
f(\Delta|Z, \phi) = f(\Delta|\mathbf{X}, \mathbf{Y}_{obs}, \phi).
\]

The missing data mechanism in this case is called missing at random (MAR). The missing mechanism becomes nonignorable or not missing at random (NMAR) if the distribution of \( \Delta \) also depends on \( \mathbf{Y}_{mis} \).

When also assuming that the joint distribution of \((Z_i, \delta_i)\) is independent across units, in which case the missingness in one observation does not depend on other observations, the MCAR condition simplifies to

\[
P(\delta_i = 1 | Z_i) = \phi,
\]

with \( \phi \) being a constant. Accordingly, the data are missing at random if

\[
P(\delta_i = 1 | Z_i) = P(\delta_i = 1 | X_i). \tag{1.1}
\]

Rosenbaum and Rubin (1983) introduced an assumption called strongly ignorable missing at random, which assumes that \( Y_i \) and \( \delta_i \) are conditionally independent given the value of the always observable variable \( X_i \), i.e.,

\[
Y_i \perp \delta_i | X_i. \tag{1.2}
\]
Condition (1.2) is stronger than (1.1), it is very useful for dealing with missing data problems in nonparametric or semiparametric settings. The probability \( P(\delta_i = 1 \mid X_i) \) is called the response propensity score, it is shown that conditioning on propensity score can remove the bias caused by missing mechanisms under strongly ignorable missing at random (Little, 1986).

In this thesis, we focus on cases of MAR (including MCAR) for parametric models and use the strongly ignorable missing at random assumption in nonparametric and semiparametric settings. NMAR is the most realistic assumption in many situations, but dealing with data NMAR is analytically difficult and requires strong assumptions on missing mechanisms. In some empirical settings, the MAR assumption has been shown to give better performance than methods based on more natural NMAR assumption (Rubin, Stern, and Vehovar, 1995).

1.2 Estimating Equations and Empirical Likelihood

The empirical likelihood method introduced by Owen (1988, 1990) is a nonparametric method of statistical inference. It provides likelihood ratio statistics for parameters by profiling a nonparametric likelihood, which is analogous to that used for parametric models. Given random sample \( Z_1, \ldots, Z_n \), the empirical likelihood function of \( F \), the CDF for \( Z \), is

\[
L(F) = \prod_{i=1}^{n} dF(z_i) = \prod_{i=1}^{n} p_i,
\]

where \( p_i = dF(z_i) = Pr(Z_i = z_i) \). Obviously, (1.3) is maximized by the empirical distribution function \( F_n(z) = n^{-1} \sum_{i=1}^{n} I(Z_i \leq z) \). The advantage of empirical likelihood is discussed in Chapter 2 and Chapter 4. We are especially interested in empirical likelihood based inference for estimating equations when data are missing.

Estimating equations provide an highly flexible tool to describe parameters and corresponding statistics (Godambe, 1991; Heyde, 1997). Suppose i.i.d. random vari-
ables $Z_i \in \mathbb{R}^d, i = 1, \ldots, n$, are from an unknown distribution $F$, which is associated with a $p$-dimensional parameter $\theta$ and a $r$-dimensional estimating function $g(Z, \theta)$ with $E g(Z, \theta) = 0$. In the just identified case, i.e., when $p = r$, the true value $\theta_0$ may be estimated by solving the estimating equations

$$\frac{1}{n} \sum_{i=1}^{n} g(Z_i, \hat{\theta}) = 0 \quad (1.4)$$

for $\hat{\theta}$. In econometrics, there is great interest to the over identified case with $r > p$. In this case (1.4) usually has no solution. The generalized method of moments (GMM, Hansen, 1982) looks for an estimator $\hat{\theta}$ that comes close to solving (1.4). The empirical likelihood method can be used to handle both just identified and over identified cases.

The construction of empirical likelihood for estimating equations as discussed in Qin and Lawless (1994) is summarized below.

The profile empirical likelihood for $\theta$ is derived by maximizing (1.3) subject to restrictions

$$p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(Z_i, \theta) = 0. \quad (1.5)$$

The maximum may be found via Lagrange multipliers. Let

$$H = \sum_{i=1}^{n} \log p_i + \lambda \left(1 - \sum_{i=1}^{n} p_i\right) - n t^r \sum_{i=1}^{n} p_i g(Z_i, \theta),$$

where $\lambda$ and $t = (t_1, \ldots, t_r)^T$ are Lagrange multipliers. Taking derivatives with respect to $p_i$, by $\partial H / \partial p_i = 0$, we can get $\lambda = n$,

$$p_i = \frac{1}{n} \frac{1}{1 + t^r g(Z_i, \hat{\theta})}.$$

By the last restriction in (1.5), we also have

$$0 = \sum_{i=1}^{n} p_i g(Z_i, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + t^r g(Z_i, \hat{\theta})} g(Z_i, \theta),$$

from which $t$ can be determined in terms of $\theta$. The profile empirical likelihood function for $\theta$ is now defined as

$$L(\theta) = \prod_{i=1}^{n} \left\{ \frac{1}{n} \frac{1}{1 + t^r(\theta) g(Z_i, \theta)} \right\},$$
and the log empirical likelihood ratio is defined as

$$\ell(\theta) = \sum_{i=1}^{n} \log \{ 1 + \ell^r(\theta)g(Z_i, \theta) \}.$$ 

Under the conditions described in Qin and Lawless (1994), $L(\theta)$ attains the maximum at $\hat{\theta}$ in a neighborhood of the true parameter value $\theta_0$. It is then shown that the maximum empirical likelihood estimator $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where

$$V = \left\{ E \left( \frac{\partial g}{\partial \theta} \right)^T (Eg g^T)^{-1} E \left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1}.$$ 

Qin and Lawless (1994) also give a nonparametric version of Wilks' theorem, that is,

$$2\ell(\theta_0) - 2\ell(\hat{\theta}) \xrightarrow{d} \chi_p^2.$$ 

The property of Bartlett correction is also applicable to empirical likelihood (DiCiccio, Hall, and Romano, 1991; Chen and Cui, 2006), which is another parallel between empirical likelihood and parametric likelihood. The empirical likelihood ratio statistic can also be calibrated by bootstrap, which often gives superior coverage properties in practice (Hall and La Scala, 1990).

1.3 Statistical Issues in Microarray Data Analysis

Microarray technology has become a commonly used tool in biological research. Experiments using Affymetrix chips or cDNA arrays can provide a snap shot of the transcriptional activities of thousands of genes in a single experiment. It is thus possible to monitor the changes of the whole transcriptome over different time points, tissue types, or physiological states, which was previously impossible. There is also great interest in using microarray as a diagnosis tool in detecting important diseases. For general
information regarding microarrays, one can consult the book edited by Speed (2003) or numerous other references.

However, microarray experiments do pose serious challenges to statistical methodology. One problem is the presence of systematic bias caused by variations in experimental conditions. Examples include variations in the labeling efficiency, DNA spotting, amount of RNA, scanning conditions, among others. Thus normalization is required to remove systematic effects of confounding factors before making inference on biological problems. Commonly used methods for normalization of microarray data can be found in Cui, Kerr, and Churchill (2002); Dudoit et al. (2002); Yang et al. (2002); Irizarry et al. (2003); Bolstad et al. (2003).

Another challenge is due to the sheer volume of information obtained in a microarray experiment. It is the norm that data are collected on thousands of genes with only a small number of biological samples. Since statistical inferences are carried out on each and every gene, extreme multiple testing problems must be dealt with. The commonly used method in this situation is the false discovery rate approach (Benjamini and Hochberg, 1995), or empirical Bayes methods (Efron, Tibshirani, Storey, and Tuscher, 2001; Storey, 2002). A Bayesian model selection method has also been proposed in balancing false rejection and false nonrejection (Ishwaran and Rao, 2003).

There are also different approaches in selecting test statistics. In making inference on each gene separately, t-tests, ANOVA, or mixed models all can be used. Methods are also proposed in order to borrow strength across genes in estimating the gene-specific variance (Baldi and Long, 2001; Wright and Simon, 2003; Smyth, 2004; Cui et al., 2005). Since the expression measurements can demonstrate significant nonnormality, semiparametric models might be of advantage in certain situations (Strimmer, 2003; Newton, Noueiry, Sarkar, and Ahlquist, 2004). Recently, there are also attempts to integrate normalization and data analysis (Fan, Peng, and Huang, 2005; Huang, Wang, and Zhang, 2005).

For practitioners, there is another challenge posed by the high cost of microarray
chips. Thus designing experiments that can achieve good power within financial constraints is an important consideration.

1.4 Quantitative Trait Locus Mapping

Quantitative trait loci (QTLs) are genetic regions on a chromosome that control certain quantitative traits, such as crop yield or body fat. QTL mapping involves construction of genomic maps and testing for association between traits and polymorphic markers. A significant association provides evidence that a QTL is near the marker. The progress in sequencing the whole genomes of various organisms greatly facilitated QTL mapping experiments as genomic maps with dense marker spacing can more easily be constructed. General reviews of QTL-mapping methods can be found in Broman (2001) and Doerge (2001).

In searching for QTLs, one can do testing on single markers or perform interval mapping. For the interval mapping method, since markers are observed at known locations, the genotypes between the locations are missing, which will lead to mixture distributions at putative loci. Statistical models for QTL mapping include linear regression models (Cowen, 1989; Haley and Knott, 1992; Moreno-Gonzales, 1992), and likelihood based mixture models (Lander and Botstein, 1989). Multiple QTL models have been studied in Jansen (1993); Zeng (1994); Kao, Zeng, and Teasdale (1999) and others. There is also research in the generalization to different experimental designs (e.g., Song et al., 1999; Zeng et al., 2000) and to multiple and categorical traits (Jiang and Zeng, 1997; Henshall and Goddard, 1999). Nonparametric methods in QTL mapping have also been proposed (Kruglyak and Lander, 1995). Zou, Fine, and Yandell (2002) proposed semiparametric mixture modeling using empirical likelihood.

Since the entire genome (or at least several regions) is tested for the presence of QTL and the test statistics are not independent among loci, there is a common problem in
determining the significance threshold of the test statistic. In addition to asymptotic approximations (Dupuis and Siegmund, 1999), Doerge and Churchill (1996) proposed permutation testing and Zeng, Kao, and Basten (1999) discussed bootstrap resampling. As it is often the case that the effect of a QTL is small, research on better methods and experimental designs for detecting association between QTL and traits is ongoing.

Jansen and Nap (2001, 2004) used the term genetical genomics to describe QTL mapping when transcriptional abundance as measured by microarrays is used as the trait. More on this topic can be found in Chapter 3 and Chapter 4.

1.5 Thesis Organization

The following three chapters consist of three manuscripts on missing data methods and applications to biological problems.

Chapter 2 is a technical report, a shorter version of which is prepared for submission to Biometrika. We propose a nonparametric imputation method for data with missing values. The inference on the parameter defined by general estimating equations is performed using an empirical likelihood method. It is shown that nonparametric imputation together with empirical likelihood inference can reduce the bias and improve efficiency of the estimate when compared with inference using only complete cases of the data set. The confidence regions obtained by empirical likelihood demonstrate good coverage properties. Since our method is valid under very weak assumptions while also possessing the flexibility inherent to estimating equations and empirical likelihood, we expect that it can be applied to a wide range of problems. An example is given using mouse eye weight and gene expression data.

Chapter 3 is a paper published in Biometrics (online early version DOI:10.1111/j.1541-0420.2005.00491.x, with the print version to appear), in which we proposed a selective transcriptional profiling approach in improving efficiency and affordability of geneti-
cal genomics research. Genetical genomics is an approach that blends QTL mapping and microarray technology, and has shown great promise in dissecting gene regulation networks. However, the high cost of microarrays tends to limit the adoption of the standard genetical genomics approach. In a missing data framework, we proposed selective transcriptional profiling, in which only a subset of objects are subjected to microarray experiments. It is shown that this approach can significantly reduce experimental cost while still achieving satisfactory power.

Chapter 4 is a paper prepared for submission, in which we developed empirical likelihood based inference for multi-sample comparison problems using data with surrogate variables. It is shown that the empirical likelihood ratio statistic still has a chi-square calibration for parameters defined by estimating equations in the presence of auxiliary data. By applying this result to selective transcriptional profiling, we show that the idea of using relatively inexpensive trait data on extra individuals to improve the power of test for association between a QTL and gene transcriptional abundance also applies to the empirical likelihood based method. Thus selective transcriptional profiling approach can also be employed using nonparametric methods.

Chapter 5 is a brief summary and general discussion.

1.6 References


Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emporium Journal of Experimental Agriculture* 1, 129–142.


CHAPTER 2. Nonparametric Imputation of Missing Values for Estimating Equation Based Inference
Dong Wang and Song Xi Chen

Abstract: We propose a nonparametric imputation procedure for data with missing values and establish empirical likelihood based inference for parameters defined by general estimating equations. The imputation is carried out multiple times via a nonparametric estimator of the conditional distribution of the missing component given the always observable component of the random vector under study. The empirical likelihood is used to construct a profile likelihood for the parameter of interest. We demonstrate that the proposed nonparametric imputation can correct the selection bias in the missingness and empirical likelihood leads to more efficient parameter estimation. The proposed method is evaluated by simulation and an empirical study on the relationship between eye weight and gene transcriptional abundance of recombinant inbred mice.

KEY WORDS: Empirical likelihood; Estimating equations; Kernel estimator; Missing at random.

2.1 Introduction

Missing data are encountered in many statistical applications. A major undertaking in biological research is to integrate data generated by different experiments and technologies. Examples include the effort by genenetwork.org and other data depositories
to combine genetics, microarray data and phenotypes accumulated over the years in the study of recombinant inbred mouse lines (Wang, Williams, and Manly, 2003). But one problem in using measurements from multiple experiments is that often times different research projects choose to perform experiments on different subsets of mouse strains. This leads to the situation that only a portion of the strains have all the measurements, while other strains have missing measurements. The current practice of using only complete measurements is undesirable since the selection bias in the missingness can cause the parameter estimators to be inconsistent. Even in the absence of the selection bias, the complete measurements based inference is generally not efficient as it throws away the data with missing values.

Substantial research has been done on methods dealing with missing data problems in survey sampling, longitudinal data analysis, and multivariate analysis. Recently, there is increasing interest in nonrandomized observational studies in which the inference on the causal effect can be viewed as a missing data problem (Rosenbaum and Rubin, 1983). Little and Rubin (2002) is an excellent source of general information on missing data methods.

Most methods in missing data analysis employ certain parametric models. When a parametric model can be defined for all the variables, including those with missing values, the maximum likelihood method can be used for inference, which has been greatly facilitated by the research on the EM algorithm (Dempster, Laird, and Rubin, 1977). Multiple imputation is another popular method which creates multiple “complete” datasets by making random draws from the predictive distribution of the missing values, each dataset is analyzed using a standard complete-data method, and all the complete-data inferences are then combined to form one inference that properly reflects the uncertainty due to nonresponse (Rubin, 1987; Little and Rubin, 2002). Though the theoretical justification of multiple imputation is Bayesian, the idea of multiple imputation has been applied to other methods, such as the hot-deck imputation (Herzog and
Kim and Fuller (2004) proposed fractional hot deck imputation in the survey setting, in which multiple values are drawn from the same imputation cell of the missing observation, and a weight is assigned to each imputed value. If one specifies prior distributions on the parameters as well as on the missing variables, a fully Bayesian inference is available. To make inference under more relaxed assumptions, Robins, Rotnitzky, and Zhao (1994, 1995), and Robins and Rotnitzky (1995) proposed the weighted estimating equation (WEE) method using parametrically estimated propensity scores. WEE is more robust against model misspecification, although correct specification of the conditional distribution of the missing variables given the observed data is often needed to achieve the optimal efficiency.

Nonparametric methods have also been proposed for missing data problems. Titterington (1977) and Titterington and Mill (1983) discussed using a kernel method to estimate a multivariate density from data with incomplete observations. When the parameter of interest is the mean of the response variable, Cheng (1994) proposed using the kernel conditional mean estimator to impute the missing values. Hahn (1998), Hirano, Imbens, and Ridder (2003) and others studied the estimation of average treatment effects using nonparametrically estimated propensity scores. Since in treatment effect problems, the response of any unit can only be observed under one treatment, it can be seen as a missing data problem. In a similar setting, Chen, Hong, and Tarozzi (2004) discussed semiparametric efficient estimation using the generalized method of moments.

In this paper we consider the estimation of parameters defined by a set of estimating equations when missing values are present. The model framework of the general estimating equations is semiparametric and contains a large number of parameters. To suit the general nature of parameters defined by the estimating equations, we propose a nonparametric imputation procedure that generates multiple copies of the missing values via a kernel estimator of the conditional distribution of the missing variables given the fully observable variables. The imputed data and the original data set are combined to form
an expanded data set on which the estimating equations are constructed. An empirical likelihood ratio statistic for the parameter is formulated based on the expanded data via estimating equations where the estimating functions involving incomplete observations are averaged over the imputed values to reduce the variability caused by imputing from the conditional distribution function. The resulting maximum empirical likelihood estimator of the parameter is consistent and more efficient than the estimator based on the complete portion of the data only. In particular, when the number of the estimating equations is the same as the dimension of the parameter, the proposed estimator attains the semiparametric efficiency bound. Empirical likelihood (EL) proposed by Owen (1988, 1990) is a technique that provides a profile likelihood in nonparametric or semiparametric settings. The empirical likelihood for parameters defined by estimating equations as established in Qin and Lawless (1994) (without missing values) is a flexible and effective device for incorporating extra data information. Wang and Rao (2002) developed an empirical likelihood inference for the mean parameter with missing responses where the missing values were imputed with a kernel estimator of the conditional mean. This paper establishes empirical likelihood based inference for general parameters in the presence of missing values.

The paper is structured as follows. The proposed nonparametric imputation method is described in Section 2.2. Section 2.3 contains the main theoretical results. Results from two simulation studies are reported in Section 2.4. Section 2.5 analyzes a dataset on recombinant inbred mice by applying our method. Theoretical derivations are deferred to the Appendix.

### 2.2 Nonparametric Imputation

Let $Z_i = (X_i^*, Y_i^*)^T$ for $i = 1, \cdots, n$ be a set of independent and identically distributed random vectors, where $X_i$'s are $d_x$-dimensional and are always observable, and
$Y_i$'s are $d_y$-dimensional and are subject to missingness. Let $\theta$ be a $p$-dimensional parameter of the unknown distribution of $Z_i$ so that $E\{g(Z_i, \theta)\} = 0$, where $g(Z, \theta)$ represents a $r$-dimensional estimating function. Here $r \geq p$, that is, the number of estimating equations, $r$, is not smaller than the dimension of $\theta$. See Godambe (1991) and Boos (1992) for comprehensive reviews of estimating equation based inference. The interest of this paper is regarding the inference on $\theta$ when some of the $Y_i$'s are missing.

Define the missing indicator $\delta_i = 1$ when $Y_i$ is observed and $\delta_i = 0$ otherwise. We assume that $\delta$ and $Y$ are conditionally independent given $X$, namely the strongly ignorable missing at random condition proposed by Rosenbaum and Rubin (1983), which has also been assumed in Cheng (1994), Wang and Rao (2002) and other studies. This assumption implies

$$P(\delta = 1 \mid Y, X) = P(\delta = 1 \mid X) =: p(X),$$

where $p(X)$ is the response propensity score given $X$. A non-constant $p(X)$ indicates the existence of selection bias in the missingness.

Let $F(y \mid X_i)$ be the conditional distribution of $Y$ given $X = X_i$. For a missing $Y_i$, we propose to impute a value $\hat{Y}_i$ that is randomly generated from the following kernel estimator of $F(y \mid X_i)$ based on the complete portion (no missing values) of the sample:

$$\hat{F}(y \mid X_i) = \frac{1}{\sum_{j=1}^{n} \delta_j W\left(\frac{X_j - X_i}{h}\right) I(Y_j \leq y)} \sum_{j=1}^{n} \delta_j W\left(\frac{X_j - X_i}{h}\right).$$ (2.1)

Here $W(\cdot)$ is a $d_y$-dimensional kernel function that is a bounded and symmetric probability density function; $h$ is the bandwidth satisfying $h \to 0$ and $nh^{d_y} \to \infty$ as $n \to \infty$; and $I(\cdot)$ is the $d_y$-dimensional indicator function. In particular, we say $Y_i \leq y$ if each component of $Y_i$ is less than or equal to the corresponding component in $y$.

We can see that $\hat{Y}_i$ is effectively drawn from a discrete distribution such that the probability of selecting a $Y_i$ with $\delta_i = 1$ is

$$\frac{W\left((X_i - X_i)/h\right)}{\sum_{j=1}^{n} \delta_j W\left((X_j - X_i)/h\right)}.$$ (2.2)
To reduce the variability of the imputed estimating function based on a single draw, we make \( k \) independent draws \( \{Y_{iv}\}_{i=1}^{k} \) from \( \hat{F}(y|X_i) \) and use

\[
\bar{g}(\bar{Z}_i, \theta) = \delta_i g(Z_i, \theta) + (1 - \delta_i)k^{-1} \sum_{\nu=1}^{k} g(X_i, \bar{Y}_{i\nu}, \theta)
\]  

(2.3)

as the estimation function for the \( i \)-th observation. Now the expanded sample consists of \( \{Z_i\}_{i=1}^{r} \) where

\[
\bar{Z}_i = \begin{cases} 
Z_i, & \text{if } \delta_i = 1; \\
(X_i, \{\bar{Y}_{i\nu}\}_{\nu=1}^{k})^T, & \text{if } \delta_i = 0.
\end{cases}
\]  

(2.4)

A popular method of imputation is to impute a missing \( Y_i \) by the conditional mean of \( Y \) given \( X = X_i \) as proposed in Yates (1933) under a parametric regression model and in Cheng (1994) and Wang and Rao (2002) via the Nadaraya-Watson kernel estimator for the conditional mean. However, it may not work for other parameters, for instance, quantiles or correlation coefficients. Neither is it generally applicable to the case of missing covariates in a regression context. The proposed nonparametric imputation from the conditional distribution is more generally applicable in the setting of estimating equations. We will show that when \( \theta \) is a mean related parameter, the proposed imputation method leads to a parameter estimator that has the same efficiency as that obtained by the conditional mean imputation.

\section*{2.3 Empirical Likelihood}

With the imputed estimation equations, the usual estimating equation approach can be used to make inference on \( \theta \), since \( \bar{g}(\bar{Z}_i, \theta) \) is asymptotically unbiased. In this article, we carry out a likelihood type inference using empirical likelihood (EL).

Empirical likelihood possesses two key properties of a conventional parametric likelihood. One is that the log likelihood ratio converges to a chi-square distribution, namely Wilks' theorem; and the other is the delicate second order property of Bartlett correction. A major advantage of empirical likelihood is that it has no predetermined
shape of the confidence region, instead it produces regions that reflect the features of the dataset. The applications of empirical likelihood are greatly expanded when it is combined with estimating equations. For complete data, Qin and Lawless (1994) established the chi-square calibration of empirical likelihood applied to general estimating equations for both just identified and over identified cases. Chen and Cui (2006) established Bartlett correction for empirical likelihood for general estimating equations in the presence of nuisance parameters. Our proposal of using empirical likelihood in conjunction with nonparametric imputation is especially attractive, since it requires very few assumptions for both imputation and inference while it retains the flexibility inherent to empirical likelihood and estimating equations.

Based on the reconstructed sample \( \{\tilde{Z}_i\}_{i=1}^n \), the profile empirical likelihood for \( \theta \) is defined as

\[
L(\theta) = \sup \left\{ \prod_{i=1}^n p_i \left| p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n \tilde{g}(\tilde{Z}_i, \theta) = 0 \right\},
\]

where \( p_i \) represents the weight placed on \( \tilde{g}(\tilde{Z}_i, \theta) \). By introducing Lagrange multipliers and following the standard derivation of EL (Qin and Lawless, 1994), it can be shown that the optimal \( p_i \) is

\[
p_i = \frac{1}{n} \frac{1}{1 + t^*(\theta) \tilde{g}(\tilde{Z}_i, \theta)},
\]

where \( t(\theta) \) satisfies

\[
\frac{1}{n} \sum_i \frac{\tilde{g}(\tilde{Z}_i, \theta)}{1 + t^*(\theta) \tilde{g}(\tilde{Z}_i, \theta)} = 0.
\]

Let \( \ell(\theta) = -\log \{L(\theta)/n^{-n}\} \) be the log empirical likelihood ratio and \( \hat{\theta} \) be the maximum empirical likelihood estimator that maximizes \( L(\theta) \).

### 2.4 Main Results

In this section we establish the asymptotic normality of \( \hat{\theta} \) when using the imputation scheme described in Section 2.2, and derive the asymptotic distribution of the log
Let \( \theta_0 \) be the true value of the parameter, 
\[
V = \left\{ E\left( \frac{\partial g}{\partial \theta} \right) \Gamma^{-1} E\left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1},
\]
\[
\hat{\Gamma} = E \{ p(X) \text{Cov}(g|X) + E(g|X)E(g^*|X) \}, \quad \text{and}
\]
\[
\Gamma = E \{ p^{-1}(X) \text{Cov}(g|X) + E(g|X)E(g^*|X) \}.
\]

The asymptotic normality and the efficiency of \( \hat{\theta} \) are established in the following theorem.

**Theorem 1.** Under conditions C1-C7 given in the Appendix, as \( n \to \infty \) and \( \kappa \to \infty \),
\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \Sigma),
\]
where \( \Sigma = V E\left( \frac{\partial g}{\partial \theta} \right) \hat{\Gamma}^{-1} \Gamma^{-1} E\left( \frac{\partial g}{\partial \theta} \right)V \).

The estimator \( \hat{\theta} \) is consistent for \( \theta_0 \) and hence the potential bias in the missingness as indicated by the propensity score \( p(x) \) has been filtered out. In the absence of missing values, \( \hat{\Gamma} = \Gamma = E(gg^*) \). Hence
\[
\Sigma_{full} =: \Sigma = \left\{ E\left( \frac{\partial g}{\partial \theta} \right)^T (Egg^*)^{-1} E\left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1},
\]
which is the same as that in Qin and Lawless (1994) and is the efficiency based on the full data. Comparing \( \Sigma \) in the presence of missing data and \( \Sigma_{full} \) shows that the efficiency of \( \hat{\theta} \) based on the proposed imputation will be close to that based on the full data if either the proportion of missing data is low, i.e., \( p(X) \) is close to 1, or if \( E\{ p^{-1}(X) \text{Cov}(g|X) \} \) is small relative to \( E\{ E(g|X)E(g^*|X) \} \), namely when \( X \) is highly “correlated” with \( Y \).

In the case of \( \theta = EY \), \( \Sigma = E\{ \sigma^2(X)/p(X) \} + Var\{ m(X) \} \), where \( \sigma^2(X) = Var(Y|X) \) and \( m(X) = E(Y|X) \). Thus in this case, \( \hat{\theta} \) is asymptotically equivalent to the estimator proposed by Cheng (1994) based on the conditional mean imputation.
When \( r = p \), namely the number of estimating equations is the same as the dimension of \( \theta \),

\[
\Sigma = \left\{ E\left( \frac{\partial g}{\partial \theta} \right)^T \Gamma^{-1} E\left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1},
\]

which is the semiparametric efficiency bound for the estimation of \( \theta \) by similar argument to Theorem 1 of Chen, Hong, and Tarozzi (2004). If \( \kappa \), the number of draws in the imputation, is fixed, we have

\[
\begin{align*}
\Gamma &= E\left\{ \left\{ p^{-1}(X) + \kappa^{-1}(1 - p(X)) \right\} \text{Cov}(g|X) + E(g|X)E(g^*|X) \right\}, \\
\tilde{\Gamma} &= E\left\{ \left\{ p(X) + \kappa^{-1}(1 - p(X)) \right\} \text{Cov}(g|X) + E(g|X)E(g^*|X) \right\}.
\end{align*}
\]

Thus better efficiency is achieved for larger \( \kappa \) and this confirms the effect of the multiple draws. Our experience built on simulation studies suggests \( \kappa = 20 \) is sufficient for most applications.

Now consider the log empirical likelihood ratio for testing \( H_0 : \theta = \theta_0 \),

\[
\mathcal{R}(\theta_0) = 2\ell(\theta_0) - 2\ell(\hat{\theta}).
\]

Let \( I_r \) be the \( r \)-dimensional identity matrix. The next theorem shows that the ratio converges to a linear combination of independent chi-square distributions.

**Theorem 2.** Under the conditions C1-C7, as \( n \to \infty \) and \( \kappa \to \infty \),

\[
\mathcal{R}(\theta_0) \overset{L}{\to} Q^T \Omega Q,
\]

where \( Q \sim N(0, I_r) \) and \( \Omega = \Gamma^{1/2} \tilde{\Gamma}^{-1/2} E \left( \frac{\partial g}{\partial \theta} \right) V E \left( \frac{\partial g}{\partial \theta} \right)^T \tilde{\Gamma}^{-1/2} \Gamma^{1/2} \).

When there are no missing values, we have \( \Gamma = \tilde{\Gamma} = E(gg^*) \) and

\[
\Omega = E(gg^*)^{-1/2} E \left( \frac{\partial g}{\partial \theta} \right) \left\{ E\left( \frac{\partial g}{\partial \theta} \right)^T \left( E(gg^*) \right)^{-1} E\left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1} E \left( \frac{\partial g}{\partial \theta} \right)^T E(gg^*)^{-1/2},
\]

which is symmetric and idempotent with \( tr(\Omega) = p \). This means that \( \mathcal{R}(\theta_0) \) converges to \( \chi^2_p \), which is a version of Wilks’ theorem established in Qin and Lawless (1994).
When there are missing values, the usual Wilks’ Theorem for empirical likelihood is no longer achievable due to a mis-match between the variance of \( n^{-1/2} \sum_{i=1}^{n} \tilde{g}(Z_i, \theta_0) \) and the limit of \( n^{-1} \sum_{i=1}^{n} \tilde{g}(Z_i, \theta_0) \tilde{g}'(Z_i, \theta_0) \). When there are no imputed values, the empirical likelihood is able to match the two via internal studentization. However, this ability is lost when imputed values are used. This phenomenon also happens to the empirical likelihood when a nuisance parameter is replaced by a plugged-in estimator as revealed by Hjort, McKeague, and Van Keilegom (2004).

In the case of \( \theta_0 = EY \), \( \mathcal{R}(\theta_0) \xrightarrow{L} (V_1(\theta_0)/V_2(\theta_0))\chi^2_1 \), where

\[
V_1(\theta_0) = E\{\sigma^2(X)/p(X)\} + \text{Var}\{m(X)\},
\]
\[
V_2(\theta_0) = E\{\sigma^2(X)p(X)\} + \text{Var}\{m(X)\}.
\]

This is the same limiting distribution as that when the missing values are imputed from a kernel estimator of the conditional mean as established in Wang and Rao (2002).

We have potentially several methods for the construction of a confidence region for \( \theta \), which can be readily transformed to test statistics for a hypothesis regarding \( \theta \). One is based on the explicit estimation of the covariance matrix \( \Sigma \) and the asymptotic normality given in Theorem 1. For example, a consistent estimator of \( \Gamma \), \( \hat{\Gamma}_n \), can be obtained by the following,

\[
\hat{p}_n(x) = \frac{\sum_{i=1}^{n} W_h(x - X_i) \delta_i}{\sum_{i=1}^{n} W_h(x - X_i)},
\]
\[
\hat{m}_{gn}(x) = \frac{\sum_{i=1}^{n} W_h(x - X_i) \delta_i g(Z_i, \theta)}{\sum_{i=1}^{n} W_h(x - X_i) \delta_i},
\]
\[
\hat{\Sigma}_n(x) = \frac{\sum_{i=1}^{n} W_h(x - X_i) \delta_i \{g(Z_i, \theta) - \hat{m}_{gn}(X_i)\}\{g(Z_i, \theta) - \hat{m}_{gn}(X_i)\}^T}{\sum_{i=1}^{n} W_h(x - X_i) \delta_i},
\]
\[
\hat{\Gamma}_n = n^{-1} \sum_{i=1}^{n} \left\{ \hat{p}_n^{-1}(X_i) \hat{\Sigma}_n(X_i) + \hat{m}_{gn}(X_i) \hat{m}_{gn}^T(X_i) \right\},
\]

where \( W_h(X) = h^{-d}W(X/h) \). Consistent estimators for \( \hat{\Gamma} \) and \( E(\partial g/\partial \theta) \) can be obtained similarly. We can thus plug in these estimators to obtain an estimator for \( \Sigma \) and a confidence region for \( \theta_0 \) based on Theorem 1. Another method is to estimate the matrix
\( \Omega \) in Theorem 2 and then use Fourier inversion or a Monte Carlo method for the linear combinations of chi-squares. Despite the loss of Wilks' theorem, empirical likelihood confidence regions based on Theorem 2 are still attractive in terms of having natural shape and orientation free of subjective intervention, which are some of the attractions of likelihood based confidence regions.

Methods like that of (2.6) involve estimating extra parameters and may encounter computational difficulties. We propose the following bootstrap procedure to profile the distribution of the weighted chi-square distribution in Theorem 2 for its simplicity and accuracy. Bootstrap for imputed survey data is discussed in Shao and Sitter (1996), in which they considered ratio (random) imputation and regression (random) imputation. We use the following bootstrap procedure in which the bootstrap data set is imputed in the same way that the original data set was imputed:

1. Draw a simple random sample \( \mathbf{X}_n^\ast = \{(\mathbf{Z}_i^\ast, \delta_i^\ast) : i = 1, \ldots, n\} \) with replacement from the sample \( \mathbf{X}_n = \{(\mathbf{Z}_i, \delta_i) : i = 1, \ldots, n\} \) defined in (2.4).

2. Let \( \mathbf{X}_{nc}^\ast = \{(\mathbf{Z}_i^\ast, \delta_i^\ast) : \delta_i^\ast = 1\} \) be the portion of \( \mathbf{X}_n^\ast \) without imputed values and \( \mathbf{X}_{nm}^\ast = \{(\mathbf{Z}_i^\ast, \delta_i^\ast) : \delta_i^\ast = 0\} \) be the set of vectors in the bootstrap sample with imputed values. Then replace all the imputed \( Y \) values in \( \mathbf{X}_{nm}^\ast \) using the proposed imputation method where the estimation of the conditional distribution is based on \( \mathbf{X}_{nc}^\ast \). Let \( \mathbf{X}_n^\ast \) be the final bootstrap resample.

3. Let \( \mathcal{L}^\ast(\hat{\theta}) \) be the empirical likelihood ratio based on the re-imputed data set \( \mathbf{X}_n^\ast \), \( \hat{\theta}_\ast \) be the corresponding maximum empirical likelihood estimator, and \( \mathcal{R}_\theta(\hat{\theta}) = 2\mathcal{L}^\ast(\hat{\theta}) - 2\mathcal{L}^\ast(\hat{\theta}_\ast) \).

4. Repeat the above steps \( B \)-times for a large integer \( B \) and obtain \( B \) bootstrap values \( \{\mathcal{R}_\theta(\hat{\theta})\}_{b=1}^B \).

Then apply standard bootstrap formulas and their Monte Carlo approximation to obtain the confidence regions for \( \theta_0 \). Similar to Shao and Sitter (1996), our method requires that \( \mathbf{X}_n^\ast \) carries the identification flag \( \delta \) to locate missing values in the bootstrap.
sample. We state the consistency of the bootstrap estimator in the following theorem.

**Theorem 3.** Under Conditions C1-C7 and conditioning on the original sample \( x_n \),

\[
\mathcal{R}^*(\hat{\theta}) \overset{c}{\rightarrow} Q^*\Omega^*Q,
\]

with \( Q \sim N(0, I_r) \), and \( \Omega^* \rightarrow \Omega \) in probability as \( n \rightarrow \infty, \kappa \rightarrow \infty \).

Since the asymptotic distribution of \( R(\theta_0) \) is determined by \( \Omega \) only, we can use the empirical distribution of \( \mathcal{R}^*(\hat{\theta}) \) based on \( \{R^*_b(\hat{\theta})\}_{b=1}^B \) in Step 4 to calibrate the distribution of \( R(\theta_0) \) and to construct confidence regions of \( \theta_0 \). The performance of this method is demonstrated in the simulation studies presented in the next section.

### 2.5 Simulation Results

We now present results from two simulation studies. In each case, the proposed empirical likelihood inference based on the nonparametric imputation method are compared with the empirical likelihood inference based on (1) the complete observations only by ignoring data with missing values and (2) the full set of observations since the missing values are known in a simulation. The usage of the nonparametric imputation will remove the selection bias present in the estimates based on the complete cases only and is more efficient due to utilizing more data information. Obtaining the full data based estimator (assuming missing values are known) allows us to gauge how far away the proposed imputation based estimator is from the ideal case.

We also compare the proposed method with a version of the inverse probability weighted generalized method of moments (IPW-GMM) described in Chen et al. (2004). In particular, it is based on the fact that

\[
E\left\{ g(Z_i, \theta_0) \frac{P(\delta_i = 1)}{p(X_i)} \Big| \delta_i = 1 \right\} = 0.
\]
Based on the usual formulation of the generalized method of moments (GMM, Hansen, 1982), the weighted GMM estimator for $\theta_0$ considered in our simulation is

$$\hat{\theta} = \arg \min_{\theta} \left\{ \frac{1}{n_c} \sum_{i=1}^{n} \delta_i g(Z_i, \theta) \frac{1}{\hat{p}(X_i)} \right\}^T A_T \left\{ \frac{1}{n_c} \sum_{i=1}^{n} \delta_i g(Z_i, \theta) \frac{1}{\hat{p}(X_i)} \right\},$$

where $n_c$ is the number of complete cases, $A_T$ is a stochastic positive semidefinite weighting matrix, and $\hat{p}(X_i)$ is a consistent estimator for $p(X_i)$. The difference between the weighted GMM method we use and that of Chen et al. (2004) is that we used a kernel based estimator for $p(X_i)$, instead of the sieve estimator described in Chen et al. (2004). The bandwidth for $\hat{p}(X_i)$ is the one that admits the smallest empirical mean square error among several bandwidths that we experimented, including the theoretically optimal value.

### 2.5.1 Correlation Coefficient

In the first simulation, the parameter $\theta$ is the correlation coefficient between two random variables $X$ and $Y$, where $X$ is always observed, but $Y$ is subject to missingness. We first generate bivariate random vectors $(X_i, U_i)^T$ from a skewed bivariate $t$-distribution (Azzalini and Capitanio, 2003) with five degrees of freedom, mean $(0,0)^T$, shape parameter $(4,1)^T$, and dispersion matrix

$$\Omega = \begin{bmatrix} 1 & .955 \\ .955 & 1 \end{bmatrix}.$$  

Then we let $Y_i = U_i - 1.2X_i I(X_i < 0)$. The vector $(X_i, Y_i)^T$ has mean $(0, .304)$ and correlation coefficient $0.676$.

We consider three missing mechanisms:

- **Missing Mechanism I.** $p(x) = (0.3 + 0.175|x|)I(|x| < 4) + I(|x| \geq 4);$  
- **Missing Mechanism II.** $p(x) = 0.65$ for all $x;$  
- **Missing Mechanism III.** $p(x) = 0.5I(x > 0) + I(x \leq 0).$
Table 2.1 Inference for the correlation coefficient with missing values. The four methods considered are empirical likelihood (EL) using the full data, EL using only complete cases, inverse probability weighting based generalized method of moments (Weight-GMM), and EL using the nonparametric imputation (N. Imputation).

The nominal coverage level of the confidence intervals is 0.95.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>MSE</th>
<th>Coverage</th>
<th>Length of CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>0.0018</td>
<td>0.0899</td>
<td>0.0081</td>
<td>0.937</td>
<td>0.3543</td>
</tr>
<tr>
<td>Missing Mechanism I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Cases</td>
<td>0.0597</td>
<td>0.1262</td>
<td>0.0195</td>
<td>0.863</td>
<td>0.4896</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>0.0057</td>
<td>0.1105</td>
<td>0.0123</td>
<td>0.761</td>
<td>0.3568</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0050</td>
<td>0.1013</td>
<td>0.0103</td>
<td>0.948</td>
<td>0.4888</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>0.0047</td>
<td>0.0605</td>
<td>0.0037</td>
<td>0.949</td>
<td>0.2514</td>
</tr>
<tr>
<td>Missing Mechanism I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Cases</td>
<td>0.0727</td>
<td>0.0776</td>
<td>0.0113</td>
<td>0.849</td>
<td>0.3269</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>0.0072</td>
<td>0.0755</td>
<td>0.0058</td>
<td>0.800</td>
<td>0.2479</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>0.0076</td>
<td>0.0695</td>
<td>0.0049</td>
<td>0.953</td>
<td>0.3239</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Mechanism II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-0.0007</td>
<td>0.0753</td>
<td>0.0057</td>
<td>0.957</td>
<td>0.3146</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.0067</td>
<td>0.0688</td>
<td>0.0048</td>
<td>0.793</td>
<td>0.2338</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0004</td>
<td>0.0648</td>
<td>0.0041</td>
<td>0.957</td>
<td>0.2841</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Mechanism III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-0.0905</td>
<td>0.1000</td>
<td>0.0182</td>
<td>0.782</td>
<td>0.3955</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.0035</td>
<td>0.0747</td>
<td>0.0060</td>
<td>0.773</td>
<td>0.2751</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0055</td>
<td>0.0677</td>
<td>0.0046</td>
<td>0.946</td>
<td>0.2862</td>
</tr>
</tbody>
</table>
Here Missing Mechanism II is missing completely at random. The other two, however, prescribe selection bias in the missingness.

Let $\mu_x$ and $\mu_y$ be the means, and $\sigma_x^2$ and $\sigma_y^2$ be the variance of $X$ and $Y$, respectively. In the construction of the EL for $\theta$ (Owen, 1990), $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ are treated as nuisance parameters.

Table 2.1 contains the bias and standard error of the four estimators considered in the simulation based on 1000 simulations with sample size $n = 100$ and 200 respectively. It also contains the EL confidence intervals using the full data set, complete cases only, and the proposed nonparametric imputation method (N. Imputation) at the nominal level of 95%. They are all based on the proposed bootstrap calibration method with $B = 2000$. When using the nonparametric imputation method, $k = 20$ values were drawn for each missing $Y_t$. The confidence interval based on the weighted GMM are calibrated using the asymptotic normal approximation with the covariance matrix estimated by a kernel method similar to that of (2.6).

It is clear from Table 2.1 that the nonparametric imputation method significantly reduces bias compared to the inference based only on complete cases when the data are missing at random but not missing completely at random. The estimator based on the completely observed data suffers quite severe bias under missing mechanisms I and III. As expected, imputation reduces the variance of the estimator relative to the method based on complete cases only under all three missing mechanisms, including the case of missing completely at random. The weighted-GMM method can also reduce the bias compared to complete case analysis, but tends to have larger variance than that of the proposed estimator. Confidence intervals based on complete case analysis and the weighted-GMM method can have severe under-coverage: the former is due to the selection bias and the latter is due to the poor performance of normal approximation. The proposed confidence intervals have satisfactory coverage close to the nominal level.
Table 2.2  Inference for parameters in a logistic regression model with co-
variates subject to missingness. The four methods considered
are empirical likelihood (EL) using the full data, EL using only
complete cases, inverse probability weighting based generalized
method of moments (Weight-GMM), and EL using the nonpara-
metric imputation (N. Imputation). The nominal level of cover-
age for confidence intervals is 0.95.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>MSE</th>
<th>Coverage</th>
<th>Length of CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_0 = -1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>-0.0035</td>
<td>1.244</td>
<td>1.549</td>
<td>0.967</td>
<td>5.380</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-1.622</td>
<td>1.489</td>
<td>4.847</td>
<td>0.901</td>
<td>6.429</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.3113</td>
<td>1.402</td>
<td>2.061</td>
<td>0.932</td>
<td>5.107</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>0.0645</td>
<td>1.279</td>
<td>1.640</td>
<td>0.953</td>
<td>5.368</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>0.0270</td>
<td>0.4270</td>
<td>0.1831</td>
<td>0.965</td>
<td>1.835</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>0.4070</td>
<td>0.4995</td>
<td>0.4152</td>
<td>0.908</td>
<td>2.308</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>0.1200</td>
<td>0.4795</td>
<td>0.2443</td>
<td>0.935</td>
<td>1.722</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0030</td>
<td>0.4346</td>
<td>0.1889</td>
<td>0.951</td>
<td>1.828</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = -1.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>-0.0766</td>
<td>0.5009</td>
<td>0.2568</td>
<td>0.976</td>
<td>2.172</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-0.0664</td>
<td>0.5539</td>
<td>0.3112</td>
<td>0.975</td>
<td>2.506</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.0663</td>
<td>0.5589</td>
<td>0.3168</td>
<td>0.837</td>
<td>1.513</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0330</td>
<td>0.5512</td>
<td>0.3049</td>
<td>0.940</td>
<td>2.037</td>
</tr>
<tr>
<td></td>
<td>$\beta_0 = -1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>-0.0591</td>
<td>0.9506</td>
<td>0.9073</td>
<td>0.951</td>
<td>3.760</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-1.634</td>
<td>1.136</td>
<td>3.960</td>
<td>0.774</td>
<td>4.621</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.2787</td>
<td>1.053</td>
<td>1.187</td>
<td>0.933</td>
<td>3.895</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0196</td>
<td>0.9728</td>
<td>0.9468</td>
<td>0.952</td>
<td>3.907</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>0.0331</td>
<td>0.3252</td>
<td>0.1069</td>
<td>0.955</td>
<td>1.338</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>0.4018</td>
<td>0.3873</td>
<td>0.3115</td>
<td>0.829</td>
<td>1.645</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>0.0999</td>
<td>0.3663</td>
<td>0.1441</td>
<td>0.925</td>
<td>1.327</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>0.0146</td>
<td>0.3378</td>
<td>0.1143</td>
<td>0.950</td>
<td>1.413</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = -1.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Data</td>
<td>-0.0384</td>
<td>0.3705</td>
<td>0.1387</td>
<td>0.968</td>
<td>1.537</td>
</tr>
<tr>
<td>Complete Cases</td>
<td>-0.0402</td>
<td>0.4153</td>
<td>0.1741</td>
<td>0.961</td>
<td>1.719</td>
</tr>
<tr>
<td>Weight-GMM</td>
<td>-0.0443</td>
<td>0.4220</td>
<td>0.1801</td>
<td>0.850</td>
<td>1.148</td>
</tr>
<tr>
<td>N. Imputation</td>
<td>-0.0159</td>
<td>0.4152</td>
<td>0.1726</td>
<td>0.967</td>
<td>1.726</td>
</tr>
</tbody>
</table>
2.5.2 Generalized Linear Models with Missing Covariates

The problem of missing covariates in generalized linear models (GLM’s) has been the focus of intensive research due to its application in clinical trials and other settings. Commonly used methods in dealing with missing data for GLM have recently been reviewed in Ibrahim, Chen, Lipsitz, and Herring (2005). Empirical likelihood for GLM’s in the complete data case was first studied by Kolaczyk (1994). Application of the empirical likelihood method to GLM’s can help overcome some difficulties with parametric likelihood, especially in the aspect of overdispersion. Here we consider a logistic regression model with binary response variable $Y$ and covariates $X = (X_1, X_2)^T$. We suppose $\text{logit}(P(Y_i = 1)) = -1 + X_{1i} - 1.5X_{2i}$, $X_{1i} \sim N(3, 0.5^2)$, and $X_{2i}$ being binary with $\text{logit}(P(X_{2i} = 1)) = -1 + 0.5X_{1i}$. Also suppose $Y_i$ and $X_{1i}$ are always observable while $X_{2i}$ is subject to missingness with the missing mechanism $\text{logit}(P(X_{2i} \text{ is missing})) = 0.5 + X_{1i} - 3Y_i$.

When no missing data are involved, the empirical likelihood analysis for the logistic model simply involves the estimating equations

$$
\sum_{i=1}^{n} X_i \{Y_i - \pi(X_i' \beta)\} = 0,
$$

(2.7)

with $\beta$ being the parameter vector and $\pi(z) = \exp(z)/(1 + \exp(z))$. We can use the nonparametric imputation method proposed in Section 2.2 to impute the missing $X_{2i}$ values. In this case, the components of the data vector that are always observed are $X_{1i}$ and $Y_i$. Though in Section 2.2, we suppose that these components are continuous, binary $Y_i$ values can be easily accommodated by splitting the data into two parts according to the value of $Y_i$, and then applying the imputation scheme to each part of data, smoothing on continuous $X_{1i}$ values. The estimator for $\beta$ based on the nonparametric imputation could be calculated using a modified version of the fitting procedure described in Chapter 2 of McCullagh and Nelder (1983). Table 2.2 shows the inference on $\beta$ using different methods, the nominal coverage level of confidence regions is 95% and the sample size
is either 150 or 250. Like the results for the correlation coefficient, empirical likelihood based analyses on the full data (no missing values), complete observations only, and the expanded dataset from nonparametric imputation, are calibrated by bootstrap with 2000 bootstrap samples taken for each analysis. Results from the weighted-GMM were derived using the asymptotic normal approximation. A total of 1000 Monte Carlo random samples were generated for this study.

For parameters $\beta_0$ and $\beta_1$, the mean squared error of the proposed estimator are several folds smaller than that based on complete observations only; the proposed method also leads to a reduction in the mean square error by as much as 20% relative to the weighted-GMM. All three methods give similar mean squared errors for the parameter $\beta_2$. The confidence intervals based on only complete observations or the weighted-GMM tend to show notable undercoverage, while the proposed confidence intervals have satisfactory coverage levels for all parameters.

### 2.6 Empirical Example

Microarray technology provides a powerful tool in molecular biology by measuring the expression levels of thousands of genes simultaneously. One problem of interest is to test whether the expression levels of genes are related to a measurement like body weight, food consumption, or bone density. This is usually the first step in uncovering roles that a gene plays in important pathways. The BXD recombinant inbred strains of mouse were derived from crosses between C57BL/6J (B6 or B) and DBA/2J (D2 or D) strains (Williams, Gu, Qi, and Lu, 2001). Around one hundred BXD strains have been established by researchers at University of Tennessee and the Jackson Laboratory. A remarkable variety of phenotype data have been accumulated for BXD mice over the years (Pierce, Lu, Gu, Silver, and Williams, 2004). Recently, microarray data on BXD strains have also been collected. When it is of interest to investigate whether a trait
measurement like body weight is related to the expression levels of certain genes, it often occurs that trait data or microarray measurements are not available for all strains. The most common practice is to use only cases with complete observations.

The trait we consider is the fresh eye weight measured on 83 BXD strains by Zhai, Lu, and Williams (ID 10799, BXD phenotype data base). The Hamilton Eye Institute Mouse Eye M430v2 (November05) RMA Data Set contains the measurements of the expression levels in the eye on 39,000 transcripts. In generating the data, pooled RNA samples were hybridized to Affymetrix M430 2.0 arrays. This particular data set was processed using the RMA protocol (Bolstad, Irizarry, Astrand, and Speed, 2003). It is of interest to test whether the fresh eye weight is related to the expression levels of certain genes. But the microarray data are only available for 45 out of the 83 BXD mouse strains for which fresh eye weight measurements are available. The web server at genenetwork.org performs a test for the significance of the correlation coefficient using only complete cases with Fisher’s z transformation, but methods utilizing all observations might be preferred to reduce bias and improve efficiency. We will use the proposed imputation method to generate an expanded data set and carry out linear regression analysis and inference on the correlation coefficient.

We conduct four separate simple linear regression analysis of the eye weight on the expression level of four genes respectively. The genes are H3071E5, Slc26a8, Rps16, and Tex9, which are identified by the corresponding probe names in Table 2.3. Here we have missing covariates in our analysis. The missing gene expression levels are imputed from a kernel estimator of the conditional distribution of the gene expression level given the fresh eye weight. The smoothing bandwidths were selected based on the cross-validation method, which is 1.5 for the first three probes in Table 2.3 and 1.8 for probe 14455835_at.

Table 2.3 reports empirical likelihood estimates of the intercept and slope parameters and their 95% confidence intervals based on the proposed nonparametric imputation method. It also contains results from a conventional parametric regression analysis us-
Parameter estimates and confidence intervals (shown in parentheses) based on a simple linear regression model using parametric method with complete cases only and the empirical likelihood method using nonparametric imputation. For parametric inference, the confidence intervals for the intercept and slope are obtained using quantiles of the t distribution, and the confidence interval for the correlation coefficient is obtained by Fisher’s z transformation.

<table>
<thead>
<tr>
<th>Probe</th>
<th>Complete Cases Only (parametric)</th>
<th>Nonparametric Imputation with EL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td></td>
</tr>
<tr>
<td>1444597.at</td>
<td>-21.99 (-40.97, -2.998)</td>
<td>-15.69 (-37.02, 5.209)</td>
</tr>
<tr>
<td>1441747.at</td>
<td>73.59 (49.45, 97.73)</td>
<td>67.28 (38.34, 95.87)</td>
</tr>
<tr>
<td>1455835.at</td>
<td>-13.52 (-31.08, 4.041)</td>
<td>-8.090 (-26.76, 10.18)</td>
</tr>
<tr>
<td>1453360.a.at</td>
<td>-23.81 (-46.12, -1.507)</td>
<td>-14.66 (-38.57, 8.776)</td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td></td>
</tr>
<tr>
<td>1444597.at</td>
<td>10.16 (5.720, 14.59)</td>
<td>8.736 (2.688, 14.21)</td>
</tr>
<tr>
<td>1441747.at</td>
<td>-6.352 (-9.294, -3.411)</td>
<td>-5.561 (-9.431, -1.471)</td>
</tr>
<tr>
<td>1455835.at</td>
<td>6.766 (3.371, 10.16)</td>
<td>5.754 (1.948, 9.236)</td>
</tr>
<tr>
<td>1453360.a.at</td>
<td>5.101 (2.588, 7.613)</td>
<td>4.094 (0.8733, 6.979)</td>
</tr>
<tr>
<td></td>
<td>Correlation Coefficient</td>
<td></td>
</tr>
<tr>
<td>1444597.at</td>
<td>0.5757 (0.3395, 0.7436)</td>
<td>0.4426 (0.1321, 0.6977)</td>
</tr>
<tr>
<td>1441747.at</td>
<td>-0.5533 (-0.7285, -0.3102)</td>
<td>-0.4319 (-0.6809, -0.0761)</td>
</tr>
<tr>
<td>1455835.at</td>
<td>0.5256 (0.2744, 0.7097)</td>
<td>0.4151 (0.0755, 0.6613)</td>
</tr>
<tr>
<td>1453360.a.at</td>
<td>0.5296 (0.2996, 0.7124)</td>
<td>0.4024 (0.1013, 0.6846)</td>
</tr>
</tbody>
</table>
ing only the complete observations, assuming independent and identically normally distributed residuals. Table 2.3 shows that these two inference methods can produce quite different parameter estimates and confidence intervals. The difference in parameter estimates is as large as 50% for the intercept and 25% for the slope parameter. Table 2.3 also reports estimates and confidence intervals of the correlation coefficients using the proposed method and Fisher’s $z$ transformation. The latter is based on the complete observations only and is the method used by genenetwork.org. We observe again differences between the two methods despite not being significant at 5% level. The largest difference of about 25% is registered at the gene represented by probe 1444597.at. As indicated earlier, part of the differences may be the estimation bias of the complete observation based estimators as they are unable to filter out selection bias in the missingness.

Though the simple linear regression model with the expression value of one gene as the single covariate is overly simplistic in most occasions, it nonetheless provides important information in revealing the data structure, and the suggested advantage of using the nonparametric imputation should also apply to more sophisticated analysis. It should be noted that in analyzing microarray data, thousands of hypotheses are tested simultaneously using the false discovery rate approach. So differences in $p$-values obtained by complete case analysis and the nonparametric imputation could decide whether a gene is selected for further study. To date, the owner of the mouse eye microarray data set has not released it for global analysis, so a study on all genes in the data set has not been attempted. It will be of interest to see how the nonparametric imputation method performs when combined with a false discovery rate selection approach when the data set is available for global analysis.
2.7 References


Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emporium Journal of Experimental Agriculture* 1, 129–142.
Define $f(x)$ as the probability density function of $X$, $d_x$ as the dimension of $X$. Also we use $\theta_0$ to represent the real value of the parameter of interest, and denote $m_y(x) = E\{g(X,Y,\theta_0)|X = x\}$, as the conditional expectation of $g(X,Y,\theta_0)$ given $X$. The following conditions are needed in the proofs of the lemmas and theorems.

C1: The functions $p(x)$, $f(x)$ and $m_y(x)$ have bounded second partial derivatives, and $\inf_x p(x) \geq c_0$ for some $c_0 > 0$.

C2: The estimating function $g(x,y,\theta_0)$ has bounded second partial derivative with regard to $x$, and $E\|g(Z,\theta_0)\|^4 < \infty$.

C3: The kernel function $W$ is a nonnegative, symmetric and bounded probability density function with finite second moments.

C4: The bandwidth $h$ satisfies $nh^{d_x} \rightarrow \infty$ and $\sqrt{n}h^2 \rightarrow 0$ as $n \rightarrow \infty$.

C5: The estimating function $g(z,\theta)$ is continuous in $\theta$ in a neighborhood of the true value $\theta_0$, $\|\partial g(z,\theta)/\partial \theta\|$ and $\|g(z,\theta)\|^3$ are bounded by some integrable functions in this neighborhood.

C6: The matrices $\Gamma$ and $\tilde{\Gamma}$ are, respectively, positive definite with the smallest eigenvalue bounded away from zero, and $E[\partial g(z,\theta)/\partial \theta]$ has full column rank $p$.

C7: The second derivative $\partial^2 g(z,\theta)/\partial \theta \partial \theta^T$ is continuous in $\theta$ in a neighborhood of the true value $\theta_0$, and $\|\partial^2 g(z,\theta)/\partial \theta \partial \theta^T\|$ is bounded by some integrable function in the neighborhood.

Conditions C5-C7 are usual conditions for empirical likelihood based inference for parameters defined by estimating equations. Requiring $p(X)$ bounded away from zero implies that data cannot be missing with probability 1 anywhere in the domain of $X$, which is required for nonparametric inference. Existence of the derivatives and condition C4 are needed for arguments using Taylor expansions in the proof. Condition C4 limits $d_x < 4$. Lemma 1 to Theorem 3 can be extended to cover $d_x \geq 4$ case if, in C1-C4, we
require that \( p(x), f(x), m_y(x), \) and \( g(x, y, \theta_0) \) have partial derivatives in \( x \) up to order \( \xi \) with \( 2\xi > d_x, \sqrt{n}h^\xi \to 0 \), and the kernel function \( W(\cdot) \) has order \( 2\xi \). Also we need to replace (2.2) and (2.3) with

\[
\frac{\delta_i|W\{(X_i - X_i)/h\}|}{\sum_{j=1}^{n} \delta_j|W\{(X_j - X_i)/h\}|}, \quad \text{and}
\]

\[
g(\tilde{Z}_i, \theta) = \delta_i g(Z_i, \theta) + (1 - \delta_i) \kappa^{-1} \sum_{\nu=1}^{\kappa} g(X_i, \tilde{Y}_i, \theta) \Lambda_i,
\]

where \( \Lambda_i = \frac{\sum_{j=1}^{n} \delta_j|W\{(X_j - X_i)/h\}|}{\sum_{j=1}^{n} \delta_j|W\{(X_j - X_i)/h\}|} \text{Sign}[W\{(X_i - X_i)/h\}].
\]

**Lemma 1.** Assume that conditions C1-C4 are satisfied, \( n \to \infty \) and \( \kappa \to \infty \), then

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(\tilde{Z}_i, \theta_0) \xrightarrow{L} N(0, \Gamma),
\]

where \( \Gamma = E\{p^{-1}(X)\text{Cov}(g|X) + E(g|X)E(g^2|X)\}. \)

Before we prove Lemma 1, we state the following proposition, which is a direct consequence of Lemma 1 in Schenker and Welsh (1988). We use it in the proof.

**Proposition 1.** Let \( \{V_i\} \) be a sequence of random variables such that, for some function \( h \), as \( n \to \infty \), \( h(V_1, \ldots, V_n) \xrightarrow{L} \Xi \), where \( \Xi \) has a distribution function \( G \). If \( \{U_i\} \) is a sequence of random variables such that

\[
P\{U_n - h(V_1, \ldots, V_n) \leq s \mid V_1, \ldots, V_n\} \to F(s)
\]

almost surely for all \( s \in \mathbb{R} \), where \( F \) is a continuous distribution functions, then

\[
P(U_n \leq t) \to (G \ast F)(t),
\]

for all \( t \in \mathbb{R} \), where \( \ast \) denotes the convolution operator.

**Proof of Lemma 1:** Let \( u \in \mathbb{R}^v \) and \( \|u\| = 1 \). Also let \( g_u(Z, \theta_0) = u^*g(Z, \theta_0) \) and \( \tilde{g}_u(\tilde{Z}, \theta_0) = u^*\tilde{g}(\tilde{Z}, \theta_0) \). First we need to show that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{g}_u(\tilde{Z}_i, \theta_0) \xrightarrow{L} N(0, u^*\Gamma u),
\]
and then use the Cramér-Wold device to prove Lemma 1. Define

\[
m_{g_0}(x) = E\{g_0(X, Y, \theta_0) | X = x\},
\]

\[
\hat{m}_{g_0}(x) = \frac{\sum_{i=1}^{n} \delta_i W\left(\frac{X_i - \bar{X}}{h}\right) g_0(x, Y_i, \theta_0)}{\sum_{i=1}^{n} \delta_i W\left(\frac{X_i - \bar{X}}{h}\right)}.
\]

Now we have

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i g_0(X_i, Y_i, \theta_0) + (1 - \delta_i) \kappa^{-1} \sum_{\nu=1}^{\kappa} g_0(X_i, Y_{i\nu}, \theta_0) \right\}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \delta_i \{g_0(X_i, Y_i, \theta_0) - m_{g_0}(X_i)\}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \left\{ \kappa^{-1} \sum_{\nu=1}^{\kappa} g_0(X_i, Y_{i\nu}, \theta_0) - \hat{m}_{g_0}(X_i) \right\}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{ \hat{m}_{g_0}(X_i) - m_{g_0}(X_i) \} + \frac{1}{n} \sum_{i=1}^{n} m_{g_0}(X_i)
\]

\[
:= S_n + A_n + T_n + R_n.
\]

Note that \(S_n\) and \(R_n\) are sums of i.i.d. random variables. Define \(\eta(x) = p(x)f(x)\) and \(\hat{\eta}(x) = \frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - x)\) as its kernel estimator, where \(W_h(s) = h^{-d} W(s/h)\). Then,

\[
T_n = \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \frac{\frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - X_i) \{g_0(X_i, Y_j, \theta_0) - m_{g_0}(X_j)\} \eta(X_i)}{\eta(X_i)}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{ \hat{m}_{g_0}(X_i) - m_{g_0}(X_i) \} \frac{\eta(X_i) - \hat{\eta}(X_i)}{\eta(X_i)}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \left\{ \frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - X_i) \{m_{g_0}(X_j) - m_{g_0}(X_i)\} \right\} \frac{1}{\eta(X_i)}
\]

\[
:= T_{n1} + T_{n2} + T_{n3}.
\]

We now derive the asymptotic distribution for \(T_{n1}\). Define

\[
\tilde{T}_{n1} = \sum_{j=1}^{n} E\{T_{n1} | (X_j, Y_j, \delta_j)\}
\]

\[
= \sum_{j=1}^{n} \delta_j E\{T_{n1} | (X_j, Y_j, \delta_j = 1)\}
\]
as the projection of $T_{n1}$ on the original observations. Then write $T_{n1} = \hat{T}_{n1} + (T_{n1} - \hat{T}_{n1})$.

First we will derive the asymptotic distribution of $\hat{T}_{n1}$, then we will show that $T_{n1} - \hat{T}_{n1} = o_p(n^{-1/2})$. For $\hat{T}_{n1}$, note that

$$T_{n1} = \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - X_i)(g_u(X_i, Y_j, \theta_0) - m_{gu}(X_j)) \eta(X_i)$$

$$= \frac{1}{n} \sum_{j=1}^{n} \delta_j (g_u(X_i, Y_j, \theta) - m_{gu}(X_j)) \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \frac{W_h(X_i - X_j)}{\eta(X_i)} \right\}.$$ 

Thus

$$\hat{T}_{n1} = \frac{1}{n} \sum_{j=1}^{n} \delta_j E \left\{ (g_u(X_i, Y_j, \theta_0) - m_{gu}(X_j)) \frac{(1 - \delta_i) W_h(X_i - X_j)}{\eta(X_i)} \right\} | X_j, Y_j$$

$$= \frac{1}{n} \sum_{j=1}^{n} \delta_j \int \left\{ (g_u(x, Y_j, \theta_0) - m_{gu}(X_j)) \frac{(1 - p(x)) W_h(x - X_j)}{\eta(x)} \right\} f(x) dx$$

$$= \frac{1}{n} \sum_{j=1}^{n} \delta_j \int \left\{ (g_u(x, Y_j, \theta_0) - m_{gu}(X_j)) \frac{(1 - p(x)) W_h(x - X_j)}{p(x)} \right\} dx$$

$$= \frac{1}{n} \sum_{j=1}^{n} \delta_j \int \left\{ (g_u(X_i + hs, Y_j, \theta) - m_{gu}(X_j)) \frac{(1 - p(X_i + hs)) W(s)}{p(X_i + hs)} \right\} ds.$$ 

Since both $g_u$ and $p(x) = (1 - p(x))/p(x)$ has bounded seconded derivative on $x$, and

$\sqrt{n}h^2 \rightarrow 0$ as $n \rightarrow \infty$, we carry out Taylor expansion around $X_j$ to conclude

$$\hat{T}_{n1} = \frac{1}{n} \sum_{j=1}^{n} \delta_j (g_u(X_j, Y_j, \theta) - m_{gu}(X_j)) \frac{1 - p(X_j)}{p(X_j)} + o_p(n^{-1/2}). \quad (2.8)$$

Now we show $T_{n1} - \hat{T}_{n1} = o_p(n^{-1/2})$. Let

$$T_{n11} = (1 - \delta_i) \frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - X_i) \{g_u(X_i, Y_j, \theta_0) - m_{gu}(X_j)\} \eta(X_i)$$

$$\hat{T}_{n11} = \sum_{j=1}^{n} E(T_{n11} | (X_j, Y_j, \delta_j = 1)).$$

Then by straight forward computation,

$$E(T_{n1} - \hat{T}_{n1})^2 = \frac{1}{n^2} \sum_{i=1}^{n} E(T_{n1i} - \hat{T}_{n1i})^2 + \frac{2}{n^2} \sum_{i \neq j} E((T_{n1i} - \hat{T}_{n1i})(T_{n1j} - \hat{T}_{n1j}))$$

$$= \frac{1}{n} E(T_{n1i} - \hat{T}_{n1i})^2.$$
The last step used the fact that \( E_{i \neq j} \{(T_{ni} - \bar{T}_{ni})(T_{nj} - \bar{T}_{nj})\} = 0 \), which can be shown by some straightforward if lengthy computation.

Thus

\[
E(T_{ni} - \bar{T}_{ni})^2 = E(T_{ni} - \bar{T}_{ni})^2 = ET_{ni}^2 - ET_{ni}^2 \leq ET_{ni}^2
\]

\[
\leq E \left\{ \frac{1}{n} \sum_{j=1}^{n} \delta_j W_h(X_j - X_i) \{g_u(X_i, Y_j, \theta_0) - m_{gu}(X_j)\} \right\}^2
\]

\[
\rightarrow 0,
\]

by similar argument for proving the consistency of the Nadaraya-Watson estimators (e.g. Stone 1977; Devroye and Wagner 1980). This suggests that \( T_{n1} = \bar{T}_{n1} + o_p(n^{-1/2}) \).

By standard argument, we can show that \( T_{n1} = o_p(n^{-1/2}) \). For \( T_{n3} \), we can use similar argument for \( T_{n1} \) to show

\[
nE \left[ T_{n3} - \sum_{j=1}^{n} \delta_j E\{T_{n3} \mid (X_j, Y_j, \delta_j = 1)\} \right]^2 = o(1).
\]

Also,

\[
\sum_{j=1}^{n} \delta_j E\{T_{n3} \mid (X_j, Y_j, \delta_j = 1)\}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \delta_i E \left\{ \frac{(1 - \delta_i)W_h(X_j - X_i)(m_{gu}(X_j) - m_{gu}(X_i))}{\eta(X_i)} \right\} \quad \hfill (2.8)
\]

\[
= o_p(n^{-1/2}),
\]

by similar augment for (2.8).

Thus, we have

\[
\sqrt{n} T_n \overset{d}{\rightarrow} N \left[ 0, E \{(1 - p(X))^2 \sigma_{gu}^2(X)/p(X)\} \right],
\]

where \( \sigma_{gu}^2(X) = Var\{g_u(X, Y, \theta) \mid X\} \).

Also note \( \sqrt{n} S_n \overset{d}{\rightarrow} N\{0, E(p(X)\sigma_{gu}^2(X))\} \) and \( \sqrt{n} R_n \overset{d}{\rightarrow} N\{0, Var(m_{gu}(X))\} \) by the Central Limit Theorem. Further, it is straightforward to show that \( nCov(S_n, T_n) = \).
\[ E\{(1 - p(X))\sigma^2_{g_n}(X)\} + o(1), \quad n\text{Cov}(R_n, S_n) = 0 \quad \text{and} \quad n\text{Cov}(R_n, T_n) = o(1). \]

It readily follows that

\[ \sqrt{n} \begin{pmatrix} S_n \\ T_n \\ R_n \end{pmatrix} \xrightarrow{d} N \left( 0, \begin{bmatrix} 1 & 0 \\ 0 & \text{Var}(m_{g_n}(X)) \end{bmatrix} \right), \]

where

\[ \Upsilon = \begin{bmatrix} E\{p(X)\sigma^2_{g_n}(X)\} & E\{(1 - p(X))\sigma^2_{g_n}(X)\} \\ E\{(1 - p(X))\sigma^2_{g_n}(X)\} & E\{(1 - p(X))^2\sigma^2_{g_n}(X)/p(X)\} \end{bmatrix}. \]

Thus we have

\[ \sqrt{n}(S_n + T_n + R_n) \xrightarrow{d} N \left[ 0, E\{\sigma^2_{g_n}(X)/p(X)\} + \text{Var}(m_{g_n}(X)) \right]. \quad (2.10) \]

Now we consider the asymptotic distribution of

\[ A_n = \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \left\{ \kappa^{-1} \sum_{\nu=1}^{n} g_n(X_i, Y_{i\nu}, \theta_0) - \hat{m}_{g_n}(X_i) \right\}. \]

Given all the original observations,

\[ n^{-1/2}(1 - \delta_i) \left\{ \kappa^{-1} \sum_{\nu=1}^{n} g_n(X_i, Y_{i\nu}, \theta) - \hat{m}(X_i) \right\}, \quad i = 1, 2, \ldots, n, \]

are independent with conditional mean zero and conditional variance

\[ (n\kappa)^{-1}(1 - \delta_i)\left\{ \hat{\gamma}_{g_n}(X_i) - \hat{m}^2_{g_n}(X_i) \right\}, \]

where \( \hat{\gamma}_{g_n}(x) = \sum_{j=1}^{n} \delta_j W_n(x - X_j) \sigma^2_{g_n}(x, Y_j, \theta_0)/\hat{\eta}(x) \), which is a kernel estimator of \( \gamma_{g_n}(x) = E\{g^2_n(X, Y, \theta_0)|X = x\} \). By verifying Lyapunov’s condition, we can show that conditioning on original observations,

\[ \sqrt{n}A_n \xrightarrow{d} N \left[ 0, (n\kappa)^{-1} \sum_{i=1}^{n} (1 - \delta_i)\left\{ \hat{\gamma}_{g_n}(X_i) - \hat{m}^2_{g_n}(X_i) \right\} \right]. \quad (2.11) \]

Note the variance

\[ (n\kappa)^{-1} \sum_{i=1}^{n} (1 - \delta_i)\left\{ \hat{\gamma}_{g_n}(X_i) - \hat{m}^2_{g_n}(X_i) \right\} \xrightarrow{p} \kappa^{-1} E\{(1 - p(X))\sigma^2_{g_n}(X)\}. \quad (2.12) \]
In Proposition 1, set $V_i$'s to be the original observations, $U_n = \sqrt{n}(S_n + T_n + R_n + A_n)$ and $h(V_1, \ldots, V_n)$ to be $\sqrt{n}(S_n + T_n + R_n)$, then by (2.10), $G$ is the normal distribution function with mean 0 and variance $E\{\sigma^2_{g_u}(X)/p(X)\} + \text{Var}\{m_{g_u}(X)\}$. By (2.11), $F$ is the normal distribution function with mean 0 and variance $(n\kappa)^{-1} \sum_{i=1}^{n}(1-\delta_i)\{\hat{g}_{g_u}(X_i) - \hat{m}^2_{g_u}(X_i)\}$. These together with (2.12), imply that $(G \ast F)(t)$ is normal with mean 0 and variance

$$V \text{ar}\{m_{g_u}(Z, \theta)\} + E\{p^{-1}(X)\sigma^2_{g_u}(X)\} = u^T u,$$ i.e.,

$$n^{-1/2} \sum_{i=1}^{n} \hat{g}_{u}(\hat{Z}_i, \theta) \xrightarrow{L} N(0, u^T u), \text{ as } n \to \infty, \kappa \to \infty.$$

Lemma 1 is then proved by using the Cramèr-Wold device.

Lemma 2. Under the conditions C1-C4, as $n \to \infty$ and $\kappa \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \hat{g}(\hat{Z}_i, \theta_0) \hat{g}^*(\hat{Z}_i, \theta_0) \xrightarrow{p} \Gamma,$$

where $\Gamma = E\{p(X)\text{Cov}(g(X)) + E(g(X)|X)E(g^*(X)|X)\}$.

Proof: Consider each element of the matrix $\frac{1}{n} \sum_{i=1}^{n} \hat{g}(\hat{Z}_i, \theta_0) \hat{g}^*(\hat{Z}_i, \theta_0)$, that is,

$$\frac{1}{n} \sum_{i=1}^{n} \hat{g}_j(\hat{Z}_i, \theta_0) \hat{g}_k(\hat{Z}_i, \theta_0), \quad 1 \leq j, k \leq r.$$

Write

$$\frac{1}{n} \sum_{i=1}^{n} \hat{g}_j(\hat{Z}_i, \theta_0) \hat{g}_k(\hat{Z}_i, \theta_0)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \delta_j \hat{g}_j(Z_i, \theta_0) g_k(Z_i, \theta_0)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} (1-\delta_i) \left\{ \kappa^{-1} \sum_{\nu=1}^{\kappa} \hat{g}_j(X_i, \tilde{Y}_\nu, \theta_0) \right\} \left\{ \kappa^{-1} \sum_{\nu=1}^{\kappa} g_k(X_i, \tilde{Y}_\nu, \theta_0) \right\}$$

$$:= T_{n1} + T_{n2}.$$
Note that

\[ T_{n1} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \{g_j(Z_i, \theta_0) - m_{g_j}(X_i)\} \{g_k(Z_i, \theta_0) - m_{g_k}(X_i)\} \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \delta_i m_{g_j}(X_i) m_{g_k}(X_i) + \frac{1}{n} \sum_{i=1}^{n} \delta_i g_j(Z_i, \theta_0) m_{g_k}(X_i) \]

\[ + \frac{1}{n} \sum_{i=1}^{n} \delta_i g_k(Z_i, \theta_0) m_{g_j}(X_i) \]

\[ := T_{n1a} + T_{n1b} + T_{n1c} + T_{n1d}. \]

It is obvious that \( T_{n1a}, T_{n1b}, T_{n1c} \) and \( T_{n1d} \) are all sums of i.i.d. random variables. By law of large numbers and the continuous mapping theorem, we can show that

\[ T_{n1} \xrightarrow{p} E \left[ p(X) \text{Cov} \{g_j(Z, \theta_0), g_k(Z, \theta_0) | X \} + p(X) m_{g_j}(X) m_{g_k}(X) \right]. \]

Also write

\[ T_{n2} = \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{\tilde{g}_j(\tilde{Z}_i, \theta_0) \tilde{g}_k(\tilde{Z}_i, \theta_0) - \tilde{m}_{g_j}(X_i) \tilde{m}_{g_k}(X_i)\} \]

\[ + \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{\tilde{m}_{g_j}(X_i) \tilde{m}_{g_k}(X_i) - m_{g_j}(X_i) m_{g_k}(X_i)\} \]

\[ + \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) m_{g_j}(X_i) m_{g_k}(X_i) \]

\[ := T_{n2a} + T_{n2b} + T_{n2c}. \]

Note that \( g_j(X_i, Y_{iv}, \theta_0) \) has conditional mean \( \tilde{m}_{g_j}(X_i) \) given the original observations \( X_n \). It can be shown that \( T_{n2a} \xrightarrow{p} 0 \) as \( \kappa \to \infty \). By argument similar to those used for (2.9), \( T_{n2b} \xrightarrow{p} 0 \) as \( n \to \infty \). Obviously \( T_{n2c} \) is the sum of i.i.d. random variables, which leads to \( T_{n2c} \xrightarrow{p} E\{(1 - p(X)) m_{g_j}(X_i) m_{g_k}(X_i)\} \). Hence we have

\[ T_{n2} \xrightarrow{p} E\{(1 - p(X)) m_{g_j}(X_i) m_{g_k}(X_i)\} \]

as \( n \to \infty, \kappa \to \infty \). Therefore,

\[ T_{n1} + T_{n2} \xrightarrow{p} E \left[ p(X) \text{Cov} \{g_j(Z, \theta_0), g_k(Z, \theta_0) | X \} + m_{g_j}(X) m_{g_k}(X) \right]. \]

This completes the proof of Lemma 2. \( \square \)
Lemma 3. Under the conditions C1-C5 and $\kappa \to \infty$,

$$\max_{1 \leq i \leq n} \|\tilde{g}(\tilde{Z}_i, \theta_0)\| = o_p(n^{1/3}).$$

Proof: First note $\max_{1 \leq i \leq n} \|g(Z_i, \theta_0)\| = o_p(n^{1/3})$ when $E\|g(Z_i, \theta_0)\|^3 < \infty$. Since

$$\max_{1 \leq i \leq n} \|\tilde{g}(\tilde{Z}_i, \theta_0)\| \leq \max_{1 \leq i \leq n} \|g(Z_i, \theta_0)\| + \max_{1 \leq i \leq n} \kappa^{-1} \sum_{\nu=1}^{\kappa} g(X_i, \bar{Y}_\nu, \theta_0),$$

we only need to show

$$\max_{1 \leq i \leq n} \kappa^{-1} \sum_{\nu=1}^{\kappa} g(X_i, \bar{Y}_\nu, \theta_0) = o_p(n^{1/3}).$$

Note that

$$\max_{1 \leq i \leq n} \kappa^{-1} \sum_{\nu=1}^{\kappa} g(X_i, \bar{Y}_\nu, \theta_0) \leq \max_{1 \leq i \leq n} \kappa^{-1} \sum_{\nu=1}^{\kappa} g(X_i, \bar{Y}_\nu, \theta_0) - \tilde{m}_g(X_i) + \max_{1 \leq i \leq n} \|\tilde{m}_g(X_i) - m_g(X_i)\| + \max_{1 \leq i \leq n} \|m_g(X_i)\| \leq T_{n1} + T_{n2} + T_{n3}.$$

Using similar argument to Lemma 1 and 2, as well as Lemma 3 of Owen (1990), we can show $T_{n1} = o_p(n^{1/3})$ and $T_{n2} = o_p(n^{1/3})$. Also note that $E\|m_g(X_i)\|^3 \leq E\|g(Z_i, \theta_0)\|^3 < \infty$ implies $T_{n3} = o_p(n^{1/3})$. These together prove the lemma. \(\square\)

We need a result from Aitchison and Silvery (1958), which is stated as the following.

Proposition 2. If $\Psi(\lambda)$ is a continuous function mapping $\mathbb{R}^k$ into itself with the property that for every $\lambda$ such that $\|\lambda\| = 1$, $\lambda^T \Psi(\lambda) < 0$, then there exists a point $\hat{\lambda}$ such that $\|\hat{\lambda}\| < 1$ and $\Psi(\hat{\lambda}) = 0$, where $\|\cdot\|$ denotes the Euclidean distance.

Now we define the following two quantities,

$$Q_{1n}(\theta, t) = \frac{1}{n} \sum_{i} \frac{1}{1 + t^T \tilde{g}(\tilde{Z}_i, \theta)} \tilde{g}(\tilde{Z}_i, \theta),$$

$$Q_{2n}(\theta, t) = \frac{1}{n} \sum_{i} \frac{1}{1 + t^T \tilde{g}(\tilde{Z}_i, \theta)} \left( \frac{\partial \tilde{g}(\tilde{Z}_i, \theta)}{\partial \theta} \right)^T t.$$
The next Lemma shows that the equation $Q_{1n}(\theta, t) = 0$ can uniquely determine $t = t(\theta)$ in the neighborhood of $\theta_0$, recall that $t$ is the Lagrange multiplier described in (2.5).

**Lemma 4.** Assume that conditions C1-C6 are satisfied, $n \to \infty$ and $\kappa \to \infty$, then with probability tending to 1, in the sphere $\{ \theta : ||\theta - \theta_0|| \leq d_n \}$, the equation $Q_{1n}(\theta, t) = 0$ has roots $t = t(\theta) = O(d_n)$, and $t(\theta)$ is continuous and differentiable when $\theta$ belongs to this sphere, where $d_n = n^{-\frac{1}{3} - \epsilon}, \frac{1}{6} > \epsilon > 0$.

**Proof:** For fixed $\theta$ such that $||\theta - \theta_0|| \leq d_n$, let

$$\Psi(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{g}(\tilde{Z}_i, \theta)}{1 + n^{-\frac{1}{3} - \epsilon/2} \lambda^r \tilde{g}(\tilde{Z}_i, \theta)}.$$

Already shown that when $E||g(Z, \theta)||^3 < \infty$, $\max_{1 \leq i \leq n} ||\tilde{g}(\tilde{Z}_i, \theta)|| = o_p(n^{1/3})$, also $\Psi(\lambda)$ is a continuous function for $||\lambda|| \leq 1$. When $||\lambda|| = 1$, we have, with probability tending to 1,

$$\lambda^r \Psi(\lambda) = \lambda^r \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\tilde{Z}_i, \theta) - n^{-\frac{1}{3} - \epsilon/2} \lambda^r \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\tilde{Z}_i, \theta) \tilde{g}^r(\tilde{Z}_i, \theta) + O(n^{-\frac{3}{2} - \epsilon})$$

$$\leq O(n^{-\frac{3}{2} - \epsilon}) - cn^{-\frac{1}{3} - \epsilon/2} + O(n^{-\frac{3}{2} - \epsilon})$$

$$< 0,$$

where $c$ is the smallest eigenvalue of $\tilde{G}$. By Proposition 2, with probability tending to 1, there exists a point $\hat{\lambda}$ such that $||\hat{\lambda}|| < 1$ and $\Psi(\hat{\lambda}) = 0$. Then the result in the lemma can be derived using the implicit-function theorem. \(\square\)

The next lemma shows the existence of solutions to $Q_{1n}(\theta, t) = 0$ and $Q_{2n}(\theta, t) = 0$. As in Lemma 4, we define $d_n = n^{-\frac{1}{3} - \epsilon}, \frac{1}{6} > \epsilon > 0$.

**Lemma 5.** Under the conditions C1-C7, as $n \to \infty$ and $\kappa \to \infty$, with probability tending to 1, $\ell(\theta) = \sum_{i=1}^{n} \log(1 + t(\theta)\tilde{g}(\tilde{Z}_i, \theta))$ attains its minimum value at some point $\hat{\theta}$ in the interior of the ball $||\theta - \theta_0|| \leq d_n$, and $\hat{\theta}$ and $\hat{t} = t(\hat{\theta})$ satisfy

$$Q_{1n}(\hat{\theta}, \hat{t}) = 0, \quad Q_{2n}(\hat{\theta}, \hat{t}) = 0.$$
Proof: Denote \( \theta = \theta_0 + ud_n \), for \( \theta \in \{ \theta \mid ||\theta - \theta_0|| = d_n \} \), where ||u|| = 1. First, we give a lower bound for \( \ell(\theta) \) on the surface of the ball. With \( Q_{1n} = 0 \) and Lemma 4, we have

\[
\ell(\theta) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\tilde{Z}_i, \theta) \tilde{g}^*(\tilde{Z}_i, \theta) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\tilde{Z}_i, \theta) \right\} + o_p(d_n)
\]

about \( \theta \in \{ \theta \mid ||\theta - \theta_0|| \leq d_n \} \).

By this and Taylor expansion, we have,

\[
\ell(\theta) = \sum_i \ell^*(\theta) \tilde{g}(\tilde{Z}_i, \theta) - \frac{1}{2} \sum_i \left\{ \ell^*(\theta) \tilde{g}(\tilde{Z}_i, \theta) \right\}^2 + o_p(n^{1/2})
\]

\[
= \frac{n}{2} \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta) \right\}^r \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta) \tilde{g}^*(\tilde{Z}_i, \theta) \right\}^{-1} \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta) \right\}
\]

+ \( o_p(n^{1/2}) \)

\[
= \frac{n}{2} \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta_0) + \frac{1}{n} \sum_i \frac{\partial \tilde{g}(\tilde{Z}_i, \theta_0)}{\partial \theta} ud_n \right\}^r \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta) \tilde{g}^*(\tilde{Z}_i, \theta) \right\}^{-1}
\]

\[
\times \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta_0) + \frac{1}{n} \sum_i \frac{\partial \tilde{g}(\tilde{Z}_i, \theta_0)}{\partial \theta} ud_n \right\} + o_p(n^{1/2})
\]

\[
= \frac{n}{2} \left\{ O_p(n^{-1/2}) + E \left( \frac{\partial \tilde{g}(\tilde{Z}_i, \theta_0)}{\partial \theta} \right) ud_n \right\}^r \tilde{\Gamma}^{-1}
\]

\[
\times \left\{ O_p(n^{-1/2}) + E \left( \frac{\partial \tilde{g}(\tilde{Z}_i, \theta_0)}{\partial \theta} \right) ud_n \right\} + o_p(n^{1/2})
\]

\[= O_p(n^{1/2}) \]

Similarly, we have

\[
\ell(\theta_0) = \frac{n}{2} \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta_0) \right\}^r \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta_0) \tilde{g}^*(\tilde{Z}_i, \theta_0) \right\}^{-1}
\]

\[
\times \left\{ \frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta_0) \right\} + o_p(1)
\]

\[= O_p(1) \]

Noting that \( \ell(\theta) \) is a continuous function about \( \theta \) as \( \theta \) belongs to the ball \( ||\theta - \theta_0|| \leq d_n \), with probability tending to 1, \( \ell(\theta) \) has minimum value in the interior of the ball, and \( \hat{\theta} \)
satisfies
\[
\left. \frac{\partial \ell(\theta)}{\partial \theta} \right|_{\theta = \hat{\theta}} = \sum_i \left\{ \frac{\partial t^* (\theta)}{\partial \theta} + \left\{ \frac{\partial \tilde{g}(\tilde{Z}_i, \theta)}{\partial \theta} \right\}^\tau t(\theta) \right\} \left. \right|_{\theta = \hat{\theta}} \\
= \frac{1}{1 + t^* (\theta) \tilde{g}(\tilde{Z}_i, \theta)} \left\{ \frac{\partial \tilde{g}(\tilde{Z}_i, \theta)}{\partial \theta} \right\}^\tau \left. t(\theta) \right|_{\theta = \hat{\theta}} \\
= 0
\]

Proof of Theorem 1: Taking the derivatives about $\theta$ and $t^*$,
\[
\frac{\partial Q_{1n}(\theta, 0)}{\partial \theta} = \frac{1}{n} \sum_i \frac{\partial \tilde{g}(\tilde{Z}_i, \theta)}{\partial \theta}, \quad \frac{\partial Q_{1n}(\theta, 0)}{\partial t^*} = -\frac{1}{n} \sum_i \tilde{g}(\tilde{Z}_i, \theta) \tilde{g}^*(\tilde{Z}_i, \theta),
\]
\[
\frac{\partial Q_{2n}(\theta, 0)}{\partial \theta} = 0, \quad \frac{\partial Q_{2n}(\theta, 0)}{\partial t^*} = \frac{1}{n} \sum_i \left\{ \frac{\partial \tilde{g}(\tilde{Z}_i, \theta)}{\partial \theta} \right\}^\tau.
\]

Expanding $Q_{1n}(\hat{\theta}, \hat{t})$, $Q_{2n}(\hat{\theta}, \hat{t})$ at $(\theta_0, 0)$, by the conditions of the theorem,
\[
0 = Q_{1n}(\hat{\theta}, \hat{t}) = Q_{1n}(\theta_0, 0) + \frac{\partial Q_{1n}(\theta_0, 0)}{\partial \theta} (\hat{\theta} - \theta_0) + \frac{\partial Q_{1n}(\theta_0, 0)}{\partial t^*} (\hat{t} - 0) + o_p(\zeta_n),
\]
\[
0 = Q_{2n}(\hat{\theta}, \hat{t}) = Q_{2n}(\theta_0, 0) + \frac{\partial Q_{2n}(\theta_0, 0)}{\partial \theta} (\hat{\theta} - \theta_0) + \frac{\partial Q_{2n}(\theta_0, 0)}{\partial t^*} (\hat{t} - 0) + o_p(\zeta_n),
\]

where $\zeta_n = ||\hat{\theta} - \theta_0|| + ||\hat{t}||$. Then we can write
\[
\begin{pmatrix}
\hat{t} \\
\hat{\theta} - \theta_0
\end{pmatrix} = S_n^{-1} \begin{pmatrix}
-Q_{1n}(\theta_0, 0) + o_p(\zeta_n) \\
o_p(\zeta_n)
\end{pmatrix},
\]

where
\[
S_n = \begin{pmatrix}
\frac{\partial Q_{1n}}{\partial \theta} & \frac{\partial Q_{1n}}{\partial t^*} \\
\frac{\partial Q_{2n}}{\partial \theta} & 0
\end{pmatrix}_{(\theta_0, 0)} \rightarrow \begin{pmatrix}
S_{11} & S_{12} \\
S_{21} & 0
\end{pmatrix} = \begin{pmatrix}
-\tilde{\Gamma} & E \left( \frac{\partial g}{\partial \theta} \right) \\
E \left( \frac{\partial g}{\partial \theta} \right) & 0
\end{pmatrix}.
\]

Note that $Q_{1n}(\theta_0, 0) = \frac{1}{n} \sum_{i=1}^n \tilde{g}(\tilde{Z}_i, \theta_0) = O_p(n^{-1/2})$, it follows that $\zeta_n = O_p(n^{-1/2})$.

After some matrix manipulation
\[
\sqrt{n} (\hat{\theta} - \theta_0) = S_{22}^{-1} S_{21} S_{11}^{-1} \sqrt{n} Q_{1n}(\theta_0, 0) + o_p(1),
\]
where
\[ V = S_{22}^{-1} = \left\{ E\left( \frac{\partial g}{\partial \theta} \right)^{\top} \Gamma^{-1} E\left( \frac{\partial g}{\partial \theta} \right) \right\}^{-1}. \]

By Lemma 1, \( \sqrt{n}Q_{1n}(\theta_0, 0) \to N(0, \Gamma) \), and the theorem follows.

**Proof of Theorem 2:**

\[ R(\theta_0) = 2 \left[ \sum_i \log \{1 + t_i \tilde{g}(\tilde{Z}_i, \theta_0)\} - \sum_i \log \{1 + t_i \tilde{g}(\tilde{Z}_i, \tilde{\theta})\} \right], \]

where \( t_0 = t(\theta_0) \). Also note that

\[ \ell(\tilde{\theta}, \tilde{\theta}) = \sum_i \log \{1 + t_i \tilde{g}(\tilde{Z}_i, \tilde{\theta})\} = -\frac{n}{2} Q_{1n}^T(\theta_0, 0)AQ_{1n}(\theta_0, 0) + o_p(1), \]

where

\[ A = S_{11}^{-1}(I + S_{12}S_{22}^{-1}S_{21}S_{11}^{-1}). \]

Under \( H_0 \),

\[ \frac{1}{n} \sum_i \frac{1}{1 + t_i \tilde{g}(\tilde{Z}_i, \theta_0)} \tilde{g}(\tilde{Z}_i, \theta_0) = 0, \]

\[ t_0 = -S_{11}^{-1}Q_{1n}(\theta_0, 0)S_{11}^{-1}Q_{1n}(\theta_0, 0) + o_p(1), \]

and

\[ \sum_i \log \{1 + t_i \tilde{g}(\tilde{Z}_i, \theta_0)\} = -\frac{n}{2} Q_{1n}^T(\theta_0, 0)S_{11}^{-1}Q_{1n}(\theta_0, 0) + o_p(1). \]

Thus we have

\[ R(\theta_0) = nQ_{1n}^T(\theta_0, 0)(A - S_{11}^{-1})Q_{1n}(\theta_0, 0) + o_p(1) \]

\[ = \sqrt{n}Q_{1n}^T(\theta_0, 0)S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1}Q_{1n}(\theta_0, 0) + o_p(1). \]

Note that

\[ S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1} \overset{p}{\rightarrow} \Gamma^{-1}E\left( \frac{\partial g}{\partial \theta} \right) \Gamma^{-1}E\left( \frac{\partial g}{\partial \theta} \right)^{\top}, \]

and by Lemma 1, \( \sqrt{n}Q_{1n}(\theta_0, 0) \to N(0, \Gamma) \) in distribution, the theorem then follows.
Proof for Theorem 3: The proof for Theorem 3 essentially involves establishing bootstrap version of Lemma 1 to Theorem 2. We only outline the main steps in proving the bootstrap version of Lemma 1 here.

Let $X^*_i$, $Y^*_i$, $Y^*_t$, $\delta^*_i$ be the counterpart to $X_i$, $Y_t$, $\hat{Y}_t$, $\delta_i$ in the bootstrap sample, $S_n(\hat{\theta})$, $A_n(\hat{\theta})$, $T_n(\hat{\theta})$ and $R_n(\hat{\theta})$ represent the quantities $S_n$, $A_n$, $T_n$ and $R_n$ with $\theta_0$ replaced by $\hat{\theta}$ respectively. Let $S^*_n(\hat{\theta})$, $A^*_n(\hat{\theta})$, $T^*_n(\hat{\theta})$ and $R^*_n(\hat{\theta})$ be their bootstrap counterpart. First we show that

$$\sqrt{n}\{S^*_n(\hat{\theta}) + T^*_n(\hat{\theta}) + R^*_n(\hat{\theta}) - S_n(\hat{\theta}) - T_n(\hat{\theta}) - R_n(\hat{\theta})\} \overset{d}{\to} \mathcal{N}[0, E_*\{\sigma^2_{g_0}(X, \hat{\theta})/p(X)\} + \text{Var}_*\{m_{g_0}(X, \hat{\theta})\}],$$

where $E_*\cdot$ and $\text{Var}_*\cdot$ represent the conditional expectation and variance given the original data respectively.

Define

$$\hat{m}_{g_0}(x, \hat{\theta}) = \frac{\sum_{i=1}^{n} \delta_i W(\frac{x - X_i}{h}) g_0(x, Y_i, \hat{\theta})}{\sum_{i=1}^{n} \delta_i W(\frac{x - X_i}{h})}, \quad \text{and} \quad \hat{m}^*_g(x, \hat{\theta}) = \frac{\sum_{i=1}^{n} \delta^*_i W(\frac{x - X^*_i}{h}) g_0(x, Y^*_i, \hat{\theta})}{\sum_{i=1}^{n} \delta^*_i W(\frac{x - X^*_i}{h})}.$$

Then

$$S^*_n(\hat{\theta}) + T^*_n(\hat{\theta}) + R^*_n(\hat{\theta}) - S_n(\hat{\theta}) - T_n(\hat{\theta}) - R_n(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[ \delta^*_i \{g_0(Z^*_i, \hat{\theta}) - m_{g_0}(X^*_i, \hat{\theta})\} - \frac{1}{n} \sum_{j=1}^{n} \delta_j \{g_0(Z_j, \hat{\theta}) - m_{g_0}(X_j, \hat{\theta})\} \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \left[ (1 - \delta^*_i) \{\hat{m}_{g_0}(X^*_i, \hat{\theta}) - \hat{m}_{g_0}(X^*_i)\} \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \left[ (1 - \delta^*_i) \{\hat{m}_{g_0}(X^*_i, \hat{\theta}) - m_{g_0}(X^*_i, \hat{\theta})\} \right]$$

$$- \frac{1}{n} \sum_{j=1}^{n} \left[ (1 - \delta_j) \{\hat{m}_{g_0}(X_j, \hat{\theta}) - m_{g_0}(X_j, \hat{\theta})\} \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \left\{ m_{g_0}(X^*_i, \hat{\theta}) - \frac{1}{n} \sum_{j=1}^{n} m_{g_0}(X_j, \hat{\theta}) \right\}$$

$$:= B_1 + B_2 + B_3 + B_4.$$
For both \( B_1 \) and \( B_4 \), we can apply the Central Limit Theorem for bootstrap samples (e.g. Shao and Tu, 1985) to derive

\[
\sqrt{n}B_1 \xrightarrow{d} N[0, E_\ast \{p(X)\sigma^2_{g_n}(X, \hat{\theta})\}],
\]

\[
\sqrt{n}B_4 \xrightarrow{d} N[0, Var_\ast \{m_{g_n}(X, \hat{\theta})\}].
\]  

(2.14)

Also it can be shown \( B_2 = o_p(n^{-1/2}) \). Use similar arguments to (2.8) to show

\[
B_3 = \frac{1}{n} \sum_{i=1}^{n} \left[ \delta_i^\ast \left\{ g_u(Z_i^\ast, \hat{\theta}) - m_{g_n}(X_i^\ast, \hat{\theta}) \right\} \frac{1 - p(X_i^\ast)}{p(X_i^\ast)} - \frac{1}{n} \sum_{j=1}^{n} \delta_j \left\{ g_u(Z_j, \hat{\theta}) - m_{g_n}(X_j, \hat{\theta}) \right\} \frac{1 - p(X_j)}{p(X_j)} \right] + o_p(n^{-1/2}).
\]

Then imitate the proof for Lemma 1 and apply the bootstrap central limit theorem to conclude (2.13).

For \( A_n^\ast(\hat{\theta}) \), given observations in the bootstrap sample that are not imputed, we have

\[
\sqrt{n}A_n^\ast(\hat{\theta}) \xrightarrow{d} N \left[ 0, (n\kappa)^{-1} \sum_{i=1}^{n} (1 - \delta_i^\ast) \{ \gamma^\ast(X_i^\ast, \hat{\theta}) - \tilde{m}_n^\ast(X_i^\ast, \hat{\theta}) \} \right].
\]

Similar to the proof of Lemma 1, by employing Proposition 1

\[
\frac{1}{\sqrt{n}} \left\{ \sum_{i=1}^{n} g_u(\tilde{Z}_i^\ast, \hat{\theta}) - n^{-1} \sum_{j=1}^{n} g_u(\tilde{Z}_j, \hat{\theta}) \right\}
\]

\[
\xrightarrow{d} N [0, E_\ast \{\sigma^2_{g_n}(X, \hat{\theta})/p(X)\} + Var_\ast \{m_{g_n}(X, \hat{\theta})\}].
\]

The bootstrap version of Lemma 1 is justified by noting

\[
E_\ast \{\sigma^2_{g_n}(X, \hat{\theta})/p(X)\} \rightarrow E \{\sigma^2_{g_n}(X)/p(X)\}, \text{ and}
\]

\[
Var_\ast \{m_{g_n}(X, \hat{\theta})\} \rightarrow Var \{m_{g_n}(X)\},
\]

as \( n \rightarrow \infty \), then employ the Cramèr-Wold device.
CHAPTER 3. Identifying Genes Associated with a Quantitative Trait or Quantitative Trait Locus via Selective Transcriptional Profiling

A paper published in *Biometrics*¹

Dong Wang and Dan Nettleton

**Abstract:** Genetical genomics is an approach that blends the mapping of quantitative trait loci (QTL) with microarray analysis. The approach can be used to identify associations between the allelic state of a genomic region and a gene’s transcript abundance. However, the large number of microarrays required for adequate power results in high material and labor cost that prevent wide adoption of the genetical genomics strategy outside of some well funded laboratories. We present a method called *selective transcriptional profiling* that involves selecting an optimal subset of individuals to microarray from a larger set of individuals for which relatively inexpensive quantitative trait and molecular marker data are available. We show how to use microarray data from the selected individuals, along with the trait and marker data from all individuals, to identify genes whose transcript abundance is associated with a quantitative trait of interest through linkage to a trait QTL or correlation with the trait. Our methods for selection and analysis are derived within a missing data framework.

**KEY WORDS:** Genetical genomics; Microarray; Missing data; Quantitative trait locus.

3.1 Introduction

Quantitative trait locus (QTL) mapping has become a widely used technique for identifying genomic regions associated with quantitative traits in plants and animals. Microarray technology has become a major tool in functional genomics research. Jansen and Nap (2001) proposed genetical genomics as an approach that blends QTL mapping and microarray analysis to identify associations between the allelic state of a genomic region and a gene's transcript abundance. Genetical genomics has been used recently to study the genetics of gene expression in yeast (Brem et al., 2002; Yvert et al., 2003), rat (Hubner et al., 2005), mouse (Schadt, Monks, and Drake 2003; Bystrykh et al. 2005; Chesler et al. 2005), maize, and human (Schadt, Monks, and Drake, 2003). Such studies are quite costly because gene expression is measured separately for each of many individuals using microarray slides (Schena et al., 1995) or Affymetrix GeneChips (Lipshutz et al., 1999) that have a high per unit cost in human effort and materials. Although genetical genomics is a powerful approach for discovering how genes work together to carry out essential biological processes, the large number of microarrays required for adequate power can make the approach unaffordable for all but a few well funded laboratories.

In this paper we propose a new method called selective transcriptional profiling that will allow researchers to identify associations between the transcriptional expression of genes and genetic loci using far fewer microarrays than required in the classical genetical genomics approach. Our method is ideal for a researcher interested in uncovering the genetic architecture of a particular quantitative trait (e.g., body fat in mice or yield in corn). Given that a QTL for the trait of interest has been identified using a full set of trait data and molecular marker data on $N$ individuals, we describe an optimal strategy for selecting a subset of the $N$ individuals whose transcriptome should be measured using microarrays. Furthermore, we show how to use this partial microarray data in conjunction with the full set of relatively inexpensive trait and molecular marker data.
to (1) identify genes whose transcript abundance is associated with a QTL for the trait of interest and (2) identify genes whose transcript abundance is conditionally correlated with the trait of interest, given the estimated QTL genotypes. We expect the expression of genes that play a role in establishing the trait values to be correlated with the trait values. Our method exploits these correlations to extract information about gene expression from all $N$ individuals even though expression is explicitly measured in only a subset of the $N$ individuals. Conventional wisdom among practitioners suggests transcript profiling for individuals with highest and lowest trait values. We show, however, that this intuitive approach for selective profiling will typically be sub-optimal.

Our approach is directly relevant to uncovering the genes involved in establishing trait values. The discovery of a QTL for the trait of interest indicates that different QTL alleles are associated with different trait values. It will often be the case the different QTL alleles lead to differences in transcriptional abundance of other genes. Such QTL are known as trans-acting modulators of gene expression, and initial evidence suggests that many loci may act in this fashion (Jansen and Nap, 2004). Genes whose expression is modulated by a trans-acting QTL are of great interest because these genes are likely to play a role in explaining trait variation. Our approach is aimed directly at uncovering such genes, and thus may be used to elucidate major components of the genetic pathways that lead to variation in the trait of interest. Other genes that play an important role in determining the trait of interest may not be regulated by the identified QTL, but we show how these genes can be discovered by testing for conditional correlation between the trait and gene expression measurements, given the estimated the QTL genotype.

Jin et al. (2004) recently proposed selective phenotyping for improving the efficiency of genetical genomics studies and QTL studies in general. Although similar in purpose to our selective transcriptional profiling approach, there are major differences between the methods. The basic idea of the selective phenotyping strategy proposed by Jin et al. (2004) is to choose a subsample of individuals that are as dissimilar as possible
with respect to marker genotypes across the genome or over genomic regions of interest. The traits of interest – which could include measures of mRNA transcript abundance – are then measured for the selected individuals and only data from selected individuals are used for analysis. In our selective transcriptional profiling approach, we use both traditional trait and marker data to select individuals for transcriptional profiling and then use available data from both selected and unselected individuals to perform our analysis.

3.2 Model and Notation

Our method can be used to test whether a QTL with two different genotypes is associated with the expression level of any given gene. The two-genotype case is relevant for many common population structures including backcross, doubled haploid, or recombinant inbred lines. Suppose a QTL known to be associated with a trait of interest has alleles \( x \) and \( y \). Suppose we have \( N = N_x + N_y \) individuals, among which \( N_x \) are of genotype \( x \) and \( N_y \) are of genotype \( y \). Because of financial constraints, only \( n_x \) individuals of genotype \( x \) and \( n_y \) individuals of genotype \( y \) will be chosen for microarray analysis to measure the expression level of each of thousands of genes. For ease of exposition, we will describe our approach for analyzing data on the expression of only one gene. In practice the same analysis strategy will be implemented separately for each of thousands of genes.

Let \( X_{i1} \) and \( X_{i2} \) denote the quantitative trait and gene expression measure for the \( i \)th individual of genotype \( x \). We assume that \((X_{i1}, X_{i2})'\) has (perhaps after suitable transformation) a bivariate normal distribution with mean \((\mu_{x1}, \mu_{x2})'\) and covariance matrix

\[
\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}
\]
There are $n_x$ complete pairs of data for expression and trait,

$$(X_{11}, X_{12}), (X_{21}, X_{22}), \cdots, (X_{n_x1}, X_{n_x2})$$

and $N_x - n_x$ observations of the trait only,

$$X_{(n_x+1)1}, X_{(n_x+2)1}, \cdots, X_{N_x1}.$$

Similarly, $Y_{i1}$ and $Y_{i2}$ are trait and expression measures of the $i$th individual of genotype $y$. We assume $(Y_{i1}, Y_{i2})'$ are bivariate normal with mean $(\mu_{y1}, \mu_{y2})'$ and the covariance matrix $\Sigma$ that is the same for both genotypes $x$ and $y$. There are $n_y$ pairs of $(Y_{i1}, Y_{i2})$ and $N_y - n_y$ observations of $Y_{i1}$ only. All data are assumed to be independent across individuals. Using this framework, our problems of interest are

(1) to test $H_0: \mu_{x2} = \mu_{y2}$ vs. $H_A: \mu_{x2} \neq \mu_{y2}$; i.e., to test for association between QTL genotype and gene expression, and

(2) to test $H_0: \rho \equiv \frac{\sigma_{x2}}{\sqrt{\sigma_{x1}^2 \sigma_{y2}^2}} = 0$ vs. $H_A: \rho \neq 0$; i.e., to test for correlation between the trait and gene expression conditional on QTL genotype.

If $n_x = N_x$ and $n_y = N_y$, the information of both trait and expression are available for all individuals, and the problems are easily handled by standard likelihood based methods. We are interested in cases where $n_x < N_x$ and $n_y < N_y$, which can be considered as a missing data problem. If the individuals used for the microarray experiment are selected completely at random regardless of the trait value, $X_{i2}$ and $Y_{i2}$ are missing completely at random (MCAR) using the nomenclature of Little and Rubin (2002). If we select individuals to measure expression using certain criteria concerning only $X_{i1}$ or $Y_{i1}$, then $X_{i2}$ and $Y_{i2}$ are missing at random (MAR), but not observed at random (OAR).
3.3 Maximum Likelihood Estimators and Tests

3.3.1 Derivation of maximum likelihood estimators and their asymptotic covariance matrix

The maximum likelihood estimators for the parameters of a population with bivariate normal distribution when some of the observations of one variable are missing under the MCAR or MAR conditions has been described by Little and Rubin (2002), Mendoza (1993) and others. For the two population case discussed here, the MLE can be derived as follows. Recall that there are \( n_x \) complete observations and \( N_x - n_x \) observations for the trait only for individuals of genotype \( x \), and for individuals of genotype \( y \), there are \( n_y \) complete observations and \( N_y - n_y \) of observations of trait only. Based on the factorization method described by Little and Rubin (2002), the likelihood function can be written as

\[
L = \left\{ \prod_{i=1}^{N_x} f(X_{i1}|\mu_{x1}, \sigma_{11}) \prod_{j=1}^{n_x} f(X_{j2}|\beta_{20x} + \beta_{21}X_{j1}, \sigma_{22.1}) \right\} \left\{ \prod_{i=1}^{N_y} f(Y_{i1}|\mu_{y1}, \sigma_{11}) \prod_{j=1}^{n_y} f(Y_{j2}|\beta_{20y} + \beta_{21}Y_{j1}, \sigma_{22.1}) \right\}.
\] (3.1)

The factors in the first brace pertain to the trait and expression level of individuals of the genotype \( x \). The first factor is the density of an independent sample of size \( N_x \) of a normal distribution with mean \( \mu_{x1} \) and variance \( \sigma_{11} \). The second factor is the density of an independent sample of size \( n_x \) from the conditional normal distributions of \( X_{j2} \) given \( X_{j1} \) with conditional mean \( \beta_{20x} + \beta_{21}X_{j1} \) and variance \( \sigma_{22.1} \) for \( j = 1, \ldots, n_x \). In a similar fashion, factors in the second brace pertain to the trait and expression level of individuals with genotype \( y \). Note that \( \beta_{21} = \sigma_{12}/\sigma_{11} \), \( \beta_{20x} = \mu_{x2} - \beta_{21}\mu_{x1} \), \( \beta_{20y} = \mu_{y2} - \beta_{21}\mu_{y1} \), and \( \sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11} \). Let

\[
\theta = (\mu_{x1}, \mu_{x2}, \mu_{y1}, \mu_{y2}, \sigma_{11}, \sigma_{22}, \sigma_{12})', \quad \text{and}
\]

\[
\phi = (\mu_{x1}, \mu_{y1}, \sigma_{11}, \beta_{20x}, \beta_{20y}, \beta_{21}, \beta_{22.1})'.
\] (3.2)
Then $\phi$ is a one-to-one mapping from $\theta$. The maximum likelihood estimators, derived by maximizing equation (3.1), are as follows:

$$
\hat{\mu}_{x1} = \bar{X}_1, \quad \hat{\mu}_{y1} = \bar{Y}_1,
$$

$$
\hat{\beta}_{21} = \frac{\sum_{i=1}^{n_x} (X_{i1} - n_x^{-1} \sum_{j=1}^{n_x} X_{j1})(X_{i2} - n_x^{-1} \sum_{j=1}^{n_x} X_{j2})}{\sum_{i=1}^{n_x} (X_{i1} - n_x^{-1} \sum_{j=1}^{n_x} X_{j1})^2 + \sum_{i=1}^{n_y} (Y_{i1} - n_y^{-1} \sum_{j=1}^{n_y} Y_{j1})^2 + \sum_{i=1}^{n_x} (Y_{i2} - n_y^{-1} \sum_{j=1}^{n_y} Y_{j2})^2},
$$

$$
\hat{\beta}_{20x} = \frac{\sum_{i=1}^{n_x} (X_{i2} - \hat{\beta}_{21} X_{i1})}{n_x}, \quad \hat{\beta}_{20y} = \frac{\sum_{i=1}^{n_y} (Y_{i2} - \hat{\beta}_{21} Y_{i1})}{n_y},
$$

$$
\hat{\sigma}^{2}_{11} = \frac{\sum_{i=1}^{n_x} (X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n_y} (Y_{i1} - \bar{Y}_1)^2}{N_x + N_y}, \quad \text{and}
$$

$$
\hat{\sigma}^{2}_{21} = \frac{\sum_{i=1}^{n_x} (X_{i2} - \hat{\beta}_{20x} - \hat{\beta}_{21} X_{i1})^2 + \sum_{i=1}^{n_y} (Y_{i2} - \hat{\beta}_{20y} - \hat{\beta}_{21} Y_{i1})^2}{n_x + n_y}.
$$

To obtain the asymptotic variance of the above MLEs, we use a similar method as described by Mendoza (1993) to derive the inverted Fisher information matrix. Details of the derivation are contained in Appendix A. The resulting asymptotic variance of $\phi$ is given by

$$
I^{-1}(\phi) = \begin{bmatrix}
\frac{\sigma_{21}}{N_x} & \frac{\sigma_{21}}{N_y} \\
\frac{\sigma_{21}}{N_y} & \frac{2\sigma_{11}^2}{N_x + N_y} \\
\frac{2\sigma_{11}^2}{N_x + N_y} & I_1^{-1}(\phi)
\end{bmatrix}
$$

(3.4)

where

$$
I_1^{-1}(\phi) = \frac{\sigma_{22,1}}{(n_x + n_y)\sigma_{11}^2} \begin{bmatrix}
(n_x + n_y)\sigma_{11}' \mu_x^2 + \mu_{x1}^2 & \mu_x^2 \mu_{y1} + \mu_x^2 \mu_y' - \mu_{x1}' \\
\mu_x^2 \mu_{y1}' & (n_x + n_y)\sigma_{11}' \mu_y^2 + \mu_y^2 \mu_{x1} - \mu_{y1}' \\
-\mu_x' & -\mu_y'
\end{bmatrix}
$$

(3.5)

$\mu_{x1}'$ and $\mu_{y1}'$ are the means of the trait for the individuals selected for transcriptional profiling with genotype $x$ and $y$, respectively, and $\sigma_{11}'$ is the variance of these selected
individuals conditioned on genotype. Here we assume that the selection mechanism is such that the variance of the selected individuals will be the same for each genotype. If the individuals used for transcriptional profiling are selected completely at random, then we have \( \mu'_{x1} = \mu_{x1} \), \( \mu'_{y1} = \mu_{y1} \), and \( \sigma'_{11} = \sigma_{11} \). Note that throughout this paper when citing asymptotic results, we require that \( N_x/N, N_y/N, n_x/N_x, \) and \( n_y/N_y \) converge to positive constants as \( N \) goes to infinity.

### 3.3.2 Testing for association between the expression of a gene and the identified QTL

As described in Section 3.2, we may identify genes whose expression is associated with QTL genotype by testing \( H_0 : \mu_{x2} = \mu_{y2} \) vs. \( H_A : \mu_{x2} \neq \mu_{y2} \). Considering the reparameterization discussed in subsection 3.3.1, this is equivalent to testing the equality

\[
\beta_{20x} + \beta_{21x1} = \beta_{20y} + \beta_{21y1}.
\]

We will consider three tests for this problem, namely a Wald test, likelihood ratio test, and score test.

Let

\[
\delta = \beta_{20x} + \beta_{21x1} - \beta_{20y} - \beta_{21y1},
\]

and let \( \hat{\delta} \) denote the maximum likelihood estimator of \( \delta \). Using the asymptotic variance of \( \hat{\delta} \) given by (3.4) and the asymptotic normality and consistency of \( \hat{\delta} \), it can be shown that

\[
\text{var}(\hat{\delta}) \cong \left\{ \left( \frac{1}{N_x} + \frac{1}{N_y} \right) \left( \frac{(1 - \rho^2)\sigma_{11}}{(n_x + n_y)\sigma_{11}^2} + \rho^2 \right) + \left( \frac{1}{n_x} + \frac{1}{n_y} \right) (1 - \rho^2) \right\} \sigma_{22}^2 \\
+ \frac{(\mu'_{x1} - \mu'_{y1} + \mu_{y1} - \mu_{x1})^2 \sigma_{22,1}}{(n_x + n_y)\sigma_{11}^2}.
\]

(3.7)

for large \( N \). This approximate variance can be derived by considering the asymptotic variances of and covariances among the terms \( \hat{\beta}_{20x}, \hat{\beta}_{21x1}, \hat{\beta}_{20y}, \) and \( \hat{\beta}_{21y1} \), while taking
advantage of the asymptotic independence of \( \hat{\beta}_{21}, \hat{\mu}_{x1}, \) and \( \hat{\mu}_{y1}. \) A Wald test statistic can then be constructed as \( W = \frac{\hat{\beta}_{21}^2}{\text{var}(\hat{\beta})}. \)

Examination of (3.7) reveals that \( \text{var}(\hat{\beta}) \) will decrease as \( \sigma'_{11} \) increases and \( (\mu'_{x1} - \mu'_{y1} + \mu_{y1} - \mu_{x1})^2 \) decreases. Thus, to minimize variance and thereby maximize power we propose to select an equal number of individuals with the highest and lowest trait values from each genotypic class. For this selection strategy, \( \sigma'_{11} \) will be maximized and

\[
\mu'_{x1} - \mu'_{y1} + \mu_{y1} - \mu_{x1} = 0. \tag{3.8}
\]

This optimal selection strategy is depicted in Figure 3.1.

For most applications, \( N_x \) and \( N_y \) will be large. Thus examination of (3.7) shows that satisfying (3.8) will be more important than maximizing \( \sigma'_{11}. \) Thus any selection strategy satisfying (3.8) would be expected to perform nearly as well as selecting individuals from both tails of the trait distribution within each genotype class. For example, selecting individuals completely at random from each genotype class should lead to good performance. Note, however, that selecting individuals from each tail of the trait distribution without consideration of genotype appears to be a poor selection scheme. With this strategy \( \sigma'_{11} \) will be small and, more importantly, \( (\mu'_{x1} - \mu'_{y1} + \mu_{y1} - \mu_{x1})^2 \) will be large when the individuals with the lowest trait values are predominantly of one genotype and individuals with the highest trait values are predominantly of the other genotype (as would be expected in our application).

To conduct the likelihood ratio test or score test, the restricted MLE under (3.6) can be obtained using the EM algorithm (see Appendix B). Once the restricted and unrestricted estimates have been obtained, the LRT can be conducted in the usual manner. For the score test, an estimate of \( I^{-1}(\phi) \) subject to the constraint (3.6) is required but may not be readily available. For the case of selection completely at random, however, an appropriate estimate of \( I^{-1}(\phi) \) is available since \( \mu'_{x1} = \mu_{x1}, \mu'_{y1} = \mu_{y1} \) and \( \sigma'_{11} = \sigma_{11}. \) For the strategy of selecting individuals with highest and lowest traits within
Figure 3.1  An idealized depiction of the joint distributions of trait and expression along with the marginal distributions of trait and expression conditional on genotype. Distributions for genotype \( x \) (\( y \)) are presented with solid (dashed) lines. We propose transcriptional profiling for individuals with the highest and lowest trait values in each genotypic class. These individuals are represented by the shaded regions on the trait axis. This will result in observing the joint distributions of trait and expression in the regions labeled with \( x \) or \( y \). The corresponding observed expression values are shaded on the expression axis.
each genotypic class, $\sigma'_{11}$ could be estimated as

$$\tilde{\sigma}_{11} = \frac{\sum_{i=1}^{n_x} (X_{i1} - \bar{\mu}_{x1})^2 + \sum_{j=1}^{n_y} (Y_{j1} - \bar{\mu}_{y1})^2}{n_x + n_y}$$

(3.9)

The Wald test, likelihood ratio test and score test are asymptotically equivalent, but the Wald test is simpler in both representation and computation. Simulations in Section 3.4 compare the three tests for finite sample sizes.

### 3.3.3 Testing for correlation between gene expression and the quantitative trait, conditional on QTL genotype

In addition to finding genes whose expression is associated with a known QTL for a traditional trait, researchers will be interested in identifying genes whose expression is correlated with the trait after accounting for the known effects of the QTL on the trait. Discovering such genes is important for understanding the genetic networks that underlie quantitative variation in the trait. Many of these genes may have expression that is associated with QTL genotype, and would therefore be identified as important via the tests described in Subsection 3.2. Other genes, however, may have expression values that are correlated with the trait but unassociated with genotype at the QTL. Such genes can be identified by testing $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$ using the test that we develop in this subsection.

Our estimator of $\rho$,

$$\hat{\rho} = \frac{\hat{\beta}_{21} \sqrt{\hat{\sigma}_{11}}}{\sqrt{\hat{\sigma}_{22.1} + \hat{\beta}_{21}^2 \hat{\sigma}_{11}}}$$

has an asymptotic normal distribution with mean $\rho$ and variance

$$\text{var}(\hat{\rho}) \cong (1 - \rho^2)^2 \left( \frac{\rho^2}{2(N_x + N_y)} + \frac{\sigma_{11}(1 - \rho^2)}{(n_x + n_y)\sigma'_{11}} + \frac{\rho^2}{2(n_x + n_y)} \right)$$

(3.10)

as derived using the Delta method and the inverse information matrix described in equations (3.4) and (3.5). To improve the normal approximation for small sample sizes,
an approach similar to the Fisher transformation can be derived using a strategy similar to Mendoza (1993) and Hotelling (1953). Briefly, let

\[ k = \frac{\sigma_{11}}{\sigma'_{11}}, \quad (3.11) \]

\[ N = n + n_y, \]

\[ n = n_x + n_y, \]

\[ t_1 = \frac{-3n^2 - 6nN + 20knN - 3N^2 + 20k^2N^2 - 28k^2N^2}{32N^2k^{\frac{5}{2}}}, \]

\[ t_2 = \frac{-n^2 - 2nN + 4knN - N^2 + 4kN^2 - 4k^2N^2}{32N^2k^{\frac{3}{2}}}, \]

\[ t_3 = \frac{3n^2 + 6nN - 20knN + 3N^2 - 20kN^2 + 60k^2N^2}{64N^2k^{\frac{5}{2}}}, \]

\[ Z(\rho) = t_1 \rho + t_2 \rho^3 + t_3 \ln \frac{1 + \rho}{1 - \rho}, \text{ and} \]

\[ Z(\hat{\rho}) \] denote the random variable obtained by replacing \( \sigma_{11}, \sigma'_{11}, \) and \( \rho \) by their estimators. Then \( Z(\hat{\rho}) \) has an asymptotic normal distribution with mean \( Z(\rho) \) and variance \( \frac{1}{n_x + n_y - 4} \) (see Appendix C for further details). This statistic is similar to that of Mendoza (1993), but the form of its variance is different due to the loss of an extra degree of freedom in the two population setting. Since \( Z(\hat{\rho}) \) approaches normality faster than \( \hat{\rho} \), it is a better statistic for testing \( H_0: \rho = 0 \). It can be verified that this test obtains maximum power when \( \sigma'_{11} \) is maximized by the selection strategy, which is the case when we select the individuals with the lowest and highest traits within each genotypic class. The value of \( \sigma'_{11} \) can be estimated using (3.9) when using this recommended strategy.

### 3.4 Simulation Study

We conducted a simulation study to evaluate the small-sample properties of the Wald test, likelihood ratio test, score test, and our proposed test for correlation. Throughout the simulations reported in this section, the difference in trait means for the two genotype classes was held constant at one unit (i.e., \( \mu_{x1} - \mu_{y1} = 1 \)), and the within-genotype
variances for trait and expression were held constant at one unit (i.e., \( \sigma_{11} = \sigma_{22} = 1 \)). Additional simulations not reported here indicate that varying these factors provides no further insight into the small-sample properties of the proposed estimators. In particular, the performance of the tests appears to be unaffected by changes in \( \mu_x - \mu_y \), and as expected, power increases as \( \sigma_{11} \) and/or \( \sigma_{22} \) decreases.

To evaluate methods for testing for association between QTL genotype and expression (\( H_0 : \mu_{x2} = \mu_{y2} \)), we examined the small-sample performance of the Wald test, likelihood ratio test, score test, and two implementations of a standard \( t \)-test. For a given difference in means \( \mu_{x2} - \mu_{y2} \) ranging from 0 to 1 unit in increments of 0.25 units, 10000 independent samples were generated. In each sample, a trait value and an expression level of a gene were simulated for \( N_x = 100 \) individuals of genotype \( x \) and for \( N_y = 100 \) individuals of genotype \( y \). The correlation between trait and expression was fixed at \( \rho = 0.85 \). Thirty individuals from each genotype were selected as the individuals with complete data for both trait and expression level (\( n_x = n_y = 30 \)). All other individuals were treated as if only trait values were available when performing the Wald test, likelihood ratio test or score test described in the previous section. Two strategies for selecting individuals for microarray experiments were considered here: (1) selection completely at random within each genotype group, and (2) selection of individuals with the top 15 and bottom 15 trait values within each genotype group. The results from the Wald, likelihood ratio, and score tests were compared with that from a two-sample \( t \)-test using only the expression data from the 30 selected individuals from each genotype group as well as a two-sample \( t \)-test using expression data for all 100 individuals. The latter method was included to determine the degree of power loss associated with selective transcriptional profiling.

Tables 3.1, 3.2, and 3.3 summarize the results for the 0.001-level, 0.01-level, and 0.05-level tests of \( H_0 : \mu_{x2} = \mu_{y2} \). For the simulation reported in Table 3.1, individuals used for the microarray experiment are selected at random from each genotype group. It is observed that the type I error and power of the Wald test, likelihood ratio test (LRT)
Figure 3.2 The power of the Wald test, LRT, score test, $t$-test using only expression levels of selected individuals (30 randomly selected from 100 in each genotypic class) and the $t$-test using the expression levels of all individuals. The expression level mean for individuals with genotype $x$ was fixed at 0. The mean for individuals with genotype $y$ varied from 0 to 1.75 in increments of 0.25. The within-genotype correlation between expression and trait was fixed at $\rho = 0.85$.

and score test are very similar, though the Wald test might have a slightly higher type I error and power than that of LRT, which in turn might be slightly higher than that of score test. All three tests are far superior to the $t$-test using only the expression level of selected individuals, especially when the mean difference is between .25 and .75. The $t$-test using the expression level of all individuals is the most powerful of all five, which is as expected. These results are also shown in Figure 3.2, which shows that the Wald test, LRT or score test offer a substantial improvement over the $t$-test using only the expression level from selected individuals.
Table 3.1 The type I error and power of the Wald test, likelihood ratio test (LRT), score test, $t$-test on the expression level of the randomly selected individuals (30 randomly selected from 100 in each genotypic class) and $t$-test on the expression level of all individuals. The trait mean is 0 for individuals with genotype $y$ and 1 for individuals with genotype $x$. The mean of expression is 0 for individuals of genotype $y$. The within-genotype correlation between expression and trait was fixed at $p = 0.85$. The type I error rate and power are reported for test size of .001, .01 and .05, and results appear in that order.

<table>
<thead>
<tr>
<th>$\mu_{x2} - \mu_{y2}$</th>
<th>Wald</th>
<th>LRT</th>
<th>Score (selected)</th>
<th>$t$-test (selected)</th>
<th>$t$-test (complete)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.0017</td>
<td>0.0012</td>
<td>0.0009</td>
<td>0.0008</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>0.0124</td>
<td>0.0110</td>
<td>0.0095</td>
<td>0.0104</td>
<td>0.0083</td>
</tr>
<tr>
<td></td>
<td>0.0576</td>
<td>0.0548</td>
<td>0.0515</td>
<td>0.0521</td>
<td>0.0483</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0322</td>
<td>0.0254</td>
<td>0.0208</td>
<td>0.0089</td>
<td>0.0613</td>
</tr>
<tr>
<td></td>
<td>0.1289</td>
<td>0.1197</td>
<td>0.1115</td>
<td>0.0496</td>
<td>0.2074</td>
</tr>
<tr>
<td></td>
<td>0.2953</td>
<td>0.2891</td>
<td>0.2806</td>
<td>0.1632</td>
<td>0.4222</td>
</tr>
<tr>
<td>0.50</td>
<td>0.3135</td>
<td>0.2845</td>
<td>0.2512</td>
<td>0.0727</td>
<td>0.5767</td>
</tr>
<tr>
<td></td>
<td>0.5790</td>
<td>0.5635</td>
<td>0.5453</td>
<td>0.2419</td>
<td>0.8235</td>
</tr>
<tr>
<td></td>
<td>0.7890</td>
<td>0.7833</td>
<td>0.7771</td>
<td>0.4681</td>
<td>0.9381</td>
</tr>
<tr>
<td>0.75</td>
<td>0.8057</td>
<td>0.7809</td>
<td>0.7478</td>
<td>0.3001</td>
<td>0.9744</td>
</tr>
<tr>
<td></td>
<td>0.9385</td>
<td>0.9338</td>
<td>0.9254</td>
<td>0.5921</td>
<td>0.9961</td>
</tr>
<tr>
<td></td>
<td>0.9836</td>
<td>0.9828</td>
<td>0.9819</td>
<td>0.8172</td>
<td>0.9995</td>
</tr>
<tr>
<td>1.00</td>
<td>0.9864</td>
<td>0.9828</td>
<td>0.9771</td>
<td>0.6511</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.9985</td>
<td>0.9983</td>
<td>0.9979</td>
<td>0.8794</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9677</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Table 3.2  The power of the Wald test, LRT and score test when selecting individuals with extreme trait values in each genotypic class (top 15 and bottom 15 of 100 in each genotypic class), compared with the t-test using only selected individuals or all individuals. All other conditions are the same as Table 3.1.

<table>
<thead>
<tr>
<th>( \mu_{x2} - \mu_{y2} )</th>
<th>Wald (selected)</th>
<th>LRT (selected)</th>
<th>Score (selected)</th>
<th>( t )-test (selected)</th>
<th>( t )-test (complete)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>0.0090</td>
<td>0.0084</td>
<td>0.0076</td>
<td>0.0000</td>
<td>0.0083</td>
</tr>
<tr>
<td></td>
<td>0.0503</td>
<td>0.0486</td>
<td>0.0465</td>
<td>0.0000</td>
<td>0.0483</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0325</td>
<td>0.0280</td>
<td>0.0247</td>
<td>0.0000</td>
<td>0.0613</td>
</tr>
<tr>
<td></td>
<td>0.1218</td>
<td>0.1159</td>
<td>0.1106</td>
<td>0.0000</td>
<td>0.2074</td>
</tr>
<tr>
<td></td>
<td>0.2923</td>
<td>0.2877</td>
<td>0.2837</td>
<td>0.0060</td>
<td>0.4222</td>
</tr>
<tr>
<td>0.50</td>
<td>0.3192</td>
<td>0.2996</td>
<td>0.2736</td>
<td>0.0000</td>
<td>0.5767</td>
</tr>
<tr>
<td></td>
<td>0.5880</td>
<td>0.5768</td>
<td>0.5630</td>
<td>0.0072</td>
<td>0.8235</td>
</tr>
<tr>
<td></td>
<td>0.8011</td>
<td>0.7968</td>
<td>0.7927</td>
<td>0.0998</td>
<td>0.9381</td>
</tr>
<tr>
<td>0.75</td>
<td>0.8155</td>
<td>0.7983</td>
<td>0.7764</td>
<td>0.0045</td>
<td>0.9744</td>
</tr>
<tr>
<td></td>
<td>0.9421</td>
<td>0.9383</td>
<td>0.9344</td>
<td>0.0989</td>
<td>0.9961</td>
</tr>
<tr>
<td></td>
<td>0.9853</td>
<td>0.9847</td>
<td>0.9838</td>
<td>0.4713</td>
<td>0.9995</td>
</tr>
<tr>
<td>1.00</td>
<td>0.9867</td>
<td>0.9835</td>
<td>0.9808</td>
<td>0.0668</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.9981</td>
<td>0.9981</td>
<td>0.9978</td>
<td>0.4634</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.8682</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

To examine the possible effect of selection strategies, a second set of simulations were performed under the same situations except that in selecting individuals for microarray experiments, we selected the top 15 and bottom 15 individuals in each genotype group according to the trait value. The results are shown in Table 3.2. The power results when selecting individuals with extreme trait values within each genotypic group are very similar to the power results obtained when using random selection as shown in Table 3.1, which is consistent with our asymptotic result. The type I error rate and power for the \( t \)-test using only selected individuals are very low in Table 3.2 due to non-normality of the small samples.

Because of the simplicity of computation and interpretation of the Wald test, it is our preferred method of testing for association between QTL genotype and gene expression.
Table 3.3 The type I error rate for the Wald test with different overall number of individuals and the number of selected individuals in each genotype. The nominal type I error rate is .001, .01 and .05 respectively. Results are based on selection of individuals with highest and lowest traits within each genotype class.

<table>
<thead>
<tr>
<th>$N_x = N_y$</th>
<th>$\frac{n_x}{N_x}$</th>
<th>$\frac{n_y}{N_y}$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.0039</td>
<td>0.0028</td>
<td>0.0017</td>
<td>0.0022</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0199</td>
<td>0.0163</td>
<td>0.0137</td>
<td>0.0157</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0687</td>
<td>0.0663</td>
<td>0.0579</td>
<td>0.0607</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.0017</td>
<td>0.0011</td>
<td>0.0013</td>
<td>0.0011</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0134</td>
<td>0.0109</td>
<td>0.0101</td>
<td>0.0097</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0592</td>
<td>0.0557</td>
<td>0.0537</td>
<td>0.0516</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.0009</td>
<td>0.0013</td>
<td>0.0014</td>
<td>0.0020</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0090</td>
<td>0.0105</td>
<td>0.0105</td>
<td>0.0112</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0500</td>
<td>0.0515</td>
<td>0.0530</td>
<td>0.0545</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3 summarizes the type I error rate of the 0.001-level, 0.01-level, and 0.05-level Wald tests for varying values of $N_x = N_y$ and $n_x = n_y$. The type I error rate tends to be higher than the nominal level with the greatest inflation in type I error rates occurring when both the overall number and the number of selected individuals are very low. Note, however, that even in the worst case, the observed type I error rate is not grossly higher than nominal.

Another question of interest is the effect of within-genotype correlation on the power of test. Results of simulations under different within-genotype correlations are shown in Figure 3.3. As expected, the power of test increases with the value of correlation, and the increase is most substantial when the correlation coefficient is at least 0.6.

In testing the significance of the within-genotype correlation coefficient, simulations were conducted under similar conditions used for Table 3.1 and Table 3.2, but the correlation coefficient was varied from 0 to 0.7. The statistic $\tilde{Z}(\hat{\rho})$ was generated to test the null hypothesis $H_0: \rho = 0$, as described in Section 3.3. The type I error rate and
Figure 3.3 The power of the Wald test under random selection with regard to different within-genotype correlation coefficient values. The means of trait and expression level for \( y \) individuals are fixed at to be \((0,0)\). The mean vector for \( x \) individuals is \((1,.75)\) or \((1,.5)\).
Table 3.4 The type I error rate and power in testing whether the within-genotype correlation coefficient is 0 using the method described in Section 3.3. The compared selection strategies are random selection within each genotype class (Random) and selection of the individuals with the most extreme quantitative traits within each genotype class (Tails).

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.0552</td>
<td>0.1310</td>
<td>0.3744</td>
<td>0.6722</td>
<td>0.9039</td>
<td>0.9883</td>
<td>0.9999</td>
<td>1.0000</td>
</tr>
<tr>
<td>Tails</td>
<td>0.0526</td>
<td>0.2296</td>
<td>0.7018</td>
<td>0.9693</td>
<td>0.9994</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The power of the test are reported in Table 3.4.

Obviously the test is more powerful when individuals with extreme trait values are selected from each genotypic class. Further simulations showed that this method gives the correct type I error rate even when the number of individuals is small (for example, $N_x = N_y = 50$ and $n_x = n_y = 10$).

### 3.5 Uncertainty in QTL Genotype

We have assumed that QTL genotypes are essentially known for all individuals in the QTL mapping population. In reality only the genotypes at genetic markers near estimated QTL positions are known. There are two related sources of uncertainty in the QTL genotype for each individual: (i) uncertainty in QTL genotype given its flanking marker genotypes, and (ii) uncertainty in the marker interval containing the QTL. The former source of uncertainty could be accounted for through a mixture model as in interval mapping of QTL (Lander and Botstein 1989; Carbonell et al. 1992). There is no conceptual difficulty in this development, though there will not be a closed form solution for the maximum likelihood estimates, and additional computational methods will be required. Variation in QTL position is potentially larger and more challenging to accommodate in a formal manner.
We conducted a simulation study to determine the probability and impact of misclassifying individual QTL genotypes when using the markers flanking the estimated QTL position to predict QTL genotypes. A single chromosome of length 1.1 Morgans was simulated with 12 markers spaced 10 centiMorgans (cM) apart. A single QTL was positioned in the center of the chromosome midway between the 6th and 7th markers. Marker and QTL genotypes were simulated using a backcross design and Haldane’s map function (Haldane, 1919). Traits were normally distributed with standard deviation 1 and a mean that depended on QTL genotype. QTL effect sizes ranging from 0.25 to 2.5 in increments of 0.25 units were considered. For each effect size, 10,000 data sets were generated with 200 individuals in each data set.

Using the approach of Haley and Knott (1992), test statistics were computed at 1 cM increments along the chromosome to scan for the presence of a QTL for each randomly generated data set. Simulation was used to determine the appropriate threshold for chromosome-wise significance at the 0.05 level. A QTL was estimated to be at position \( p \) if the test statistic corresponding to position \( p \) exceeded the threshold for significance and was the largest test statistic on the chromosome. The genotypes of the markers flanking an estimated QTL position were used to predict the QTL genotype of each individual with non-recombinant flanking marker genotypes. Estimated mean error rates ranged from 0.099 for effect size 0.25 to 0.003 for effect size 2.5. Estimated median error rates were considerably smaller ranging from 0.055 for effect size 0.25 to 0 for effect sizes greater than 1. Complete results are provided in Table 3.5.

To determine the impact of QTL genotype misclassification on the power of our transcriptional profiling approach, we examined power of our Wald test for both selection strategies as a function of \( \mu_{x2} - \mu_{y2} \) (effect of QTL on transcription) for QTL trait effect sizes \( \mu_{x1} - \mu_{y1} = 0, \pm 0.25, \cdots, \pm 1.25 \) and corresponding misclassification rates slightly higher than estimated in our simulation. Loss of power was greatest when the effects of the QTL on trait and expression were both small. For the worst-case scenario when the
Table 3.5  Mean and median misclassification error rates and power for QTL detection for varying QTL effect sizes.

<table>
<thead>
<tr>
<th>QTL Effect</th>
<th>Mean Error Rate</th>
<th>Median Error Rate</th>
<th>Detection Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.099</td>
<td>0.055</td>
<td>0.269</td>
</tr>
<tr>
<td>0.50</td>
<td>0.050</td>
<td>0.034</td>
<td>0.811</td>
</tr>
<tr>
<td>0.75</td>
<td>0.023</td>
<td>0.006</td>
<td>0.993</td>
</tr>
<tr>
<td>1.00</td>
<td>0.013</td>
<td>0.005</td>
<td>1.000</td>
</tr>
<tr>
<td>1.25</td>
<td>0.008</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.50</td>
<td>0.006</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.75</td>
<td>0.004</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2.00</td>
<td>0.003</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2.25</td>
<td>0.003</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2.50</td>
<td>0.003</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The effect of the QTL on expression and trait was 0.25 in magnitude, power dropped from just under 30% (see Tables 1 and 2) to just over 20%. Power loss for other scenarios was not as extreme. For example the power achieved in the presence of misclassification error was typically more than 90% of that achieved assuming no misclassification error, and the loss was negligible when the QTL effect on expression was 1 or greater. The full set of results are provided in Table 3.6. We have seen that misclassification probabilities and the impact of misclassification on power are greatest when the effect of the QTL on the trait is small. Our transcriptional profiling approach begins after a major trait QTL has been successfully mapped. Thus, in practice we will be dealing with QTL whose effects on the trait were large enough to be detected. The approximate power for picking up the 0.25 trait effect was estimated to be slightly less than 30% in our QTL simulation. Thus, although the misclassification rates were high and power cut by nearly a third in this case, other cases with larger effects, higher detection powers, and less substantial misclassification errors are more likely to be encountered in practice.

Impact of misclassification on our proposed test for within-genotype correlation between expression and trait was also examined via simulation. Loss of power relative to
Table 3.6 Estimated power of the Wald test for detecting association between QTL genotype and gene expression for varying misclassification error rates, selection strategies, and QTL effects on trait and expression. Power estimates are based on 2000 simulation replications of the nominal 0.05 level Wald test with 30 individuals selected from 100 individuals for each QTL genotype. The selection strategies are random selection within each genotype class (random) and selection of the individuals with the most extreme quantitative traits within each genotype class (tails).

<table>
<thead>
<tr>
<th>Error Rate</th>
<th>Selection Strategy</th>
<th>QTL Effect</th>
<th>$\mu_{x2} - \mu_{y2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>random</td>
<td>NA</td>
<td>0.061</td>
</tr>
<tr>
<td>0.00</td>
<td>tails</td>
<td>NA</td>
<td>0.055</td>
</tr>
<tr>
<td>0.10</td>
<td>random</td>
<td>-0.25</td>
<td>0.060</td>
</tr>
<tr>
<td>0.10</td>
<td>random</td>
<td>0.25</td>
<td>0.051</td>
</tr>
<tr>
<td>0.10</td>
<td>tails</td>
<td>-0.25</td>
<td>0.062</td>
</tr>
<tr>
<td>0.10</td>
<td>tails</td>
<td>0.25</td>
<td>0.060</td>
</tr>
<tr>
<td>0.05</td>
<td>random</td>
<td>-0.50</td>
<td>0.057</td>
</tr>
<tr>
<td>0.05</td>
<td>random</td>
<td>0.50</td>
<td>0.060</td>
</tr>
<tr>
<td>0.05</td>
<td>tails</td>
<td>-0.50</td>
<td>0.059</td>
</tr>
<tr>
<td>0.05</td>
<td>tails</td>
<td>0.50</td>
<td>0.058</td>
</tr>
<tr>
<td>0.02</td>
<td>random</td>
<td>-0.75</td>
<td>0.052</td>
</tr>
<tr>
<td>0.02</td>
<td>random</td>
<td>0.75</td>
<td>0.054</td>
</tr>
<tr>
<td>0.02</td>
<td>tails</td>
<td>-0.75</td>
<td>0.053</td>
</tr>
<tr>
<td>0.02</td>
<td>tails</td>
<td>0.75</td>
<td>0.057</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>-1.00</td>
<td>0.056</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>1.00</td>
<td>0.049</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>-1.00</td>
<td>0.052</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>1.00</td>
<td>0.059</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>-1.25</td>
<td>0.051</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>1.25</td>
<td>0.065</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>-1.25</td>
<td>0.057</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>1.25</td>
<td>0.053</td>
</tr>
</tbody>
</table>
Table 3.7 Estimated power of the test for within-QTL-genotype correlation between expression and trait for varying misclassification error rates, selection strategies, and true correlations. Power estimates are based on 2000 simulation replications of the nominal 0.05 level test with 30 individuals selected from 100 individuals for each QTL genotype. The selection strategies are random selection within each genotype class (random) and selection of the individuals with the most extreme quantitative traits within each genotype class (tails).

<table>
<thead>
<tr>
<th>Error Rate</th>
<th>Selection Strategy</th>
<th>QTL Effect</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>tails</td>
<td>NA</td>
<td>0.049</td>
<td>0.233</td>
<td>0.711</td>
<td>0.958</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.00</td>
<td>random</td>
<td>NA</td>
<td>0.054</td>
<td>0.113</td>
<td>0.329</td>
<td>0.616</td>
<td>0.881</td>
<td>0.979</td>
<td>0.999</td>
</tr>
<tr>
<td>0.10</td>
<td>tails</td>
<td>0.25</td>
<td>0.063</td>
<td>0.289</td>
<td>0.754</td>
<td>0.977</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.10</td>
<td>tails</td>
<td>-0.25</td>
<td>0.066</td>
<td>0.145</td>
<td>0.559</td>
<td>0.913</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.10</td>
<td>random</td>
<td>0.25</td>
<td>0.065</td>
<td>0.159</td>
<td>0.376</td>
<td>0.697</td>
<td>0.910</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td>0.10</td>
<td>random</td>
<td>-0.25</td>
<td>0.065</td>
<td>0.093</td>
<td>0.267</td>
<td>0.548</td>
<td>0.839</td>
<td>0.969</td>
<td>0.998</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>1.00</td>
<td>0.055</td>
<td>0.292</td>
<td>0.752</td>
<td>0.974</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.01</td>
<td>tails</td>
<td>-1.00</td>
<td>0.056</td>
<td>0.186</td>
<td>0.622</td>
<td>0.954</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>1.00</td>
<td>0.057</td>
<td>0.144</td>
<td>0.364</td>
<td>0.673</td>
<td>0.918</td>
<td>0.993</td>
<td>1.000</td>
</tr>
<tr>
<td>0.01</td>
<td>random</td>
<td>-1.00</td>
<td>0.059</td>
<td>0.109</td>
<td>0.293</td>
<td>0.617</td>
<td>0.877</td>
<td>0.984</td>
<td>1.000</td>
</tr>
</tbody>
</table>

that reported in Table 3.4 was relatively minor. There was no detectable loss of power when the slope of the line segment connecting the points \((\mu_{x1}, \mu_{x2})\) and \((\mu_{y1}, \mu_{y2})\) had the same sign as the within-genotype correlation coefficient (as in Figure 3.1). Power typically dropped by a few percentage points when the sign of the slope and the correlation coefficient differed. In the worst case, power was slightly more than 70% of that achieved with no misclassification error. The details are shown in Table 3.7.

3.6 Discussion

Much effort has been devoted to mapping QTL that control various traits, but to further understand the genetic mechanisms through which QTL may affect traits, it is necessary to test whether a given QTL affects the expression of other genes. The
genetical genomics approach proposed by Jansen and Nap (2001) offers a wealth of data that can be used to identify genes whose expression is linked to a QTL and to address many other interesting and important questions. The cost of measuring expression in a large number of individuals, however, prevents many researchers from attempting full-scale genetical genomics studies. The work that we present here has been motivated by our interactions with scientists who wish to understand the genetic mechanisms through which a QTL affects a trait of interest but cannot afford to measure expression of all individuals in their QTL mapping population. They often ask the question: “Given that we cannot measure the expression of all individuals, which subset of individuals should we measure?” We have proposed selective transcriptional profiling as an answer to this question.

In this paper we have described an optimal selection strategy for the goal of identifying genes whose expression is linked to a QTL or correlated with the trait of interest. Furthermore, we have developed statistical tests that make appropriate use of the selected data to identify genes of interest. This is a first step in the very complex problem of identifying genetic pathways that are linked to a QTL of interest. Selective transcriptional profiling is ideal for this kind of study as it provides power that approaches that of full-scale genetical genomics at far less cost.

We compared the Wald test, likelihood ratio test, and score test as methods for using the selected data to identify linkages between QTL and gene expressions. The performance of the three tests was quite similar, but we stated a preference for the Wald test due to its simplicity in form and in computation. In practice the test would need to be executed for several thousand genes; thus computational efficiency is an important consideration.

It is also important to consider simultaneous inference issues when conducting thousands of tests. Such issues have been dealt with extensively in recent microarray literature where conducting one test for each of thousands of genes has become routine.
Selective transcriptional profiling will produce two sets of $p$-values where each set contains one $p$-value for each of several thousand genes. One set will consist of $p$-values corresponding to tests of linkage between gene expression and the QTL while the other set will contain $p$-values for the tests of correlation between gene expression and trait, conditional on QTL genotype. A false discovery rate (FDR) approach as described by Storey and Tibshirani (2003) could be used to identify significant tests within each set while managing FDR at a level specified by the investigator.

### 3.7 Future Work

Results are beginning to become available from a few large-scale genetical genomics studies involving quantitative traits, molecular markers, and gene expression from hundreds of individuals. It will be important to compare results obtained from analysis of the full data set to the results that would be obtained by examining only the subset of the data that would be available if our selective transcriptional profiling approach were employed. Certainly the full data will provide more information than can be obtained with our selective transcriptional profiling approach. However, we expect analysis of real data to show that much can be discovered with our selective transcriptional profiling approach, making it attractive when resource limitations preclude full-scale genetical genomics.

We have focused on the two-genotype case in this paper, but extending the work to handle multiple groups is important for two reasons. First many QTL mapping studies are conducted in $F_2$ populations where three genotypes occur at any given locus. Second researchers may be interested in simultaneously studying multiple QTL for a given trait. For $q$ different QTL, each with $m$ possible genotypes, a total of $qm$ classes should be considered. A variety of contrasts among the $qm$ expression group means may be of interest in this case. The extension of the methodology described in this paper
to the multiple group case is straightforward, but a substantial amount of derivation is involved. As an example, consider the three genotype scenario encountered in an experiment using an $F_2$ population. Let $Z_{i1}$ and $Z_{i2}$ designate the trait and expression value for individuals in the third genotype group; all other notation is analogous to previous notation. The interest now is to test $H_0$: $\mu_{x2} = \mu_{y2} = \mu_{z2}$, or equivalently $\gamma = 0$, where $\gamma = (\mu_{x2} - \mu_{y2}, \mu_{y2} - \mu_{z2})'$. The Wald test statistic is then constructed as $W = \hat{\gamma}'B^{-1}\hat{\gamma}$, with $B$ denoting the asymptotic covariance matrix of $\hat{\gamma}$. The maximum likelihood estimates are the natural generalization of those provided in equations (3.3) for the two-genotype case. Now let

$$
\Delta_{xy} = \mu_{x1}' - \mu_{y1}' + \mu_{y1} - \mu_{x1},
\Delta_{yz} = \mu_{y1}' - \mu_{z1}' + \mu_{z1} - \mu_{y1},
\Delta_{xz} = \mu_{z1}' - \mu_{x1}' + \mu_{x1} - \mu_{z1}.
$$

It can be shown that $|B| = \frac{(n_x + n_y + n_z)\hat{\sigma}^2_{21}}{n_x n_y n_z} + b_1 + b_2 + b_3 + b_4$, where

$$
b_1 = \frac{N_x + N_y + N_z}{N_x N_y N_z} \hat{\sigma}^2_{21},
b_2 = \left( \frac{1}{N_x n_y} + \frac{1}{N_y n_z} + \frac{1}{N_z n_x} + \frac{1}{N_x n_y} + \frac{1}{N_z n_x} + \frac{1}{N_z n_y} \right) \hat{\sigma}_{11} \hat{\sigma}_{22} \hat{\beta}_{21},
b_3 = \frac{\hat{\sigma}_{11} \hat{\sigma}_{22} \hat{\beta}_{21}}{(n_x + n_y + n_z)\hat{\sigma}_{11}} \left( \frac{\Delta_{xy}^2}{N_x} + \frac{\Delta_{yz}^2}{N_y} + \frac{\Delta_{xz}^2}{N_z} \right),
b_4 = \frac{\hat{\sigma}_{22}^2}{(n_x + n_y + n_z)\hat{\sigma}_{11}} \left( \frac{\Delta_{xy}^2}{n_x} + \frac{\Delta_{yz}^2}{n_y} + \frac{\Delta_{xz}^2}{n_z} \right).
$$

It is obvious that $|B|$ is minimized by a selection strategy for which $\Delta_{xy}$, $\Delta_{yz}$ and $\Delta_{xz}$ are zero. Thus selection strategies analogous to those proposed for the two-genotype case are recommended when three QTL genotypes are present. The likelihood ratio and score tests can also be extended in a similar fashion.

We have assumed the joint distribution of trait and expression to be bivariate normal. While the assumption may be viable in many cases and approximately correct
following transformation in others, there will be situations where this assumption is unreasonable. Deriving nonparametric methods for testing the hypotheses that we consider is feasible. In particular, we are developing an approach based on empirical likelihood (Owen, 2001) that can be used with selective transcriptional profiling. The empirical likelihood approach will work for general distributions encountered in practice and can achieve power very similar to that of our parametric method when the bivariate normal assumption happens to be satisfied. We plan to report on this work in progress in a future publication.

3.8 References


3.9 Appendix

Derivation of the Information Matrix

Let \( \phi_1 = (\mu_{x1}, \mu_{y1}, \sigma_{11}) \) and \( \phi_2 = (\beta_{30x}, \beta_{30y}, \beta_{21}, \sigma_{22}) \). Also let \( D(\phi) \) represent the matrix of the second derivatives of the log-likelihood with regard to \( \phi \), then

\[
D(\phi) = \begin{bmatrix}
D(\phi_1) & 0 \\
0 & D(\phi_2)
\end{bmatrix},
\]

since all the derivatives involving both parameters in \( \phi_1 \) and \( \phi_2 \) are zero. Also

\[
D(\phi_1) = \begin{bmatrix}
-\frac{N_x}{\sigma_{11}} & 0 & -\frac{\sum_{i=1}^{N_x}(X_{i1}-\mu_{x1})}{\sigma_{11}^2} \\
0 & -\frac{N_y}{\sigma_{11}} & -\frac{\sum_{j=1}^{N_y}(Y_{j1}-\mu_{y1})}{\sigma_{11}^2} \\
\frac{\sum_{i=1}^{N_x}(X_{i1}-\mu_{x1})}{\sigma_{11}^2} & \frac{\sum_{j=1}^{N_y}(Y_{j1}-\mu_{y1})}{\sigma_{11}^2} & D(\phi_1)_{33}
\end{bmatrix},
\]
where

\[ D(\phi_1)_{33} = \frac{N_x + N_y}{2\sigma_{11}^2} - \sum_{i=1}^{N_x} \frac{(X_{i1} - \mu_{x1})^2}{\sigma_{11}^2} + \sum_{j=1}^{N_y} \frac{(Y_{j1} - \mu_{y1})^2}{\sigma_{11}^2}. \]

For the elements in \( D(\phi_2) \),

\[ D(\phi_2)_{11} = -\frac{n_x}{\sigma_{22.1}}, \]

\[ D(\phi_2)_{12} = 0, \]

\[ D(\phi_2)_{13} = -\sum_{i=1}^{n_x} X_{i1}, \]

\[ D(\phi_2)_{14} = -\sum_{i=1}^{n_x} \frac{(X_{i2} - \beta_{20x} - \beta_{21}X_{i1})}{\sigma_{22.1}^2}, \]

\[ D(\phi_2)_{22} = -\frac{n_y}{\sigma_{22.1}}, \]

\[ D(\phi_2)_{23} = -\sum_{j=1}^{n_y} Y_{j1}, \]

\[ D(\phi_2)_{24} = -\sum_{j=1}^{n_y} \frac{(Y_{j2} - \beta_{20y} - \beta_{21}Y_{j1})}{\sigma_{22.1}^2}, \]

\[ D(\phi_2)_{33} = -\sum_{i=1}^{n_x} \frac{X_{i1}^2}{\sigma_{22.1}^2} + \sum_{j=1}^{n_y} \frac{Y_{j1}^2}{\sigma_{22.1}^2}, \]

\[ D(\phi_2)_{34} = \sum_{i=1}^{n_x} \frac{(X_{i2} - \beta_{20x} - \beta_{21}X_{i1})X_{i1}}{\sigma_{22.1}^2} + \sum_{j=1}^{n_y} \frac{(Y_{j2} - \beta_{20y} - \beta_{21}Y_{j1})Y_{j1}}{\sigma_{22.1}^2}, \]

and

\[ D(\phi_2)_{44} = \sum_{i=1}^{n_x} \frac{(X_{i2} - \beta_{20x} - \beta_{21}X_{i1})^2}{\sigma_{22.1}^3} + \sum_{j=1}^{n_y} \frac{(Y_{j2} - \beta_{20y} - \beta_{21}Y_{j1})^2}{\sigma_{22.1}^3} + \frac{n_x + n_y}{2\sigma_{22.1}^2}. \]

The information matrix can then be derived by taking expectations,

\[ I(\phi_1) = \begin{bmatrix} \frac{N_x}{\sigma_{11}} & 0 \\ 0 & \frac{N_y}{\sigma_{11}} \end{bmatrix}, \]

\[ I(\phi_2) = \begin{bmatrix} \frac{n_x}{\sigma_{22.1}} & 0 & \frac{n_x\mu_{x1}}{\sigma_{22.1}} & 0 \\ 0 & \frac{n_y}{\sigma_{22.1}} & \frac{n_y\mu_{y1}}{\sigma_{22.1}} & 0 \\ \frac{n_x\mu_{x1}}{\sigma_{22.1}} & \frac{n_y\mu_{y1}}{\sigma_{22.1}} & \frac{(n_x+n_y)\sigma_{11}^3 + n_x\mu_{x1}^2 + n_y\mu_{y1}^2}{\sigma_{22.1}^3} & 0 \\ 0 & 0 & 0 & \frac{n_x+n_y}{2\sigma_{22.1}^2} \end{bmatrix}, \]

and

\[ I(\phi) = \begin{bmatrix} I(\phi_1) & 0 \\ 0 & I(\phi_2) \end{bmatrix}. \]
Its inverse is shown in equation (3.4).

**Using the EM Algorithm to Obtain Restricted Maximum Likelihood Estimates**

Computation of the likelihood ratio and score test statistics for testing \( H_0 : \mu_{x2} = \mu_{y2} \) requires maximum likelihood estimates of \( \mu_{x2} \) and \( \mu_{y2} \) subject to the restriction \( \mu_{x2} = \mu_{y2} \). The estimates can be obtained via the EM algorithm using the common approach for exponential families (see Little and Rubin (2002)). Under the restriction \( \mu_{x2} = \mu_{y2} = \mu_2 \), we define the parameter vector

\[
\theta_r = (\mu_{x1}, \mu_{y1}, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12}).
\]

Given the current parameters estimates \( \theta_r^{(h)} \), the \((h+1)\)th iteration of EM is as follows:

**E step:** For incomplete observations, compute

\[
\begin{align*}
A_{x}^{(h)} &= \sum_{i=1}^{n_x} X_{i2} + \sum_{i=n_x+1}^{n_x+n_y} a_{xi}^{(h)}, \\
B_{x}^{(h)} &= \sum_{i=1}^{n_x} X_{i2}^2 + \sum_{i=n_x+1}^{n_x+n_y} b_{xi}^{(h)}, \\
C_{x}^{(h)} &= \sum_{i=1}^{n_x} X_{i1}X_{i2} + \sum_{i=n_x+1}^{n_x+n_y} c_{xi}^{(h)}, \\
A_{y}^{(h)} &= \sum_{i=1}^{n_y} Y_{i2} + \sum_{i=n_x+1}^{n_x+n_y} a_{yi}^{(h)}, \\
B_{y}^{(h)} &= \sum_{i=1}^{n_y} Y_{i2}^2 + \sum_{i=n_x+1}^{n_x+n_y} b_{yi}^{(h)}, \\
C_{y}^{(h)} &= \sum_{i=1}^{n_y} Y_{i1}Y_{i2} + \sum_{i=n_x+1}^{n_x+n_y} c_{yi}^{(h)},
\end{align*}
\]

Let
\[ A_{y}^{(h)} = \sum_{i=1}^{n_y} Y_{12} + \sum_{i=n_y+1}^{N_y} a_{yi}^{(h)}, \]

\[ B_{y}^{(h)} = \sum_{i=1}^{n_y} Y_{12}^2 + \sum_{i=n_y+1}^{N_y} b_{yi}^{(h)}, \] and

\[ C_{y}^{(h)} = \sum_{i=1}^{n_y} Y_{12} Y_{12} + \sum_{i=n_y+1}^{N_y} c_{yi}^{(h)}. \]

M step: Compute the new estimates of \( \theta_{y}^{(h+1)} \) by

\[ \mu_{x1}^{(h+1)} = \bar{X}_{1}, \]

\[ \mu_{y1}^{(h+1)} = \bar{Y}_{1}, \]

\[ \mu_{2}^{(h+1)} = \frac{A_{x}^{(h)} + A_{y}^{(h)}}{N_x + N_y}, \]

\[ \sigma_{11}^{(h+1)} = \sum_{i=1}^{N_x} X_{i1}^2 - N_x \bar{X}_{1}^2 + \sum_{j=1}^{N_y} Y_{j1}^2 - N_y \bar{Y}_{1}^2, \]

\[ \sigma_{22}^{(h+1)} = \frac{B_{x}^{(h)} - A_{x}^{(h)} / N_x + B_{y}^{(h)} - A_{y}^{(h)} / N_y}{N_x + N_y}, \] and

\[ \sigma_{12}^{(h+1)} = \frac{C_{x}^{(h)} - A_{x}^{(h)} \bar{X}_{1} + C_{y}^{(h)} - A_{y}^{(h)} \bar{Y}_{1}}{N_x + N_y}. \]

Also compute \( \sigma_{22,1}^{(h+1)} = \sigma_{22}^{(h+1)} - \sigma_{12}^{(h+1)} \sigma_{11}^{(h+1)} \).

Derivation of the Variance Stabilizing Transformation of \( \rho \)

From Subsection 3.3.3

\[ \text{var}(\rho) \approx (1 - \rho^2)^2 \left( \frac{\rho^2}{2N} + \frac{k(1 - \rho^2)}{n} + \frac{\rho^2}{2n} \right). \]

A function \( f(\rho) \) will provide a variance stabilizing transformation similar to the well known Fisher’s Z transformation of Pearson’s correlation coefficient if

\[
\frac{d}{d\rho} f(\rho) = \frac{1}{\sqrt{n \cdot \text{var}(\rho)}} = \frac{(2N)^{\frac{1}{2}}}{(1 - \rho^2) \sqrt{n \rho^2 + 2Nk(1 - \rho^2) + N \rho^2}}.
\] (3.12)
To derive \( f(\rho) \) note that using a Taylor expansion about \( \rho = 0 \) yields,

\[
\frac{1}{\sqrt{n\rho^2 + 2Nk(1 - \rho^2) + N\rho^2}} = \frac{1}{(2Nk)^{\frac{1}{2}}} \frac{n - 2Nk + N}{2^{\frac{3}{2}}(Nk)^{\frac{3}{4}}} \rho^2 + \frac{3(n - 2Nk + N)^2}{2^{\frac{3}{2}}(2Nk)^{\frac{5}{4}}} \rho^4 + O(\rho^6).
\]

Substituting (3.13) into (3.12) and integrating yields

\[
f(\rho) \equiv t_1 \rho + t_2 \rho^3 + t_3 \ln \frac{1 + \rho}{1 - \rho},
\]

which leads to the variance stabilizing transformation shown in (3.11).
Abstract: Selective transcriptional profiling is an attractive approach for alleviating the high cost of genetical genomics research. The method described in Wang and Nettleton (2005) is based on parametric models that might not be appropriate for all experiments. In this article, we derive empirical likelihood based inference for multi-sample comparison when information is available on surrogate variables. This provides a nonparametric method for studies using selective transcriptional profiling approach. It is shown that when testing for the association between the transcription abundance of a given gene and a known QTL, the principle of using relatively inexpensive trait data on extra individuals to improve the efficiency of the test also applies to an empirical likelihood based method. Thus we can extend the application of selective transcriptional profiling to situations where the assumptions of normality and equal variance between genotypes may be problematic. The theoretical result is derived in the framework of estimating equations, so it might be applied to other multi-sample comparison problems as well.

KEY WORDS: Empirical Likelihood, Microarray; Quantitative Trait Locus; Selective Transcriptional Profiling; Surrogate Variable.
4.1 Introduction

Selective transcriptional profiling was proposed by Wang and Nettleton (2005) as an approach to improve the efficiency and affordability of genetical genomics. In common genetical genomics studies (Jansen and Nap, 2001; Brem et al., 2002; Yvert et al., 2003), the transcriptional abundance of each gene as measured by microarray experiments is treated as a traditional trait like crop yield or body fat, and the usual QTL mapping strategy is carried out for the expression level of each gene. Since it is often the case that hundreds of microarrays are needed to achieve the desirable power, financial considerations limit this approach to few well funded researchers. Selective transcriptional profiling involves selecting an optimal subset of individuals to microarray from a larger set of individuals for which relatively inexpensive quantitative trait and molecular marker data are available. It is shown that by using microarray data from the selected individuals, along with the trait and marker data from all individuals, one can identify genes whose transcript abundance is associated with a quantitative trait of interest through linkage to a trait QTL with far fewer microarrays than the traditional genetical genomics approach. Conversely, one can achieve much greater power in the test for QTL association with the same number of microarrays.

The method described in Wang and Nettleton (2005) is derived in a missing data framework, assuming that the expression abundance of a given gene and the value of a traditional trait have a bivariate normal distribution. Also it is assumed that the covariance matrix is the same for individuals in different genotypic groups. Though these assumptions are reasonable over a wide range of experiments (perhaps after transformation), there are situations where these assumptions may be hard to justify. Thus it is desirable to develop a nonparametric method for analyzing data obtained by selective transcriptional profiling when strict parametric assumptions are questionable.

The problem of selective transcriptional profiling is related to the study of surrogate
endpoints or auxiliary outcome data. The study of surrogate endpoints is of particular interest to the medical community. One important example involves the use of CD4 count as a surrogate variable for subsequent survival time for human immunodeficiency virus (HIV) patients. In fishery studies, the tonnage of catch reported by a fishing boat captain is used as a surrogate for the tonnage of catch reported by government appointed observers. In social economic studies, economic status of a household from previous surveys might be used if the status is not reported in the most recent survey; income reported by a survey respondent may be used if a valid social security number is not available to match tax records.

A surrogate variable study is usually formulated as follows. Data are composed of a validation sample \( (V) \) and a nonvalidation sample \( (\bar{V}) \). The validation sample is made up of observations with measurement on the variable of interest \( (Y) \), as well as a surrogate or auxiliary variable \( (S) \) and also some covariate information \( (X) \). The nonvalidation sample is made up of observations with only information on \( S \) and \( X \). The goal is to make inference on some parameter \( \beta \), which defines some characteristics regarding \( X \) and \( Y \). The main statistical problem in analysis with a surrogate variable is how to utilize the information carried in \( S \) about the parameter \( \beta \).

Considerable research has been done to better utilize the information in the surrogate or auxiliary variable. Pepe (1992) proposed a semiparametric approach based on augmented likelihood, and Fleming et al. (1994) extended this approach to survival data and at the same time proposed an augmented score approach. Robins and Rotnitzky (1995) proposed using weighted estimating equations, in which the contribution of an observation in the validation sample to the estimating equations is inflated by the inverse of the probability of selection into the validation sample. In econometrics literature, Hahn (1998), Hirano, Imbens, and Ridder (2003) and others studied the treatment effect problem, where the probability of being included in a specific sample is estimated with a nonparametric estimator. Tarozzi (2004) considered a measurement error model.
with validation data using a similar method.

Chen, Leung, and Qin (2003) proposed using empirical likelihood for inference about \( \beta \) based on estimating equations from both the validation and the nonvalidation samples. Empirical likelihood is a nonparametric method of inference introduced by Owen (1988, 1990), which has properties analogous to that of parametric likelihood, including Wilks' theorem and Bartlett correction. Chen, Leung, and Qin (2003) show that using the nonvalidation sample leads to more precise estimation of \( \beta \), and the gain in efficiency increases with the size of the nonvalidation sample. Moreover, the method is more efficient when the information provided by the surrogate is highly "correlated" to the true endpoint.

The problem of selective transcriptional profiling can be treated as a multi-sample comparison problem when auxiliary data are available. Thus methods developed for surrogate endpoints can potentially be extended to the selective transcriptional profiling approach. The measurement of gene expression as obtained by microarray technology is the true endpoint that is part of the validation data in this setting, which are available for a subset of individuals within each genotypic group. The measurement on a trait of interest, such as crop yield or body fat, is the surrogate for gene expression or nonvalidation data. The question of interest is to test whether the mean of the expression level of a certain gene is different among genotypes. When the relatively inexpensive trait value is highly correlated with gene expression, one can potentially improve the efficiency of the test for association by using the extra trait data, which is achieved by selective transcriptional profiling approach as presented in Wang and Nettleton (2005) with a parametric model. In this paper, we extend empirical likelihood based inference for auxiliary data to the multi-sample comparison problem, then apply the general results to the selective transcriptional profiling approach. It is shown that as in the parametric model setting, using relatively inexpensive trait data on extra individuals can significantly increase the power in detecting genes whose expression level is associated with a QTL when using
empirical likelihood based inference. Alternatively one can achieve similar power using far fewer microarray chips. Like in other examples, the empirical likelihood method exhibits properties analogous to those based on parametric likelihood as discussed in Wang and Nettleton (2005).

4.2 Empirical Likelihood Based Multi-sample Testing with Auxiliary Data

In this section we demonstrate that the empirical likelihood ratio statistic for a multi-sample test with auxiliary data has a chi-square calibration. For the sake of simplicity in notation, we present the result in two-sample tests. Results for cases with more than two samples can be similarly derived with necessary modifications.

Let \( X_i = (X_{i1}, X_{i2}) \), \( i = 1, \ldots, n_x + m_x \) be independent observations from the same population. The values for \( X_{i1} \)'s are available for all \( n_x + m_x \) observations, but values for \( X_{i2} \)'s are only available for \( i = 1, \ldots, n_x \), i.e. the validation sample. Accordingly \( \{X_{(n_x+k),i}\}_{k=1}^{m_x} \) are the nonvalidation sample. We assume that \( n_x / (n_x + m_x) \to c_x \) for a constant \( c_x \in (0,1) \) as \( n_x \to \infty \). Suppose that the information for the distribution of \( X_i \) is summarized in an unknown \( p \)-dimensional parameter \( \beta_x \) via a \( p \)-dimensional estimating function \( U(X_i, \beta_x) \), with \( E\{U(X_i, \beta_x)\} = 0 \). Furthermore, the auxiliary information is summarized in a \( r \)-dimensional zero-mean estimating function \( g(X_{i1}, \gamma_x) \) with a \( r \)-dimensional unknown parameter \( \gamma_x \). Note here \( X_{i1} \) potentially includes both surrogate variables for \( X_{i2} \) and covariates that are always observable. Also we suppose that \( Y_j = (Y_{j1}, Y_{j2}), j = 1, \ldots, n_y + m_y \) are independent observations from another population. Here \( \{Y_j\}_{j=1}^{n_y} \) are the validation sample, and \( \{Y_{(n_y+t),i}\}_{t=1}^{m_y} \) are the nonvalidation sample. We define \( U(Y_j, \beta_y), g(Y_{j1}, \gamma_y), \) and \( c_y \) analogously with regard to \( U(X_i, \beta_x), g(X_{i1}, \gamma_x), \) and \( c_x \).

The interest here is to test \( H_0 : \beta_x = \beta_y \) v.s. \( H_a : \beta_x \neq \beta_y \). In the special case
where $\beta_x$ and $\beta_y$ are the means of $X_{i2}$ and $Y_{j2}$ respectively, the problem is to test whether the two populations have the same mean when auxiliary data are present. Parameters $\beta_x$ and $\beta_y$ can also be population quantiles, regression coefficients, or other parameters. To formulate empirical likelihood in this setting, denote $\psi \in \{x, y\}$ and $\Psi \in \{X, Y\}$, let $p_{\psi 1}, \ldots, p_{\psi n_\psi}$ be the nonnegative weights placed on the validation samples $\{X_i\}_{i=1}^{n_x}$ or $\{Y_j\}_{j=1}^{n_y}$. Also let $q_{\psi(n_\psi+1)}, \ldots, q_{\psi(n_\psi+m_\psi)}$ be the nonnegative weights placed on the nonvalidation samples $\{X((n_\psi+k)\ell)\}_{k=1}^{m_x}$ or $\{Y((n_\psi+l)\ell)\}_{l=1}^{m_y}$ respectively. The empirical likelihood for parameter vector $(\beta_x, \gamma_x, \beta_y, \gamma_y)$ is

$$L(\beta_x, \gamma_x, \beta_y, \gamma_y) = \max \prod_{i=1}^{n_x} p_{\psi i} \prod_{k=1}^{m_x} q_{\psi k} \prod_{j=1}^{n_y} p_{\psi j} \prod_{l=1}^{m_y} q_{\psi l}$$

subject to

$$\sum_{i=1}^{n_\psi} p_{\psi i} = 1,$$ (4.1)

$$\sum_{k=1}^{m_\psi} q_{\psi k} = 1,$$

$$\sum_{i=1}^{n_\psi} p_{\psi i} \left( U^\tau(\psi_i, \beta_\psi), g^\tau(\psi_i, \gamma_\psi) \right)^\tau = 0,$$

and

$$\sum_{k=1}^{m_\psi} q_{\psi k} g(\psi((n_\psi+k)\ell), \gamma_\psi) = 0.$$ We denote $U_{\psi i}(\beta_\psi) = U(\psi_i, \beta_\psi)$, $U_{\psi i}'(\beta_\psi) = \partial U_{\psi i}(\beta_\psi)/\partial \beta_\psi$, $g_{\psi i}(\gamma_\psi) = g(\psi_i, \gamma_\psi)$, and $g_{\psi i}'(\gamma_\psi) = \partial g_{\psi i}(\gamma_\psi)/\partial \gamma_\psi$. Using Lagrange multipliers as in Qin and Lawless (1994), the optimal weights for the parameter $(\beta_x, \gamma_x, \beta_y, \gamma_y)$ can be shown to be

$$p_{\psi i} = \frac{1}{n_\psi} \frac{1}{1 + t_{\psi i}^* U_{\psi i}(\beta_\psi) + t_{\psi i}' g_{\psi i}(\gamma_\psi)} \quad \text{for } i = 1, \ldots, n_\psi,$$ (4.2)

$$q_{\psi k} = \frac{1}{m_\psi} \frac{1}{1 + t_{\psi k}^* g_{\psi((n_\psi+k)\ell)}(\gamma_\psi)} \quad \text{for } k = 1, \ldots, m_\psi.$$ (4.3)
where the Lagrange multipliers \( t_{\psi}, \nu = 1, 2, 3 \), satisfy the following equations:

\[
\sum_{i=1}^{n_{\psi}} \frac{U_{\psi i}(\beta_{\psi})}{1 + t_{\psi}^{*i}U_{\psi i}(\beta_{\psi}) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi})} = 0, \quad (4.4)
\]

\[
\sum_{i=1}^{n_{\psi}} \frac{g_{\psi i}(\gamma_{\psi})}{1 + t_{\psi}^{*i}U_{\psi i}(\beta_{\psi}) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi})} = 0, \quad (4.5)
\]

\[
\sum_{k=1}^{m_{\psi}} \frac{g_{\psi(n_{\psi}+k)}(\gamma_{\psi})}{1 + t_{\psi 3}^{*}g_{\psi(n_{\psi}+k)}(\gamma_{\psi})} = 0. \quad (4.6)
\]

Using (4.2) and (4.3), we can write the negative log empirical likelihood as

\[
\ell(\beta_x, \gamma_x, \beta_y, \gamma_y) = -2 \log \{ L(\beta_x, \gamma_x, \beta_y, \gamma_y) \} = 2 \sum_{i=1}^{n_{\psi}} \sum_{i=1}^{n_{\psi}} \log \{ 1 + t_{\psi}^{*i}U_{\psi i}(\beta_{\psi}) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi}) \} + 2 \sum_{i=1}^{m_{\psi}} \sum_{k=1}^{m_{\psi}} \log \{ 1 + t_{\psi 3}^{*}g_{\psi(n_{\psi}+k)}(\gamma_{\psi}) \} + 2 \sum_{i=1}^{m_{\psi}} \{ n_{\psi} \log(n_{\psi}) + m_{\psi} \log(m_{\psi}) \}. \quad (4.7)
\]

When the null hypothesis is true, we denote

\[
\beta_x = \beta_y := \beta. \quad (4.8)
\]

Then differentiate (4.7) with regard to \( \beta, \gamma_x \) and \( \gamma_y \) and use the results (4.4)-(4.6) to get

\[
\sum_{\psi} \left\{ t_{\psi 1}^{*} \sum_{i=1}^{n_{\psi}} \frac{U_{\psi i}''(\beta)}{1 + t_{\psi 1}^{*}U_{\psi i}(\beta) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi})} \right\} = 0, \quad (4.9)
\]

\[
t_{\psi 2}^{*} \sum_{i=1}^{n_{\psi}} \frac{g_{\psi i}(\gamma_{\psi})}{1 + t_{\psi 1}^{*}U_{\psi i}(\beta) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi})} + t_{\psi 3}^{*} \sum_{k=1}^{m_{\psi}} \frac{g_{\psi(n_{\psi}+k)}(\gamma_{\psi})}{1 + t_{\psi 3}^{*}g_{\psi(n_{\psi}+k)}(\gamma_{\psi})} = 0. \quad (4.10)
\]

Let \((\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y, \tilde{\ell}_x, \tilde{\ell}_y, \tilde{\ell}_3, \tilde{\ell}_x, \tilde{\ell}_y, \tilde{\ell}_3, \tilde{\ell}_y)\) be the solutions to (4.4)-(4.6) and (4.8)-(4.10). Accordingly, we have the maximum empirical likelihood estimator of the parameter under \( H_0 \) as \((\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y)\). On the other hand, if we allow the estimators for \( \beta_x \) and \( \beta_y \) to take different values and replace (4.9) with

\[
t_{\psi 1}^{*} \sum_{i=1}^{n_{\psi}} \frac{U_{\psi i}''(\beta_{\psi})}{1 + t_{\psi 1}^{*}U_{\psi i}(\beta_{\psi}) + t_{\psi 2}^{*}g_{\psi i}(\gamma_{\psi})} = 0, \quad (4.11)
\]
we can obtain the maximum empirical likelihood estimator without constraint,

$$(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y),$$

with the corresponding Lagrange multiplier $(\ell_{x1}, \ell_{x2}, \ell_{x3}, \ell_{y1}, \ell_{y2}, \ell_{y3})$, by solving (4.4)-(4.6), (4.10) and (4.11). The log empirical likelihood ratio statistic for testing $H_0: \beta_x = \beta_y$ is

$$R_0 = \ell(\hat{\beta}, \hat{\gamma}_x, \hat{\gamma}_y) - \ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y).$$

The following theorem is a nonparametric version of the Wilks' theorem in parametric likelihood theory for the empirical likelihood ratio statistic.

**Theorem 1.** Under the conditions given in the Appendix and $\beta_x = \beta_y$, $R_0 \xrightarrow{d} \chi^2_\psi$ as $\min(n_x, m_y) \to \infty$, $\psi \in \{x, y\}$.

This theorem can be used to provide calibration for the empirical likelihood ratio statistic when testing $H_0: \beta_x = \beta_y$. Note that Wilks' theorem has been shown to hold for the empirical likelihood ratio statistic in a wide range of problems. Owen (1991) discussed empirical likelihood for ANOVA. Jing (1995) and Adimari (1995) considered the comparison of the mean for two populations. Qin and Lawless (1995) considered the case for parameters under constraints, and the multi-sample problem was discussed in Chapter 12 of Owen (2002). Our results and those in Chen, Leung, and Qin (2003) suggest that this property of empirical likelihood also applies to data with surrogate variables. While Chen et al. (2003) deals with one population case, we concentrate on multi-sample comparison problems, while also allowing missing at random with known observation probabilities as shown in Section 4.3.

To further simplify the notation, we may omit parameters $\beta_\psi$ and $\gamma_\psi$ in $U_\psi(\beta_\psi)$ and $g_\psi(\gamma_\psi)$ respectively. Also define $V(U_\psi) = E(U_\psi U_\psi^\top)$ and $V(g_\psi) = E(g_\psi g_\psi^\top)$.

Consider the local alternative $\beta_y = \beta_x + n_\psi^{-1/2} u$, where $0 < \|u\| < \infty$. Here we suppose that $n_x$ and $n_y$ are of the same order. It can be shown that the asymptotic
power for the test defined in Theorem 1 is \( P(\chi^2_\alpha(\lambda) > \chi^2_{p,1-\alpha}) \), where \( \alpha \) is the size of the test and \( \lambda \) is the noncentrality parameter of the noncentral chi-square distribution with value

\[
\lambda = u^T \left[ \sum_{\psi} E^{-1}(U_{\psi}^\top) \{ V(U_{\psi}) - (1 - c_{\psi}) E(U_{\psi} g_{\psi}^\top) V^{-1}(g_{\psi}) E(g_{\psi} U_{\psi}^\top) \} E^{-1}(U_{\psi}^\top) \right]^{-1} u.
\]

When the nonvalidation data are not utilized and the inference is based only on the validation data, the asymptotic power function becomes \( P(\chi^2_p(\lambda) > \chi^2_{p,1-\alpha}) \), where the noncentrality parameter of the noncentral chi-square distribution is

\[
\lambda = u^T \left[ \sum_{\psi} E^{-1}(U_{\psi}^\top) \{ V(U_{\psi}) \} E^{-1}(U_{\psi}^\top) \right]^{-1} u.
\]

Since \( \lambda > \lambda \), using nonvalidation data improves the power of the multi-sample test. The improvement in power is higher when \( c_{\psi} \) is smaller, that is, when the nonvalidation sample is large. Since the term \( E(U_{\psi} g_{\psi}^\top) V^{-1}(g_{\psi}) E(g_{\psi} U_{\psi}^\top) \) can be seen as a measure of correlation between estimating functions \( U \) and \( g \), it also suggests that the power is higher when the information contained in \( g \) is highly correlated to that contained in \( U \).

A simple variation of Theorem 1 can be used to test \( H_0: \beta_x = \beta_y = \beta_0 \), whether the parameters for two populations have the same specific value. The empirical likelihood ratio statistic for this case is

\[
R(\beta_0) = \ell(\beta_0, \tilde{\gamma}_x, \tilde{\gamma}_y) - \ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y),
\]

which converges to \( \chi^2_{2p} \). The test statistic for more than two populations can also be derived in a straightforward fashion.

### 4.3 Extension to Missing at Random with Known Probability of Observation

In the previous section, we assume that the validation sample and the corresponding nonvalidation sample are both random samples from the same population, or \( X_{i2} \) and \( Y_{i2} \).
in the nonvalidation samples are missing completely at random (MCAR). In practice, one might select objects to measure \( X_{i2} \) or \( Y_{i2} \) using some criteria based on \( X_{i1} \) or \( Y_{i1} \). For example, in Wang and Nettleton (2005), transcriptional abundance of genes measured by microarrays might be obtained using individuals with most extreme trait values within each genotypic group. In these settings, the values of \( X_{i2} \) and \( Y_{i2} \) in the nonvalidation samples are missing at random (MAR), but not MCAR. In this section, we extend Theorem 1 to cover situations where the probability of observation is known.

The case that the probability of observation needs to be estimated will be discussed in the Conclusion and Discussion.

Let \( \delta_{\psi i} = 1 \) if \( \Psi_{i2} \) is observed, i.e., \( \Psi_i \) is in the validation sample, and \( \delta_{\psi i} = 0 \) if \( \Psi_i \) is in the nonvalidation sample, in which case we only know the value of \( \Psi_{i1} \). Let \( \Psi_1 \) denote \( \{\Psi_{i1}\}_{i=1}^{m_{\psi}+m_{\overline{\psi}}} \). We define the probability that the \( i \)th individual is selected for the validation sample given \( \Psi_1 \) as

\[
\pi_i(\Psi_1) = P(\delta_{\psi i} = 1 | \Psi_1).
\]

Accordingly, \( 1 - \pi_i(\Psi_1) \) is the probability that \( \Psi_i \) is included in the nonvalidation sample. We also assume that \( \delta_{\psi i} \) and \( \Psi_{i2} \) are conditionally independent given \( \Psi_1 \), i.e.

\[
\delta_{\psi i} \perp \Psi_{i2} \mid \Psi_1
\]

(4.12)

and \( 0 < \pi_i(\Psi_1) < 1 \). Note that if \( \delta_i \) does not depend on observations on other units, \( \pi_i(\Psi_1) \) becomes the response propensity score \( p(\psi_{i1}) := P(\delta_{\psi i} = 1 | \psi_{i1}) \), and (4.12) reduces to the strongly ignorable missing at random condition in Rosenbaum and Rubin (1983). We use the weaker condition here to allow selection based on all \( \psi_{i1} \) values.

In the weighted estimating equation method (Robins and Rotnitzky, 1995) as well as the treatment effect problems considered by Hahn (1998); Hirano, Imbens, and Ridder (2003) and others, unbiased estimates are obtained by inflating the estimating equations for complete observations with the inverse of the propensity score. Similar techniques can be used here for validation and nonvalidation samples.
Denote \( r_\psi = P(\delta_{\psi i} = 1) \), and define the following quantities as estimating functions adjusted with the probability of being included in a specific sample:

\[
\tilde{U}_{\psi i}(\beta_\psi) = \frac{r_\psi U_{\psi i}(\beta_\psi)}{\pi_i(\Psi_1)} \quad \text{for} \ i = 1, \ldots, n_\psi,
\]

\[
\tilde{g}_{\psi i}(\gamma_\psi) = \frac{r_\psi g_{\psi i}(\gamma_\psi)}{\pi_i(\Psi_1)} \quad \text{for} \ i = 1, \ldots, n_\psi,
\]

\[
\tilde{g}_{\psi (n_\psi + k)}(\gamma_\psi) = \frac{(1 - r_\psi)g_{\psi (n_\psi + k)}(\gamma_\psi)}{1 - \pi_i(\Psi_1)} \quad \text{for} \ k = 1, \ldots, m_\psi.
\]

Note that by (4.12), we have, using iterated integration,

\[
E\{\tilde{U}_{\psi i}(\beta_\psi) | \delta_{\psi i} = 1\} = E\{U_{\psi i}(\beta_\psi)\} = 0. \quad (4.13)
\]

Similarly we can show

\[
E\{\tilde{g}_{\psi i}(\gamma_\psi) | \delta_{\psi i} = 1\} = 0, \quad \text{and} \quad E\{\tilde{g}_{\psi (n_\psi + k)}(\gamma_\psi) | \delta_{\psi (n_\psi + k)} = 0\} = 0.
\]

Thus \( \tilde{U}_{\psi i}(\beta_\psi), \tilde{g}_{\psi i}(\gamma_\psi) \) and \( \tilde{g}_{\psi (n_\psi + k)}(\gamma_\psi) \) are also unbiased estimating functions conditioning on the sample assignment. By using \( \tilde{U}_{\psi i}(\beta_\psi), \tilde{g}_{\psi i}(\gamma_\psi) \) and \( \tilde{g}_{\psi (n_\psi + k)}(\gamma_\psi) \) to replace \( U_{\psi i}(\beta_\psi), g_{\psi i}(\gamma_\psi) \) and \( g_{\psi (n_\psi + k)}(\gamma_\psi) \) in (4.4)-(4.6) and (4.8)-(4.10) respectively, we can derive the empirical likelihood ratio statistic for testing \( H_0 \) under missing at random but with known probability of observation. The chi-square calibration can also be shown to hold in this case.

### 4.4 Applications to Selective Transcriptional Profiling

First we consider the case in which the QTL only has two genotypes, which is the case for back cross, doubled haploid or recombinant inbred lines. Using the notation of Section 4.2, in the selective transcriptional profiling problem, \( X_{i1} \) and \( Y_{i1} \) are the values of a traditional trait for the \( i \)th individual in genotypic group \( x \) and \( y \) respectively. Variables \( X_{i2} \) and \( Y_{i2} \) are the expression levels of a gene of interest measured by microarrays for the corresponding individual. The problem of testing for association between the expression
level of a certain gene and the QTL is equivalent to testing whether $X_{i2}$ and $Y_{i2}$ have the same mean. Thus, $\beta_x$ and $\beta_y$ are the means of expression levels for individuals in the two genotypic groups respectively, $\gamma_x$ and $\gamma_y$ are the corresponding means of the traditional trait known to be associated with the QTL. The estimating functions are

$$U_{x1}(\beta_x) = X_{i2} - \beta_x, \quad g_{x1}(\gamma_x) = X_{i1} - \gamma_x,$$

$$U_{y1}(\beta_y) = Y_{j2} - \beta_y, \quad g_{y1}(\gamma_y) = Y_{j1} - \gamma_y.$$

The empirical likelihood based test for association between transcriptional abundance and QTL genotype involves deriving the maximum empirical likelihood estimator

$$(\hat{\beta}, \hat{\gamma}_x, \hat{\gamma}_y)^T$$

under $H_0: \beta_x = \beta_y := \beta$, and the maximum empirical likelihood estimator $(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y)^T$ with no constraint on $\beta_x$ and $\beta_y$. Theorem 1 shows that the empirical likelihood ratio statistic can be compared to $\chi^2_1$ distribution to obtain a p-value.

Define the covariance matrix of the trait value and the expression level for each individual as

$$Var(\Psi_i) = \begin{bmatrix} \sigma_{x11} & \sigma_{x12} \\ \sigma_{x12} & \sigma_{x22} \end{bmatrix}.$$ 

Recall the local alternative discussed in Section 4.2, suppose $\beta_y - \beta_x = n^{-1/2}_\psi u$ with $0 < u < \infty$. The asymptotic power of the test using extra trait data is given by $P(\chi^2_1(\lambda) > \chi^2_{1,1-a})$, where the noncentrality parameter for the noncentral chi-square distribution is

$$\lambda = \frac{u^2}{\sum_{\psi} \left\{ \sigma_{\psi 22} - (1 - c_\psi) \sigma_{x12}^2 / \sigma_{x11} \right\}}.$$ 

When the nonvalidation data (extra trait data) are not used and the empirical likelihood is based only on individuals with expression data, the noncentrality parameter in the asymptotic power function becomes

$$\tilde{\lambda} = \frac{u^2}{\sum_{\psi} \sigma_{\psi 22}}.$$
Thus, we achieve better power for the test of association between transcriptional abundance and the QTL of interest by including the extra trait data (nonvalidation data). This is also the case in Wang and Nettleton (2005) when bivariate normal distribution is assumed for the expression abundance and trait value.

Wang and Nettleton (2005) also considered the case of selecting individuals with most extreme trait values in each genotypic group for microarray experiments. The results in Section 4.3 assume that $0 < \pi_i(\Psi_1) < 1$, which implies that any individual can possibly be selected or excluded for microarray experiments. So the method of Wang and Nettleton (2005) of selecting equal number of individuals with the highest and lowest trait values for microarrays cannot be applied directly, since it necessitates that $\pi_i(\Psi_1)$ be either 0 or 1. However, it is possible to preferentially select individuals with more extreme trait values for microarrays and use the result in Section 4.3.

For cases with more than two genotypes, such as in F2 populations, results analogous to Theorem 1 can be obtained in a straight forward fashion. The general idea of using extra trait values to enhance the power of test still applies.

The performance of the empirical likelihood based method for finite sample size is studied with simulation presented in Section 4.5.

### 4.5 Simulation Study

First we suppose that the expression level of a certain gene and the trait value known to be associated with a QTL have a bivariate normal distribution as in Wang and Nettleton (2005). There are 100 individuals with QTL genotype $x$ and $y$ respectively. The trait means are $\gamma_x = 0$ and $\gamma_y = 1$ for genotypes $x$ and $y$. Suppose $\sigma_{\psi11} = \sigma_{\psi22} = 1$ and $\sigma_{\psi12} = .85$ for both genotypes. In this notation $\sigma_{\psi11}$ is the variance for the trait value for individuals in genotypic group $\psi$, and $\sigma_{\psi22}$ is the variance of expression abundance for genotype $\psi$ individuals. Here we suppose that the two genotypic groups share the
same covariance matrix. The trait value $\theta_{i1}$ is supposed to be known for all individuals, while the expression level as measured by microarray experiments are available on 30 randomly selected individuals in each genotype. By simulation results not shown here, the value of $\gamma_p$ does not appear to affect the performance of the test, while the test is more powerful if $\sigma_{\theta 11}$ or $\sigma_{\theta 22}$ is smaller.

Here we examine the performance of the empirical likelihood based test with auxiliary data when compared to the empirical likelihood based test using only the expression data from 30 individuals in each group. We also present results obtained using the Wald test as proposed in Wang and Nettleton (2005). Note that in this case all the parametric assumptions for the Wald test are satisfied. For a given difference in means $\beta_y - \beta_x$, ranging from 0 to 1 unit in increments of 0.25 units, 1,000 independent samples were generated. The results are summarized in Table 4.1.

It can be seen from Table 4.1 that the Wald test is the most powerful of these three tests, which is expected as the parametric model assumptions for the Wald test are satisfied. But it is notable that the empirical likelihood based test with a surrogate variable (trait value) can achieve power that is only slightly lower than that of the Wald test under conditions most favorable to the parametric method. It is also obvious that using the trait value in empirical likelihood inference leads to far superior performance than that of empirical likelihood test for two samples using only the expression data. This confirms that the principal of using extra trait data to improve power also applies to our nonparametric method. The result for the size 0.05 test is also shown in Figure 4.1.

Obviously, the interest in using the empirical likelihood method as opposed to using fully parametric models is that empirical likelihood does not require specific distributions. To explore conditions where the parametric assumptions in Wang and Nettleton (2005) no longer hold, we perform simulations for two such cases with results summarized in Tables 4.2 and 4.3.
Table 4.1 The type I error rate and power of the Wald test, the empirical likelihood based test with a surrogate variable (trait), and the empirical likelihood based test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class. Within each genotypic group, expression level and trait are assumed to have a bivariate normal distribution. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

<table>
<thead>
<tr>
<th>$\beta_y - \beta_x$</th>
<th>Wald</th>
<th>EL (with trait)</th>
<th>EL (expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.005</td>
<td>.008</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>.048</td>
<td>.047</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>.090</td>
<td>.099</td>
<td>.111</td>
</tr>
<tr>
<td>.25</td>
<td>.121</td>
<td>.122</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>.289</td>
<td>.280</td>
<td>.182</td>
</tr>
<tr>
<td></td>
<td>.392</td>
<td>.390</td>
<td>.258</td>
</tr>
<tr>
<td>.50</td>
<td>.592</td>
<td>.590</td>
<td>.283</td>
</tr>
<tr>
<td></td>
<td>.820</td>
<td>.809</td>
<td>.549</td>
</tr>
<tr>
<td></td>
<td>.884</td>
<td>.878</td>
<td>.669</td>
</tr>
<tr>
<td>.75</td>
<td>.944</td>
<td>.938</td>
<td>.643</td>
</tr>
<tr>
<td></td>
<td>.986</td>
<td>.983</td>
<td>.839</td>
</tr>
<tr>
<td></td>
<td>.996</td>
<td>.995</td>
<td>.913</td>
</tr>
<tr>
<td>1.00</td>
<td>.997</td>
<td>.995</td>
<td>.881</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>.984</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>.984</td>
</tr>
</tbody>
</table>
Figure 4.1 The power of the Wald test, the empirical likelihood based test using a surrogate (trait), and the empirical likelihood based test using only expression data. The expression level for individuals with genotype $x$ was fixed at 0. The mean for individuals with genotype $y$ varies from 0 to 1.25 in increments of 0.25. The size of the test is .05. These results are also reported in Table 4.1.
Table 4.2  The type I error rate and power of the Wald test, the empirical likelihood based test with a surrogate variable (trait), and the empirical likelihood based test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class. Within each genotypic group, expression level and trait are assumed to have a bivariate normal distribution, but the covariance matrices are different between genotypes. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

<table>
<thead>
<tr>
<th>$\beta_y - \beta_x$</th>
<th>Wald</th>
<th>EL (with trait)</th>
<th>EL (expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.013</td>
<td>.012</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>.053</td>
<td>.059</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>.111</td>
<td>.115</td>
<td>.113</td>
</tr>
<tr>
<td>.25</td>
<td>.045</td>
<td>.045</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>.130</td>
<td>.139</td>
<td>.112</td>
</tr>
<tr>
<td></td>
<td>.219</td>
<td>.228</td>
<td>.205</td>
</tr>
<tr>
<td>.50</td>
<td>.208</td>
<td>.228</td>
<td>.164</td>
</tr>
<tr>
<td></td>
<td>.411</td>
<td>.432</td>
<td>.381</td>
</tr>
<tr>
<td></td>
<td>.533</td>
<td>.544</td>
<td>.509</td>
</tr>
<tr>
<td>.75</td>
<td>.518</td>
<td>.529</td>
<td>.430</td>
</tr>
<tr>
<td></td>
<td>.720</td>
<td>.738</td>
<td>.652</td>
</tr>
<tr>
<td></td>
<td>.808</td>
<td>.821</td>
<td>.765</td>
</tr>
<tr>
<td>1.00</td>
<td>.789</td>
<td>.794</td>
<td>.720</td>
</tr>
<tr>
<td></td>
<td>.924</td>
<td>.933</td>
<td>.880</td>
</tr>
<tr>
<td></td>
<td>.958</td>
<td>.968</td>
<td>.936</td>
</tr>
</tbody>
</table>
Table 4.3 The type I error rate and power of the Wald test, the empirical likelihood based test with a surrogate variable (trait), and the empirical likelihood based test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class. For genotype $y$ individuals, the trait and expression have a bivariate normal distribution; while for genotype $x$ individuals, a skewed bivariate t-distribution is assumed. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

<table>
<thead>
<tr>
<th>$\beta_y - \beta_x$</th>
<th>Wald (with trait)</th>
<th>EL (trait)</th>
<th>EL (expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.018</td>
<td>.013</td>
<td>.015</td>
</tr>
<tr>
<td>0.069</td>
<td>.059</td>
<td>.059</td>
<td></td>
</tr>
<tr>
<td>0.121</td>
<td>.109</td>
<td>.124</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.066</td>
<td>.075</td>
<td>.043</td>
</tr>
<tr>
<td>.156</td>
<td>.198</td>
<td>.127</td>
<td></td>
</tr>
<tr>
<td>.219</td>
<td>.276</td>
<td>.199</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>.218</td>
<td>.307</td>
<td>.157</td>
</tr>
<tr>
<td>.383</td>
<td>.551</td>
<td>.321</td>
<td></td>
</tr>
<tr>
<td>.491</td>
<td>.677</td>
<td>.450</td>
<td></td>
</tr>
<tr>
<td>.75</td>
<td>.447</td>
<td>.703</td>
<td>.375</td>
</tr>
<tr>
<td>.673</td>
<td>.877</td>
<td>.622</td>
<td></td>
</tr>
<tr>
<td>.786</td>
<td>.931</td>
<td>.744</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>.733</td>
<td>.925</td>
<td>.652</td>
</tr>
<tr>
<td>.886</td>
<td>.984</td>
<td>.855</td>
<td></td>
</tr>
<tr>
<td>.924</td>
<td>.993</td>
<td>.907</td>
<td></td>
</tr>
</tbody>
</table>
For the results in Table 4.2, the trait value and expression level have a bivariate normal distribution for both genotypic groups, but with different covariance matrices. The distribution of trait and expression for genotype $y$ individuals are the same as in the simulation for Table 4.1. For genotypic group $x$, however, the trait and expression have a bivariate normal distribution with mean $(0, 0)$, $\sigma_{x11} = .5$, $\sigma_{x22} = 2$, and $\sigma_{x12} = .01$. For the results in Table 4.3, the distribution for the trait and expression values for genotype $y$ individuals is the same as before, but for individuals in genotypic group $x$, trait and expression values have a skewed bivariate t-distribution (Azzalini and Capitanio, 2003) with shape parameters $\alpha = (4, 1)$, degrees of freedom of five, dispersion matrix

$$
\begin{bmatrix}
1.00 & -1.30 \\
-1.30 & 2.25
\end{bmatrix}
$$

and mean $(0, 0)$. From Table 4.2 and Table 4.3, it can be seen that all three tests have satisfactory Type I errors, but the Wald test is less powerful than the empirical likelihood test with auxiliary information. Thus the empirical likelihood based test using extra trait information can outperform the Wald test when the parametric assumptions for the Wald test are violated. Using extra trait data still improves the power relative to tests using expression values alone. The results in Table 4.3 are also shown in Figure 4.2 for test size .05.

As discussed in Section 4.4, we could preferentially select individuals for microarray experiments based on the value of the trait. The result in Section 4.3 require that $0 < \pi_i(\Psi_1) < 1$, i.e., it is possible for any individual to be selected or excluded for microarray experiment. For the simulation results summarized in Table 4.4, in each genotypic group, ten individuals are randomly selected from the 15 individuals with the highest or the lowest trait value respectively, another ten individuals come from the 70 individuals in the middle. The empirical likelihood inference is carried out as described in Section 4.3, with the estimating function inflated by the inverse of the probability of
Figure 4.2 The power of the Wald test, the empirical likelihood based test using a surrogate (trait), and the empirical likelihood ratio test using only expression data for the simulation shown in Table 4.3. The size of the test is .05. The expression level for individuals with genotype $x$ is fixed at 0. The mean for individuals with genotype $y$ varies from 0 to 1.25 in increments of 0.25.
Table 4.4 The type I error rate and power of the Wald test, the empirical likelihood based test with a surrogate variable (trait), and the empirical likelihood based test using only expression values, when the individuals are preferentially selected for microarrays. Individuals with more extreme trait values are given preference for microarray experiments. The trait and expression have a bivariate normal distribution as in the case of Table 4.1. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

<table>
<thead>
<tr>
<th>$\beta_y - \beta_x$</th>
<th>Wald</th>
<th>EL (with trait)</th>
<th>EL (expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.014</td>
<td>.016</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>.053</td>
<td>.058</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>.107</td>
<td>.110</td>
<td>.102</td>
</tr>
<tr>
<td>.25</td>
<td>.118</td>
<td>.095</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td>.265</td>
<td>.217</td>
<td>.170</td>
</tr>
<tr>
<td></td>
<td>.379</td>
<td>.297</td>
<td>.255</td>
</tr>
<tr>
<td>.50</td>
<td>.600</td>
<td>.416</td>
<td>.228</td>
</tr>
<tr>
<td></td>
<td>.795</td>
<td>.640</td>
<td>.418</td>
</tr>
<tr>
<td></td>
<td>.859</td>
<td>.745</td>
<td>.550</td>
</tr>
<tr>
<td>.75</td>
<td>.941</td>
<td>.804</td>
<td>.501</td>
</tr>
<tr>
<td></td>
<td>.988</td>
<td>.934</td>
<td>.728</td>
</tr>
<tr>
<td></td>
<td>.992</td>
<td>.959</td>
<td>.822</td>
</tr>
<tr>
<td>1.00</td>
<td>1.000</td>
<td>.978</td>
<td>.796</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>.992</td>
<td>.930</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>.998</td>
<td>.958</td>
</tr>
</tbody>
</table>

As in cases discussed before, empirical likelihood based inference using extra trait information has more power than that based solely on expression. But when compared to Table 4.1, EL test with extra trait information in this case has less power than when individuals are selected completely at random, as the weighting by the inverse of the selection probability introduced more variance. Thus selection completely at random is preferred in this case.
4.6 Conclusion and Discussion

Selective transcriptional profiling as proposed by Wang and Nettleton (2005) can significantly reduce the cost associated with a large number of microarray chips in genetical genomics study. However, there will be occasions when the bivariate normal distribution and equal covariance assumption in Wang and Nettleton (2005) might be problematic. In this article, we develop empirical likelihood based inference for multi-sample comparison with auxiliary data and apply it in the selective transcriptional profiling approach setting. It is shown that the idea of using extra relatively inexpensive trait data to improve the power in testing the association between a known QTL and transcriptional abundance also applies to nonparametric inference with empirical likelihood. When the model assumptions in Wang and Nettleton (2005) are satisfied, the empirical likelihood method using auxiliary data is only slightly less powerful than the Wald test, and can be more powerful when the model assumptions are violated. On the other hand, the Wald test often gives satisfactory Type I error rates even when the model assumptions do not hold exactly. Thus for ease of computation and explanation, the Wald test should be preferred if data are nearly normal, and the variance structures for the two genotypes are reasonably close. The empirical likelihood method utilizing extra trait data should be used if the data suggest substantial departures from the assumptions used by Wang and Nettleton (2005).

There is recent interest in applying empirical likelihood methods to genetic mapping problems. Zou, Fine, and Yandell (2002) and Zou and Fine (2002) developed a partial empirical likelihood method for QTL mapping. It will be of interest to see if their method can be combined with the selective transcriptional profiling approach when the location of the QTL is not certain.

The theoretical results of Section 4.2 are not limited to the mean functional, though we considered only the mean case for selective transcriptional profiling. Similar methods
can be employed to compare quantiles, regression coefficients, or other parameters of interest between multiple samples when surrogate variables are available. Section 4.3 further extends the results to missing at random conditions when the probability of observation is known. In missing data settings where the propensity score, \( p(\Psi_{\psi_i}) \), is unknown, one natural approach is to estimate \( p(\Psi_{\psi_i}) \) with a parametric method similar to Rotnitzky and Robins (1995), or with a nonparametric method similar to Hahn (1988). It is also shown that estimating the propensity score can in fact improve the efficiency. The drawback is that when an estimator for \( p(\Psi_{\psi_i}) \) is used in the estimating functions, we can show that the chi-square calibration for the empirical likelihood ratio no longer holds. Instead, the empirical likelihood ratio converges to a linear combination of chi-squares. This property is similar to that for the nonparametric imputation method studied in Wang and Chen (2006). Wang and Chen (2006) developed a bootstrap calibration for empirical likelihood ratio which converges to a linear combination of chi-squares. It will be interesting to derive a similar approach for the multi-sample comparison problem.

### 4.7 References


4.8 Appendix

Suppose that \((\beta_{x0}, \gamma_{x0}, \beta_{y0}, \gamma_{y0})\) are the true values of the parameters. Recall that we use \(\psi \in \{x, y\}\) and \(\Psi \in \{X, Y\}\). The following conditions are needed for Theorem 1.

C1: Both \(V_\psi(U) = E\{U(\beta_{\psi 0})U^T(\beta_{\psi 0})\}\) and \(V_\psi(g) = E\{g(\gamma_{\psi 0})g^T(\gamma_{\psi 0})\}\) are positive definite, and the ranks of \(E(\frac{\partial V_\psi}{\partial \beta_\psi})\) and \(E(\frac{\partial V_\psi}{\partial \gamma_\psi})\) are \(p\) and \(r\) respectively.

C2: \(\frac{\partial^2 V_\psi(\beta_\psi)}{\partial \beta_\psi \partial \beta_\psi^T}\) is continuous in a neighborhood of \(\beta_{\psi 0}\), and both \(\|\frac{\partial V_\psi(\beta_\psi)}{\partial \beta_\psi}\|\) and \(\|U_\psi(\beta_\psi)\|^3\) are bounded in this neighborhood.

C3: \(\frac{\partial^2 V_\psi(\gamma_\psi)}{\partial \gamma_\psi \partial \gamma_\psi^T}\) is continuous in a neighborhood of \(\gamma_{\psi 0}\), and both \(\|\frac{\partial V_\psi(\gamma_\psi)}{\partial \gamma_\psi}\|\) and \(\|g_\psi(\gamma_\psi)\|^3\) are bounded in this neighborhood.

C4: \(n_\psi\) and \(m_\psi\) \(\to\) \(\infty\), and \(n_\psi/(n_\psi + m_\psi)\) \(\to\) \(c_\psi\) \(\in\) \((0, 1)\) as \(\min(n_\psi, m_\psi) \to \infty\). \(n_x\) and \(n_y\) are of the same order.

Proof of Theorem 1: We first derive the expansion of \(\ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y)\), in which the values of \(\hat{\beta}_x\) and \(\hat{\beta}_y\) can change freely with regard to each other. By similar derivation to that of Qin and Lawless (1994), we can show that \(\hat{t}_\psi = O_p(n_\psi^{-1/3})\) for \(\nu = 1, 2, \text{ and } 3\). Using Taylor expansion on the left sides of (4.4)-(4.6), (4.10), and (4.11) around \((\beta_{x0}, \gamma_{x0}, \beta_{y0}, \gamma_{y0}, \mathbf{0}_6)\), where \(\mathbf{0}_6\) denotes the zero vector of length six, and ignoring the terms of \(O_p(n_\psi^{-1/3})\),

\[
\sum_{i=1}^{n_\psi} U_{\psi i}(\beta_{\psi 0}) \hat{t}_{\psi 1} = 0, \tag{4.14}
\]

\[
\sum_{i=1}^{n_\psi} \gamma_{\psi i}(\gamma_{\psi 0}) \hat{t}_{\psi 2} + \sum_{k=1}^{m_\psi} \gamma_{\psi(n_\psi+k)}(\gamma_{\psi 0}) \hat{t}_{\psi 3} = 0, \tag{4.15}
\]

\[
\sum_{i=1}^{n_\psi} U_{\psi i}(\beta_{\psi 0})(\hat{\beta}_\psi - \beta_{\psi 0}) - \sum_{i=1}^{n_\psi} U_{\psi i}(\beta_{\psi 0})U_{\psi i}(\beta_{\psi 0}) \hat{t}_{\psi 1} + \sum_{i=1}^{n_\psi} U_{\psi i}(\beta_{\psi 0})g_{\psi i}(\gamma_{\psi 0}) \hat{t}_{\psi 2} + \sum_{i=1}^{n_\psi} U_{\psi i}(\beta_{\psi 0}), \tag{4.16}
\]
\[ \sum_{i=1}^{n_\varphi} g_{\psi_i}(\gamma_{\varphi_0})(\bar{\gamma}_{\varphi} - \gamma_{\varphi_0}) - \sum_{i=1}^{n_\varphi} g_{\psi_i}(\gamma_{\varphi_0})U^\varphi_{\psi_i}(\beta_{\varphi_0})\bar{t}_{\psi_1} \]

\[ - \sum_{i=1}^{n_\varphi} g_{\psi_i}(\gamma_{\varphi_0})g^\varphi_{\psi_i}(\gamma_{\varphi_0})\bar{t}_{\psi_2} = - \sum_{i=1}^{n_\varphi} g_{\psi_i}(\gamma_{\varphi_0}), \quad (4.17) \]

\[ \sum_{k=1}^{m_\psi} g'_{\psi(n_\varphi + k)}(\gamma_{\varphi_0})(\bar{\gamma}_{\varphi} - \gamma_{\varphi_0}) - \sum_{k=1}^{m_\psi} g_{\psi(n_\varphi + k)}(\gamma_{\varphi_0})g^\varphi_{\psi(n_\varphi + k)}(\gamma_{\varphi_0})\bar{t}_{\psi_3} \]

\[ = - \sum_{k=1}^{m_\psi} g_{\psi(n_\varphi + k)}(\gamma_{\varphi_0}). \quad (4.18) \]

Let

\[ s_{12} = \begin{bmatrix} s_{12x} \\ s_{12y} \end{bmatrix}, \quad s_{22} = \begin{bmatrix} s_{22x} \\ s_{22y} \end{bmatrix}, \quad \text{where} \]

\[
s_{12\psi} = \begin{bmatrix} n_\psi^{-1} \sum_{i=1}^{n_\psi} U^\varphi_{\psi_i}(\beta_{\varphi_0}) & 0 & 0 \\ 0 & n_\psi^{-1} \sum_{i=1}^{n_\psi} g^\varphi_{\psi_i}(\gamma_{\varphi_0}) & n_\psi^{-1} \sum_{k=1}^{m_\psi} g_{\psi(n_\varphi + k)}(\gamma_{\varphi_0}) \end{bmatrix},
\]

\[
s_{22\psi} = \begin{bmatrix} \frac{1}{n_\psi} \sum_{i=1}^{n_\psi} U^\varphi_{\psi_i}(\beta_{\varphi_0})U^\psi_{\psi_i}(\beta_{\varphi_0}) & \frac{1}{n_\psi} \sum_{i=1}^{n_\psi} U^\varphi_{\psi_i}(\beta_{\varphi_0})g^\varphi_{\psi_i}(\gamma_{\varphi_0}) & 0 \\ \frac{1}{n_\psi} \sum_{i=1}^{n_\psi} g_{\psi_i}(\gamma_{\varphi_0})U^\psi_{\psi_i}(\beta_{\varphi_0}) & \frac{1}{n_\psi} \sum_{i=1}^{n_\psi} g_{\psi_i}(\gamma_{\varphi_0})g^\varphi_{\psi_i}(\gamma_{\varphi_0}) & 0 \\ 0 & 0 & s_{22\psi,3} \end{bmatrix},
\]

\[
s_{22\psi,3} = \frac{1}{n_\psi} \sum_{k=1}^{m_\psi} g_{\psi(n_\varphi + k)}(\gamma_{\varphi_0})g^\varphi_{\psi(n_\varphi + k)}(\gamma_{\varphi_0}),
\]

and

\[ S_n = \begin{bmatrix} 0 & s_{12} \\ s_{12}^\top & -s_{22} \end{bmatrix}. \]

Denote \( t_\varphi = (t_{\psi_1}, t_{\psi_2}, t_{\psi_3})^\top \). Equations (14)-(18) imply that

\[ S_n \left( (\beta_x - \beta_{x_0})^\top, (\beta_y - \beta_{y_0})^\top, (\beta_z - \beta_{z_0})^\top, (\gamma_x - \gamma_{x_0})^\top, (\gamma_y - \gamma_{y_0})^\top, (\gamma_z - \gamma_{z_0})^\top, \bar{t}_x, \bar{t}_y \right)^\top = -Q_n, \]

where

\[ Q_n = (0, 0, 0, 0, T^\varphi_{nx}, T^\varphi_{ny})^\top, \]

\[ T^\varphi_n = \left( \frac{n_\psi^{-1}}{n_\psi} \sum_{i=1}^{n_\psi} U^\varphi_{\psi_i}(\beta_{\varphi_0}), \frac{n_\psi^{-1}}{n_\psi} \sum_{i=1}^{n_\psi} g^\varphi_{\psi_i}(\gamma_{\varphi_0}), \frac{n_\psi^{-1}}{n_\psi} \sum_{k=1}^{m_\psi} g^\varphi_{\psi(n_\varphi + k)}(\gamma_{\varphi_0}) \right)^\top. \]
By the manipulation of block matrices,
\[
S_{n}^{-1} = \begin{bmatrix}
\left(s_{12}s_{22}^{-1}s_{12}^{-1}\right) & \left(s_{12}s_{22}^{-1}s_{12}^{-1}s_{12}s_{22}^{-1}\right) \\
\left(s_{22}^{-1}s_{12}s_{22}^{-1}s_{12}^{-1}\right) & -s_{22}^{-1} + s_{22}^{-1}s_{12}s_{22}^{-1}s_{12}^{-1}s_{12}s_{22}^{-1}
\end{bmatrix}.
\]

Also define the following quantities,
\[
\Sigma_{\mathbf{12}} = \begin{bmatrix}
\Sigma_{12x} \\
\Sigma_{12y}
\end{bmatrix}, \quad \Sigma_{\mathbf{22}} = \begin{bmatrix}
\Sigma_{22x} \\
\Sigma_{22y}
\end{bmatrix}, \quad \text{where}
\]
\[
\Sigma_{\mathbf{12}} = \begin{bmatrix}
E U_{\psi_{1}}(\beta_{\psi_{0}}) & 0 & 0 \\
0 & E g_{\psi_{1}}(\gamma_{\psi_{0}}) & \frac{1-c_{\psi}}{c_{\psi}} E g_{\psi_{1}}(\gamma_{\psi_{0}})g_{\psi_{1}}(\gamma_{\psi_{0}})
\end{bmatrix},
\]
\[
\Sigma_{\mathbf{22}} = \begin{bmatrix}
E\{U_{\psi_{1}}(\beta_{\psi_{0}})U_{\psi_{1}}(\beta_{\psi_{0}})\} & E\{U_{\psi_{1}}(\beta_{\psi_{0}})g_{\psi_{1}}(\gamma_{\psi_{0}})\} & 0 \\
E\{g_{\psi_{1}}(\gamma_{\psi_{0}})U_{\psi_{1}}(\beta_{\psi_{0}})\} & E\{g_{\psi_{1}}(\gamma_{\psi_{0}})g_{\psi_{1}}(\gamma_{\psi_{0}})\} & 0 \\
0 & 0 & \frac{1-c_{\psi}}{c_{\psi}} E\{g_{\psi_{1}}(\gamma_{\psi_{0}})g_{\psi_{1}}(\gamma_{\psi_{0}})\}
\end{bmatrix},
\]

and
\[
\Sigma_{n} = \begin{bmatrix}
0 & \Sigma_{12} \\
\Sigma_{12}^{\top} & -\Sigma_{22}
\end{bmatrix}.
\]

Standard argument shows that, as \(\min(n_{\psi}, m_{\psi}) \to \infty\),
\[
s_{12} P \to \Sigma_{12}, \quad s_{22} P \to \Sigma_{22}, \quad S_{n} P \to \Sigma_{n}. \tag{4.19}
\]

Now for \(l(\beta_{x}, \gamma_{x}, \beta_{y}, \gamma_{y})\), using Taylor expansion on (7) and ignoring the constant term \(\sum_{\psi}\{2n_{\psi} \log(n_{\psi}) + 2m_{\psi} \log(m_{\psi})\}\), we have
\[
l(\beta_{x}, \gamma_{x}, \beta_{y}, \gamma_{y}) = l(\beta_{x}, \gamma_{x}) + l(\beta_{y}, \gamma_{y}),
\]

where
\[
l(\beta_{\psi}, \gamma_{\psi}) = 2\tilde{t}_{\psi_{1}} \sum_{i=1}^{n_{\psi}} U_{\psi_{1}}(\beta_{\psi_{i}}) + 2\tilde{t}_{\psi_{2}} \sum_{i=1}^{m_{\psi}} g_{\psi_{i}}(\gamma_{\psi_{i}}) - \tilde{t}_{\psi_{1}} \sum_{i=1}^{n_{\psi}} U_{\psi_{1}}(\beta_{\psi_{i}})U_{\psi_{1}}^{\top}(\beta_{\psi_{i}})\tilde{t}_{\psi_{1}}
\]
\[
-2\tilde{t}_{\psi_{2}} \sum_{i=1}^{n_{\psi}} U_{\psi_{1}}(\beta_{\psi_{i}})g_{\psi_{i}}(\gamma_{\psi_{i}})\tilde{t}_{\psi_{2}} - \tilde{t}_{\psi_{2}} \sum_{i=1}^{m_{\psi}} g_{\psi_{i}}(\gamma_{\psi_{i}})g_{\psi_{i}}^{\top}(\gamma_{\psi_{i}})\tilde{t}_{\psi_{2}}
\]
\[
+2\tilde{t}_{\psi_{3}} \sum_{k=1}^{m_{\psi}} g_{\psi_{i}}(\tilde{\gamma}_{\psi_{i}})\tilde{t}_{\psi_{3}} - \tilde{t}_{\psi_{3}} \sum_{k=1}^{m_{\psi}} g_{\psi_{i}}(\tilde{\gamma}_{\psi_{i}})g_{\psi_{i}}^{\top}(\tilde{\gamma}_{\psi_{i}})\tilde{t}_{\psi_{3}}
\]
\[
+o_{p}(1). \tag{4.20}
\]
Simplifying (4.20) with (4.16)-(4.18) leads to

\[
\ell(\hat{\beta}_0, \hat{\gamma}_0) = -2\hat{\ell}_{\psi 1} \sum_{i=1}^{n_{\psi}} U'_{\psi i}(\hat{\beta}_0)(\hat{\beta}_0 - \beta_{00}) - 2\hat{\ell}_{\psi 2} \sum_{i=1}^{n_{\psi}} g'_{\psi i}(\hat{\gamma}_0)(\hat{\gamma}_0 - \gamma_{00})
\]

\[
-2\hat{\ell}_{\psi 3} \sum_{k=1}^{m_{\psi}} g'_{\psi(n_{\psi}+k)}(\hat{\gamma}_0)(\hat{\gamma}_0 - \gamma_{00}) + \ell'_{\phi 1} \sum_{i=1}^{n_{\psi}} U_{\psi i}(\hat{\beta}_0)U'_{\psi i}(\hat{\beta}_0)\hat{\ell}_{\psi 1}
\]

\[
+2\hat{\ell}_{\phi 1} \sum_{i=1}^{n_{\psi}} U_{\psi i}(\hat{\beta}_0)g'_{\psi i}(\hat{\gamma}_0)\hat{\ell}_{\psi 2} + \hat{\ell}_{\phi 2} \sum_{i=1}^{n_{\psi}} g_{\psi i}(\hat{\gamma}_0)g'_{\psi i}(\hat{\gamma}_0)\hat{\ell}_{\psi 2}
\]

\[
+\ell'_{\phi 3} \sum_{k=1}^{m_{\psi}} g_{\phi(n_{\psi}+k)}(\hat{\gamma}_0)g'_{\phi(n_{\psi}+k)}(\hat{\gamma}_0)\hat{\ell}_{\phi 3} + o_p(1)
\]

\[
= n_{\psi}\hat{\ell}_{\psi} s_{22}\hat{\ell}_{\psi} - 2\hat{\ell}_{\psi} s_{12}\{(\hat{\beta}_0 - \beta_{00})^T, (\hat{\gamma}_0 - \gamma_{00})^T\} + o_p(1). \tag{4.21}
\]

By ignoring the terms of \(o_p(n_{\psi}^{-1/3})\), we have

\[
\hat{\ell}_{\psi} = n_{\psi}^{-1}\{-s_{22}^{-1} + s_{22}^{-1}s_{12}^{-1}(s_{12}^{-1}s_{22}^{-1}s_{12}^{-1})^{-1}s_{12}^{-1}s_{22}^{-1}\}T_{n\psi},
\]

and \(\hat{\ell}_{\phi} s_{12}\psi = 0\). Thus

\[
\ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y) = (\sqrt{n}_{\beta_x}\hat{\ell}_x, \sqrt{n}_{\beta_y}\hat{\ell}_y)\{s_{22}^{-1} + s_{22}^{-1}s_{12}^{-1}(s_{12}^{-1}s_{22}^{-1}s_{12}^{-1})^{-1}s_{12}^{-1}s_{22}^{-1}\}
\]

\[
\times(\sqrt{n}_{\beta_x}\hat{\ell}_x, \sqrt{n}_{\beta_y}\hat{\ell}_y)\} + o_p(1). \tag{4.22}
\]

For \(\ell(\hat{\beta}, \hat{\gamma}_x, \hat{\gamma}_y)\), the derivation is similar. We need replace \(s_{12}\) with

\[
s_{12}^* = \begin{bmatrix} s_{12x}^* & s_{12y}^* \end{bmatrix}, \text{ where }
\]

\[
s_{12x} = \begin{bmatrix} n_{\psi}^{-1} \sum_{i=1}^{n_{\psi}} U'_{\psi i}(\hat{\beta}_0) & 0 & 0 \\
0 & n_{x}^{-1} \sum_{i=1}^{n_{x}} g'_{x}(\gamma_{00}) & n_{x}^{-1} \sum_{i=1}^{n_{x}} g'_{x(n_{x}+k)}(\gamma_{00}) \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & n_{y}^{-1} \sum_{i=1}^{n_{y}} g'_{y}(\gamma_{00}) & n_{y}^{-1} \sum_{i=1}^{n_{y}} g'_{y(n_{y}+k)}(\gamma_{00}) & n_{y}^{-1} \sum_{i=1}^{n_{y}} g'_{y(n_{y}+k)}(\gamma_{00}) \\
\end{bmatrix}
\]

\[
s_{12y} = \begin{bmatrix} n_{\psi}^{-1} \sum_{i=1}^{n_{\psi}} U'_{\psi i}(\hat{\beta}_0) & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & n_{y}^{-1} \sum_{i=1}^{n_{y}} g'_{y}(\gamma_{00}) & n_{y}^{-1} \sum_{i=1}^{n_{y}} g'_{y(n_{y}+k)}(\gamma_{00}) \\
\end{bmatrix}
\]
Accordingly, define
\[
\Sigma_{12}^* = \begin{bmatrix}
\Sigma_{12x}^* & \Sigma_{12y}^*
\end{bmatrix}, \quad \text{where}
\]
\[
\Sigma_{12x}^* = \begin{bmatrix}
E U_x'(\beta_0) & 0 & 0 \\
0 & E g_x'(\gamma_{x0}) & \frac{1-\epsilon_x}{c_x} E g_x'(\gamma_{x0}) \\
0 & 0 & 0
\end{bmatrix},
\]
\[
\Sigma_{12y}^* = \begin{bmatrix}
E U_y'(\beta_0) & 0 & 0 \\
0 & 0 & 0 \\
0 & E g_y'(\gamma_{y0}) & \frac{1-\epsilon_y}{c_y} E g_y'(\gamma_{y0})
\end{bmatrix}.
\]

We can then show
\[
\ell(\hat{\beta}, \hat{\gamma}_x, \hat{\gamma}_y) = (\sqrt{n}_x T_{nx}^\tau, \sqrt{n}_y T_{ny}^\tau) \{\Sigma_{12x}^{-1} - \Sigma_{12x}^{-1} \Sigma_{12}^* (\Sigma_{12}^* \Sigma_{12x}^{-1} \Sigma_{12x}^* - 1) \Sigma_{12}^* \Sigma_{12x}^{-1} \}
\]
\[
\times (\sqrt{n}_x T_{nx}^\tau, \sqrt{n}_y T_{ny}^\tau)^\tau + o_p(1). \quad (4.23)
\]

Thus
\[
\ell(\hat{\beta}, \hat{\gamma}_x, \hat{\gamma}_y) - \ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y) = (\sqrt{n}_x T_{nx}^\tau, \sqrt{n}_y T_{ny}^\tau) \Sigma_{22}^{-1/2} W \Sigma_{22}^{-1/2} (\sqrt{n}_x T_{nx}^\tau, \sqrt{n}_y T_{ny}^\tau)^\tau,
\]
where
\[
W = \Sigma_{22}^{-1/2} \left\{ \Sigma_{12x}^* (\Sigma_{12} \Sigma_{12x}^{-1} \Sigma_{12x}^* - 1) \Sigma_{12}^* - \Sigma_{12x}^* (\Sigma_{12} \Sigma_{12x}^{-1} \Sigma_{12x}^* - 1) \Sigma_{12}^* \right\} \Sigma_{22}^{-1/2}.
\]

Note that \(\Sigma_{22}^{-1/2} (\sqrt{n}_x T_{nx}^\tau, \sqrt{n}_y T_{ny}^\tau)^\tau \overset{d}{\rightarrow} N(0, I_{2p+4r})\), \(\text{tr}(W) = p\) and that \(W\) is symmetric and idempotent. This suggests that \(R_0\) has a asymptotic \(\chi_p^2\) distribution. \(\square\)
CHAPTER 5. General Discussion and Future Research

Since the early 1970's, statistical methods for missing data have seen great advances, but there are still areas in need of further research. The most popular method for missing data problems as of now is probably the multiple imputation method proposed by Rubin (1987). It should be noted that though both multiple imputation and the nonparametric imputation method described in this thesis involve generating multiple copies of imputed observations for each missing value, there are important differences. In multiple imputation, several "complete" data sets are generated through imputation, and each data set is analyzed using methods for complete data. Then the results from these multiple inferences are combined to provide the final estimate and the corresponding standard error. In the nonparametric imputation method discussed in Chapter 2, the value of the estimating function for each incomplete observation is the average over several imputed values, and only one inference is made using the empirical likelihood method. The theoretical root of multiple imputation is Bayesian, where the imputed value is drawn from the predictive distribution of the missing variable. In practice, deriving the predictive distribution can sometimes be very difficult or impossible. Practitioners often use various versions of hot deck imputation or matching method to make draws for missing observations. In one sense, the usage of a kernel estimator of the conditional distribution \( F(y|X) \) in the nonparametric imputation provides a formal framework for "matching" donors of the imputed value. Using empirical likelihood methods gives the advantage that both imputation and inference require very few assumptions. Future research includes extending the nonparametric imputation method to more complex
missing patterns. Particularly for monotone missing patterns, it will be interesting to see whether we could impute different variables sequentially with increasing missing proportions. The empirical likelihood based inference for multi-sample testing with surrogate variables can be applied to parameters other than the mean functional. In the selective transcriptional profiling study, chi-square calibration is sufficient for the calibration of the empirical likelihood ratio. But for cases more complex than that of the mean, a bootstrap procedure similar to that of Chen et al. (2003) is probably necessary. Extension to include estimated propensity scores is another interesting problem.

Missing data methods are relevant to biological research for two main reasons. One is that as biologists increasingly deal with huge datasets, missing data become more prevalent. When data come from different experiments or research projects, it is often not possible to have a complete data matrix. Flexible methods with weak model assumptions can help in efficiently utilizing information in the data set. A second aspect of missing data problems in biological research involves missingness by design. Since many measurements in biological research are financially expensive or technically difficult to obtain, researchers naturally tend to obtain more data on measurements easy to obtain, but only get a small number of observations for measurements that are expensive in money or time. Selective transcriptional profiling is an example of this type of experiment.

5.1 References


ACKNOWLEDGEMENTS

I would like to express my gratitude to my major professors, Drs. Nettleton and Chen, for their guidance and tremendous efforts afforded to my research projects. I want to thank my committee members, Drs. Dekkers, Koehler, Lahiri, and Marasinghe, for their advice and encouragement. Thanks to Dr. Marasinghe and Mr. Stuart Gardner for providing FORTRAN code examples. Thanks to Dr. Kenneth Manly for bringing to my attention the resources at genenetwork.org. Also thanks to Dr. Isaacson and various faculty, staff, and students in the Department of Statistics whose help and assistance will be too numerous to acknowledge in detail. Finally, thank my family for the great support.