COMBINING TEST STATISTICS AND INFORMATION CRITERIA FOR HIGH DIMENSIONAL DATA INTEGRATION

YAWEN XU

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

April 2015

Abstract

This research is focused on high dimensional data integration by combing test statistics or information criteria. Our research contains four projects.

In the first project, we propose an integration method to perform hypothesis testing and biomarkers selection based on multi-platform data sets observed from normal and diseased populations. The types of test statistics can vary across the platforms and their marginal distributions can be different. The observed test statistics are aggregated across different data platforms in a weighted scheme, where the weights take into account different variabilities possessed by test statistics. The overall decision is based on the empirical distribution of the aggregated statistic obtained through random permutations. In both simulation studies and real biological data analyses, our proposed method has better control over false discovery rates and higher positive selection rates than the uncombined method.

In mixed data clustering project, we propose a non-parametric clustering method for handling mixed data with both continuous and discrete random variables. The product space of the continuous and discrete sample space is transformed into a new product space based on adaptive quantization on the continuous part. Cluster patterns are detected locally by using a weighted modified Chi-squared test. Results from simulation studies and real data analysis have shown that our method out-performs the benchmark method, AutoClass, in various settings.

In the multiple data sets model selection project, we propose weighted integrative AICs as a model selection criterion. Our method combines AICs with different weights across multiple data sets. The weights are chosen to minimize the variance of integrative AICs. In the simulation studies, we compare our method with individual AIC method and integrative AICs with equal weights method. Our method has the better performance over false negative numbers and false detected numbers of the selected variables.

In the last project, we extend Linharts and Shirmodarias test statistics under composite likelihood function with local alternatives for correlated data set. Comparing to first order method, our simulation results show that our second order method improves the accuracy for estimating the variance of difference of AICs and reduces the error probability when conduct model comparison test.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisors, Professor Xin Gao and Professor Steven Wang, whose excellent guidance, caring patience, and advice throughout all the stages of this dissertation. I have been extremely lucky to have two supervisors who cared so much about my work and who responded my questions and queries so promptly. I also thank them for being open persons to ideas, and for encouraging and helping me to shape my interests and ideas.

I would like to express my deepest appreciation to my committee members, Professor Augustine Wong, who always generously offered his support and helpful suggestions, which makes my research life smooth. His help as well as friendship is a bountiful treasure to me forever.

My special thanks goes to Professor Zeny Feng, Professor Georges Monette and other professors whose courses and comments have been of great help to my research.

I wish to thank my family, especially my parents and my son, for all their love, encouragement and support. Thank for my parents always believing in me, for their continuous

love and their supports in my decisions. Their love provide my inspiration and is my driving force. For my son Yuhao, I owe him lots and lots of fun hours. I couldn't imagine doing my PhD without him; he really gave me the reason to continue. Words would never say how grateful I am to all of you.

Finally, I offer my regards and thanks to all of those who supported me in any respect during the completion of the dissertation.

Table of Contents

List of Tables

List of Figures

1 Combining two t-test statistics.

1.1 Introduction

In gene expression experiments, the expression levels of thousands of genes are simultaneously monitored to study the underlying biological process. In proteomic data, the protein levels or protein counts are measured for thousands of genes simultaneously. In addition, there are other types of genomic data with different sizes, formats and structures. Each distinct data type, such as gene expression, protein counts, or single nucleotide polymorphisms, provide potentially valuable and complementary information regarding the involvement of a given gene in a biological process. Many biomarkers that play important roles in biological processes behave differently in treatment versus control groups; this phenomenon can be observed consistently across various data platforms. Therefore, integrating related data sets from different sources is crucial to correctly identify the significant underlying biomarkers. Integrative analysis of multiple data types would improve the identification of biomarkers of clinical end points (Reif *et al.*, 2004). However, the integration of data from different sources poses a number of challenges. First, genomic data come in a wide variety of data formats. For example, expression data are recorded as continuous measurements, whereas proteomic data often consist of discrete counting variables. One may wish to convert data into a common format and common dimension, but this is not always practical or feasible (Hamid *et al.*, 2009). Second, different data sets are collected under different experimental settings. Therefore, the distribution of the measurements as well as the quality of the experiments may vary from data set to data set. Third, measurements obtained across different data platforms could be collected from the same or related biological samples. Therefore, measurements across different data types could have complicated dependency relationships.

The practice of combining different data sources to perform classification analysis has been considered in the literature. Efforts to integrate data and improve classification accuracy are widely seen in recent studies (Lanckriet *et al.*, 2004; Daemen *et al.*, 2008; Buness *et al.*, 2009). In contrast to performing classification on biological samples, our main objective is to select important biomarkers for an underlying biological process. Correlation analysis has been proposed to integrate diverse data types and assimilate them into biological models for the prediction of cellular behaviour and clinical outcome. Tian *et al.* (2004) performed a correlation analysis of protein and mRNA expression data using the cosine correlation metric for comparison. Bussey *et al.* (2006) integrated data on DNA copy number with gene expression levels and drug sensitivities in cancer cell lines based on Pearson's correlation coefficients. Adourian *et al.* (2008) presented a cross-compartment correlation network approach to integrate proteomic, metabolomic, and transcriptomic data for selecting circulating biomarkers; partial pairwise Pearson's correlations controlling for treatment group means were calculated. The markers with concordant RNA and protein expression were included in the prediction models, while discordant ones were excluded. However, this approach might miss some important biological information, such as protein-protein interactions and protein-gene interactions (Ma *et al.*, 2009). Another limitation is that correlation analysis mainly captures the strength of the correlation among measurements across different platforms; however, strong correlation only demonstrates consistent outcome across different platforms and does not directly translate to significant involvement in a biological process. Furthermore, statistical evidence from complicated data sets, such as factorial experiments, times series, or longitudinal data, cannot be summarized.

The problem of how to reliably combine data from different experiment platforms to identify significant biomarkers has recently received considerable attention in the bioinformatics literature. The rank aggregation method (Aerts *et al.*, 2006) has been proposed for ranking genes by similarity to the disease genes in Gene Ontology, pathways, transcription factor binding sites, and sequence, then aggregating this rankings to get the final result. Rhodes *et al.* (2004) combined four independent data sets to identify genes deregulated in prostate cancer. For each gene in each data set, a p-value was obtained as an indication of the probability that the gene was differentially expressed. P-values for different data sets were subsequently aggregated to provide an overall estimate of the genes' significance of being differentially expressed during prostate cancer. However, combining genes' ranks in the rank aggregation approach or p-values in the meta-profiling method ignores the underlying multivariate distributions of the ranks or p-values. Furthermore, data quality may vary across different data sources. The two aggregation methods detailed above essentially give equal weights to different data sets. Thus, we propose to combine statistical evidence across different platforms through summary statistics instead of raw data. For each experimental platform, we formulate a null hypothesis and construct the summary test statistic. By randomization, we obtain the null distribution of the vector of statistics across different platforms. The test statistics are summarized across different platforms in a weighted scheme, where the weights take into account different variabilities possessed by the statistics. The method allows the use of different types of summary statistics from different platforms, which gives great flexibility and generality with respect to its application.

The proposed method is similar in spirit to a meta-analysis. Both methods combine statistical evidence across multiple data sets. However, in meta-analysis different data sets are based on the same type of experiments or observational studies, and therefore the measurements are the same variables. Across different data sets, the quality of the data may vary. The goal of meta-analysis is to fully utilize all the information from different data sets and construct a weighted estimate of the effect size. Different weighting schemes are available depending on the statistical models (Hu *et al.*, 2006). On the other hand, data integration focuses on integrating statistical evidence across different experimental types. There is no common effect size to estimate across various data sets. In our proposed method, we use a weighted average of the test statistics across different data platforms, but the test statistics are summaries of evidence towards different sub-hypotheses rather than summaries of common effect size as in fixed effect meta-analysis. The proposed integration method does not check for differences across the platforms.

1.2 Methods

The aim of our multi-platform integration method is to select a set of significant biomarkers that are involved in a biological process and thus behave differently in the treatment group and the control group. In order to combine statistical evidence across different platforms, our method requires that analogous hypotheses based on the features being measured are formulated for each platform. Each null analogous hypothesis specifies the unrelatedness of the biomarker in that particular experimental setting, but all of them infer the unrelatedness of the biomarker to the biological process being investigated. Based on the set of Q analogous hypotheses for Q data sources, we construct a set of Q corresponding test statistics for each type of data. The test statistics can be different and tailored to the specific experimental settings. For example, if the microarray experiment has a multifactorial design, the appropriate test statistic can be an F statistic based on an ANOVA test. If the proteomics experiment generates counting data for diseased versus normal groups, the appropriate test statistic can be a nonparametric Wilcoxon rank sum test. A vector of observed statistics across multi-platforms is obtained. We then randomly permute data across diseased and control groups. All measurements from different platforms are permuted. In this way, we obtain an empirical null distribution of the vector of test statistics. In order to pool the randomized values of the statistics across the biomarkers to form the empirical null distribution, we assume data from different biomarkers are independent or have an exchangeable correlation structure. For the validity of the randomization procedure, we assume an exchangeable covariance structure for the measurements within each platform. Finally, we construct a weighted sum of the test statistics across different platforms with the weights being the inverse of the empirical standard deviation of each statistic. We determine a set of significant biomarkers based on the aggregated test statistic.

In the following, we demonstrate our method by integrating microarray expression data and proteomic data as an example. We consider two experiments, the first having microarray expression data measured on l_1 diseased samples and l_2 control samples and the second having proteomic data measured on m_1 diseased samples and m_2 control samples. The objective is to find biomarkers significantly involved in disease development.

Step 1): Define two analogous null hypotheses. For microarray data, the null hypothesis would be H_{01} : the gene's mRNA level is the same in diseased and normal populations; for proteomic data, the null hypothesis would be H_{02} : the protein level is the same in diseased and normal populations.

Step 2): Based on the hypotheses, construct two test statistics, t_m and t_p, tailored to each type of data. Consequently, we obtain a vector of two observed statistics $(t_m, t_p)^T$ across two data platforms. The test statistics can be of any type as long as they summarize information from the data and can be used to assess the statistical significance of the data toward the hypotheses. Let $x_1 = (x_{11}, \ldots, x_{1l_1})^T$ denote the l_1 gene expression measurements in the disease group, $x_2 = (x_{21}, \ldots, x_{2l_2})^T$ denote the l_2 gene expression measurements in the control group, $\overline{x}_1 = \sum_{j=1}^{l1} x_{1j}/l_1$, and $\overline{x}_2 = \sum_{j=1}^{l2} x_{2j}/l_2$. Similarly, $y_1 = (y_{11}, \ldots, y_{1m_1})^T$ denotes the m_1 protein measurements in the disease group and $y_2 = (y_{21}, \ldots, y_{2m_2})^T$ denotes the m_2 protein measurements in the control group, $\overline{y}_1 = \sum_{j=1}^{m_1} y_{1j}/m_1$, and $\overline{y}_2 = \sum_{j=1}^{m_2} y_{2j}/m_2$. For illustration purpose, we adopt Behrens-Fisher test statistics for each of the data:

$$t_m = \frac{\overline{x}_2 - \overline{x}_1}{\sqrt{\frac{s^2(x_1)}{l_1} + \frac{s^2(x_2)}{l_2}}},$$

and

$$t_p = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\frac{s^2(y_1)}{m_1} + \frac{s^2(y_2)}{m_2}}},$$

where s^2 denotes the sample variance. The test statistics should be formulated so that a larger test statistic in the positive direction indicates more evidence towards the alternative hypotheses. For example, if Student's t-statistic is used, then a one-sided alternative hypothesis corresponds to a one-sided t-statistic, whereas the two-sided alternative leads to the absolute value of the t-statistic. Consider n genes being measured in the experiments and we obtain n vectors of test statistics $(t_{mi}, t_{pi})^T$, $i = 1, \ldots, n$, from the data sets.

Step 3): The samples are randomly permuted across diseased and control groups. If the same sample is being measured across different platforms, all the measurements from the different platform are permuted simultaneously. The simultaneous permutation preserves the dependency relationship among the measurements from different platforms. Based on random permutation, we obtain an empirical null distribution of the vector $(t_m, t_p)^T$.

Step 4): The aggregated test statistic will be:

$$t_A = \frac{t_m}{\hat{\sigma}_1} + \frac{t_p}{\hat{\sigma}_2},$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimated standard deviations of t_m and t_p based on the empirical null distribution, and t_m and t_p are the test statistics or the absolute values of the test statistics based on the direction of the alternative hypotheses. The average estimated weights can be included because it reflects the variability of the test statistic. Our weights are chosen as the standard deviations. The weights allow to assign larger weight to the test statistic with smaller variation, and assign smaller weight to the test statistic with larger variation. At significance level α, we choose a threshold C_α, such that $P_{H_{01} \cap H_{02}}(t_A > C_\alpha) = \alpha$. Specifically, C_α is the $100(1-\alpha)\%$ percentile of t_A, which can be obtained from the empirical null distribution. Construct a decision line that separates selected significant biomarkers and nonsignificant biomarkers. The resulting separation line is:

$$\frac{t_m}{\hat{\sigma}_1} + \frac{t_p}{\hat{\sigma}_2} = C_\alpha.$$

All the biomarkers with (t_m, t_p) above the separation line will be declared as significantly involved in the disease development.

In the more general case, suppose we have Q data platforms with the observed test statistics $(t_1, \ldots, t_Q)^T$. From random permutation, we obtain the joint empirical distribution of this vector of test statistics under the global null hypothesis. Let $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_Q^2$ denote the estimated variance of the individual test statistics.The aggregated test statistic takes the form:

$$t_A = \sum_{i=1}^{Q} \frac{t_i}{\hat{\sigma}_i}.$$

The resulting critical region will take the form:

$$\frac{t_1}{\hat{\sigma}_1} + ... + \frac{t_Q}{\hat{\sigma}_Q} > C_\alpha,$$

where C_α is the $100(1-\alpha)\%$ percentile of t_A. Any biomarker with $t_A > C_\alpha$ will be selected as behaving significantly differently between the diseased group and control group.

Our method aggregates actual values of the test statistics across different data platforms, which preserves more information compared to the rank aggregation method. Moreover, our method assigns different weights to each data set according to the variability of the test statistics: the larger the variation in the test statistic, the smaller the weight assigned to it, and vice versa. The threshold C_α is determined based on the empirical null distribution of the aggregated test statistics, which implicitly takes into account the dependency relationships among the test statistics. Furthermore, our method can deal with different data types and formats generated by various experimental settings.

There are two major ways to perform the multiplicity adjustment. The first is the Bonferroni correction. If we wish to control the familywise type I error rate at α^*, then the individual level $\alpha = \alpha^*/n$, where n is the total number of biomarkers. When n is large, the Bonferroni correction leads to very stringent tests with α being very small. Alternatively, we can control the number of false discoveries. To set the number of false discoveries to be equal to or less than f, then $\alpha = f/(n\hat{\pi})$, where $\hat{\pi}$ is the estimated proportion of non-differentially expressed biomarkers. If there is no $\hat{\pi}$ available, we use $\hat{\pi} = 1$ and that gives a conservative value for α.

Different platforms can be used to test different sub-hypothesis. All of these sub-hypotheses should be concordant in supporting the overall biological hypothesis. For example, the involvement of a gene in disease development can be supported by both mRNA expression level changes and proteomic level changes. In most cases, changes in measurements from different platforms are expected to occur in the same direction. However, our method is also applicable even if the changes are in different directions, as long as the statistical evidence from both sources can be combined. For example, consider H_{10} : mRNA is increasing in normal group; H_{20}: antibody count is decreasing in normal group. Even though the actual measurements from two platforms are negatively correlated, we can construct the test statistics t_1 and t_2 so that the positive value of the statistics supports the alternative hypotheses and the weighted average can be used as combined evidence of the involvement of the biomarker in the process.

1.3 Simulation Study Results

1.3.1 Results on Simulated Data

In this section, we examine the performance of our proposed method by examining its positive selection rates and false discovery rates under various testing scenarios. We simulate data sets from Q different platforms. The number Q is set to be either 2 or 5. For the qth experiment, the data set is denoted as X_q. For each data set, we assume that n different biomarkers are measured, $X_q = (X_{q1}^T, ..., X_{qn}^T)^T$. For the ith biomarker, $X_{qi} = (X_{qi1}^T, X_{qi2}^T)^T$, where X_{qi1} denotes data from the control group with mean μ_{qi1} and X_{qi2} denotes data from the diseased group with mean μ_{qi2}. The total number of biomarkers is set to be $n = 1000$. Among the n biomarkers, let g denote the number of biomarkers that are related to the biological process of interest, i.e. $\mu_{qi1} \neq \mu_{qi2}$. The number g of differentially expressed (DE) biomarkers is set to be 200. The number of measurements for each biomarker obtained from each platform is set to be 10, in which 5 are from the control group and the other 5 are from the disease group. We also consider different effect sizes. For continuous data, we generate $X_{qi} \sim$ MVN($(\mu_{qi1}^T, \mu_{qi2}^T)^T$, Σ), where Σ has an exchangeable correlation structure with correlation ρ. The correlation ρ is set to be either 0 or 0.5. For differentially expressed markers, $\mu_{qi1} = 0 \times \mathbf{1}_m$, $\mu_{qi2} = e \times \mathbf{1}_m$, where e is the effect size and $m = 5$ is number of measurements. Discrete data X_{qi} is generated from a Poisson(λ) distribution, where $\lambda_{qi1} = \mu_{qi1}$ for the control group and $\mu_{qi2} = \mu_{qi1} + e$ for the diseased group. The g differentially expressed markers are divided into two groups with $g_1 = 100$ and $g_2 = 100$. Each group is assigned a different effect size e. For each platform, the alternative hypothesis can be either left-sided, right-sided or two-sided. The number of permutation is 100. All of the permuted values from the n biomarkers are pooled together to form the empirical null distribution. The results are summarized for 100 simulated data sets.

To compare our multi-platform integration method with the individual platform analysis method, the positive selection rate (PSR) and false discovery rate (FDR) are calculated to assess the performance of each method for selecting the differentially expressed biomarkers:

$$\text{PSR} = \frac{\text{\# of correctly identified DE biomarkers}}{\text{\# of DE biomarkers}}$$

and

$$\text{FDR} = \frac{\text{\# of falsely identified DE biomarkers}}{\text{\# of identified DE biomarkers}}$$

Tables 1.1, 1.2, and 1.3 provide detailed simulation settings and results at the $\alpha = 0.05$ significance level. From the results, we can see that our multi-platform integration method has the highest PSR and the lowest FDR with the smallest variance compared to all other individual platform analyses in all scenarios. In addition, such advantage is consistently observed regardless of whether or not there is correlation among the measurements obtained for each biomarkers. Table 1.1 summarizes the results for the integrative analysis based on two different platforms. Given different effect sizes, one or two sided alternatives, and different correlations, the increase in PSR is consistently about 40% and the decrease in FDR is about 30% compared to the results from individual platforms. Table 1.2 summarizes the results for the integrative analysis based on five different platforms. Given different simulation scenarios, the increase in PSR for most cases is about 60% and the decrease in FDR is about 40% compared to the results from individual platforms. This shows that by integrating more data from different sources, we are improving the sensitivity and selectivity of the proposed method. Table 1.3 summarizes the results for the integrative analysis based on two different platforms, where the first consists of continuous data and the second consists of discrete data. Similar to the setting with two continuous data sets, the increase in PSR is about 40% and the decrease in FDR is about 30% compared to the results from individual platforms.

Figure 1.1 demonstrates decision lines from different methods. The plot is constructed based on the results from one simulated data set and contains three decision lines: the vertical line using data from the first individual platform, the horizontal line using data from the second individual platform, and the dashed line based on our multi-platform integration method. Our decision line provides a greatly improved separation of the differentially and non-differentially expressed biomarkers. Moreover, the individual platform analysis misidentifies some of the data points compared to our method.

As we examine a large number of biomarkers, we need to investigate the control of the false discovery rate of the proposed method with regards to multiple hypothesis testing (Gao, 2006). Given a fixed cut-off value of α, we obtain the realized false discovery rate $FDR = (FP)/(T\hat{P})$ and its estimates $F\hat{D}R = (F\hat{P})/(T\hat{P})$, where FP denotes the number of false positive biomarkers, $F\hat{P} = n\pi\alpha$ is the estimated number of false positive biomarkers, $T\hat{P}$ is the total number of biomarkers claimed as positive, π is the proportion of non-differentially expressed genes, and $\hat{\pi}$ is its estimator. We can control the estimated number of false positive discoveries by selecting the significance level of the approaches. We expect that the estimated $F\hat{P}$ should be close to the true FP; the $F\hat{D}R$ should be close to the

Table 1.1: The simulation results for two platforms with continuous data.

		Methods		
		multi-platform	1st Platform	2nd Platform
Scenario 1:	ρ = 0; g = g_1 + g_2 = 200			
Right-side	Platform1:	e = 0.5 for g_1 = 100; e = 2 for g_2 = 100		
	Platform2:	e = 1.5 for g_1 = 100; e = 1 for g_2 = 100		
	PSR Mean	0.7895	0.5372	0.5588
	PSR Var	0.0007	0.0007	0.0010
	FDR Mean	0.1907	0.2680	0.2600
	FDR Var	0.0007	0.0013	0.0009
Left-side	Platform1:	e = -0.5 for g_1 = 100; e = -2 for g_2 = 100		
	Platform2:	e = -1.5 for g_1 = 100; e = -1 for g_2 = 100		
	PSR Mean	0.7908	0.5330	0.5556
	PSR Var	0.0006	0.0006	0.0012
	FDR Mean	0.1891	0.2673	0.2649
	FDR Var	0.0006	0.0009	0.0011
Two-sided	Platform1:	e = -1 for g_1 = 100; e = 1.5 for g_2 = 100		
	Platform2:	e = 2 for g_1 = 100; e = -1 for g_2 = 100		
	PSR Mean	0.6988	0.4113	0.5403
	PSR Var	0.0011	0.0011	0.0010
	FDR Mean	0.2145	0.3202	0.2694
	FDR Var	0.0007	0.0016	0.0012
Scenario 2:	ρ = 0.5; g = g_1 + g_2 = 200			
Right-side	Platform1:	e = 0.5 for g_1 = 100; e = 2 for g_2 = 100		
	Platform2:	e = 1.5 for g_1 = 100; e = 1 for g_2 = 100		
	PSR Mean	0.9405	0.6319	0.7819
	PSR Var	0.0003	0.0005	0.0007
	FDR Mean	0.1560	0.2410	0.2051
	FDR Var	0.0005	0.0009	0.0007
Left-side	Platform1:	e = -0.5 for g_1 = 100; e = -2 for g_2 = 100		
	Platform2:	e = -1.5 for g_1 = 100; e = -1 for g_2 = 100		
	PSR Mean	0.9400	0.6316	0.7871
	PSR Var	0.0002	0.0004	0.0006
	FDR Mean	0.1605	0.2419	0.2024
	FDR Var	0.0005	0.0007	0.0006
Two-sided	Platform1:	e = -1 for g_1 = 100; e = 1.5 for g_2 = 100		
	Platform2:	e = 2 for g_1 = 100; e = -1 for g_2 = 100		
	PSR Mean	0.9377	0.6670	0.7327
	PSR Var	0.0003	0.0010	0.0007
	FDR Mean	0.1622	0.2270	0.2122
	FDR Var	0.0005	0.0009	0.0007

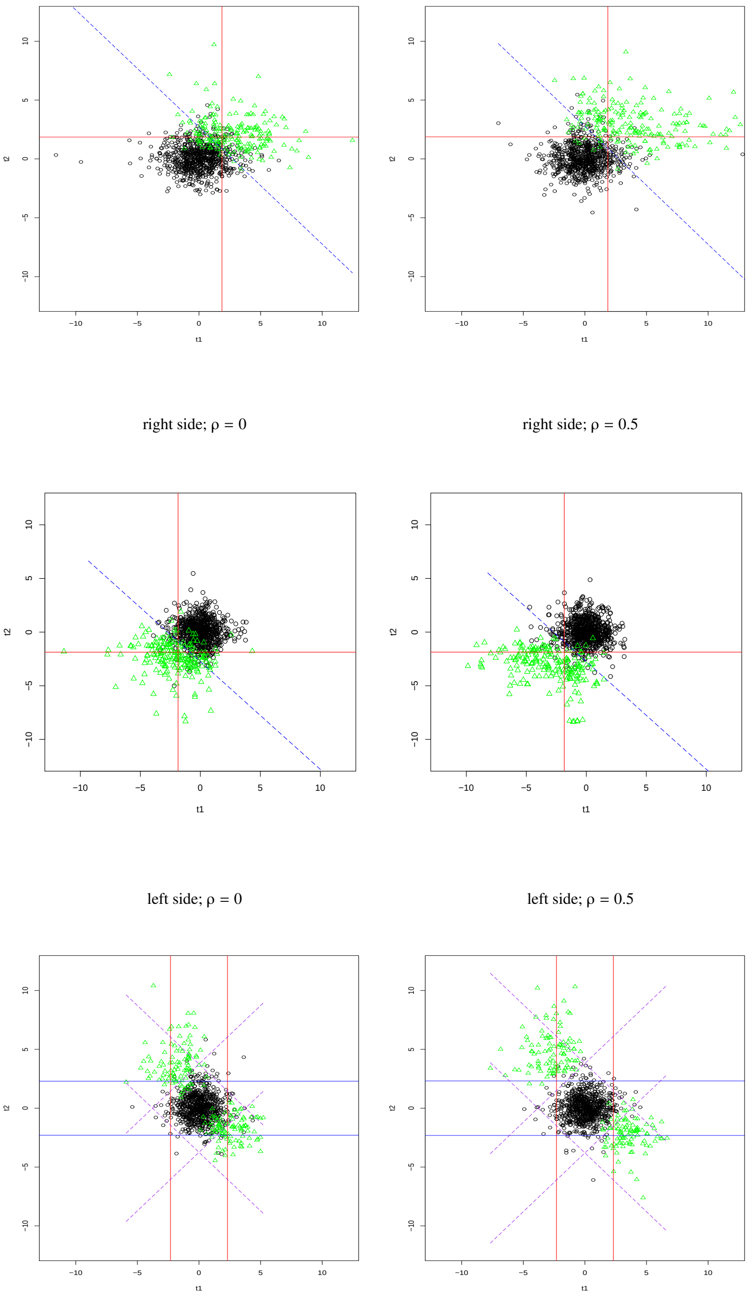
Table 1.2: The simulation settings and results for five platforms with continuous data.

Method	Multi-plat	1st Platform.	2nd Platform.	3rd Platform.	4th Platform.	5th Platform.
Scenario 1:	ρ = 0; g = g_1 + g_2 = 200					
	Platform1:	e = 1.5 for g = 200				
	Platform2:	e = 1.5 for g_1 = 100; e = 1 for g_2 = 100				
	Platform3:	e = -0.5 for g_1 = 100; e = -2 for g_2 = 100				
	Platform4:	e = -1 for g_1 = 100; e = 1.5 for g_2 = 100				
	Platform5:	e = 2 for g_1 = 100; e = -1 for g_2 = 100				
$PSR\ Mean$	0.9517	0.5601	0.4130	0.4464	0.4213	0.4471
$PSR\ Var$	0.0002	0.0012	0.0011	0.0004	0.0010	0.0005
$FDR\ Mean$	0.1572	0.2605	0.3299	0.3108	0.3205	0.2727
$FDR\ Var$	0.0004	0.0011	0.0018	0.0009	0.0010	0.0010
Scenario 2:	ρ = 0.5; g = g_1 + g_2 = 200					
	Platform1:	e = 1.5 for g = 200				
	Platform2:	e = 1.5 for g_1 = 100; e = 1 for g_2 = 100				
	Platform3:	e = -0.5 for g_1 = 100; e = -2 for g_2 = 100				
	Platform4:	e = -1 for g_1 = 100; e = 1.5 for g_2 = 100				
	Platform5:	e = 2 for g_1 = 100; e = -1 for g_2 = 100				
$PSR\ Mean$	0.9998	0.8360	0.6655	0.5682	0.6712	0.5699
$PSR\ Var$	2.7e-06	0.0006	0.0010	0.0004	0.0010	0.0008
$FDR\ Mean$	0.1281	0.1898	0.2217	0.2593	0.2314	0.2093
$FDR\ Var$	0.0004	0.0006	0.0009	0.0007	0.0007	0.0008

Table 1.3: The simulation settings and results for two platforms with continuous data and discrete data.

	Methods		
	multi-platform	1st Platform	2nd Platform
Platform1:	Continuous; $\rho = 0$; e = 0.5 for $g_1 = 100$; e = 2 for $g_2 = 100$		
Platform2:	Discrete; $\mu_{qn1} = 5$, e = 3 for g = 200		
PSR Mean	0.7356	0.5327	0.5228
PSR Var	0.0008	0.0004	0.0012
FDR Mean	0.1967	0.2702	0.2763
FDR Var	0.0008	0.0012	0.0012

Figure 1.1: Decision lines for comparing methods.



right side; ρ = 0

right side; ρ = 0.5

left side; ρ = 0

left side; ρ = 0.5

two-sided; ρ = 0

two-sided; ρ = 0.5

Vertical lines use data from the first individual platform, horizontal lines use data from the second individual platform, and dashed lines use our multi-platform integration method. Circles represent non-differentially expressed biomarkers and triangles represent differentially expressed biomarkers. Plots are based on one simulated data set and 100 permutations.

true *FDR* as well. Under the simulation setting of scenario 2 left-sided case in Table 1.4, the control of the false discovery rate of our proposed method under different significance levels is examined and presented in Table 1.4. With $\pi = 0.8$ and $\alpha = 0.005$, $\hat{FP}$ is aimed to be controlled at 4. On average, our method produces 3.84 false positives, whereas the first and second individual platform analyses have 4.65 and 5.00 false positives, respectively. The corresponding average $F\hat{D}R$ of our method is 0.0225, which is close to the true *FDR* of 0.0214. This demonstrates the integrative analysis yields satisfactory control of false discovery rate, which is improved compared to individual platform analyses.

An ongoing problem in proteomics is that extremely small sample sizes often occur, largely due to biological reasons. To investigate the performance of our method in such situations, we consider a case for each platform where in the control and the diseased groups each has only two measurements. Our method is applied and the simulation results are shown in Table 1.5, scenario 1. Due to the small sample size, the positive selection rate is rather low and the false discovery rate rather high. Nevertheless, the combined method still outperforms the single platform method.

We also consider the situation in which data on the same biomarker from n platforms have a multivariate distribution and the data from the diseased group are independent of those from the control group. The new simulation results are summarized in Table 1.5, scenario 2. The correlation between the platforms is set to 0.5, and the other parameters

Table 1.4: True positives and false discovery rates with $\pi = 0.8$.

Methods	α	0.05	0.01	0.005
	$\hat{FP}$	40	8	4
multi-platform	$\hat{TP}$	224	165	143
	(std)	6.5547	6.0820	5.5202
	FP	44.8125	8.0250	3.8375
	(std)	7.3348	3.4778	2.263
	FDR	0.1563	0.0386	0.0214
	(std)	0.0219	0.0161	0.0125
	$F\hat{D}R$	0.1428	0.0388	0.0225
	(std)	0.0041	0.0014	0.0009
1st individual	$\hat{TP}$	165	107	91
	(std)	8.8797	5.3066	4.9031
	FP	50.5125	9.9000	4.6500
	(std)	8.9101	3.4982	2.1766
	FDR	0.2431	0.0736	0.0406
	(std)	0.0326	0.0246	0.0183
	$F\hat{D}R$	0.1940	0.0600	0.0353
	(std)	0.0103	0.0030	0.0019
2nd individual	$\hat{TP}$	197	106	79
	(std)	7.2442	8.2303	6.3222
	FP	48.9250	9.6000	5.000
	(std)	7.1862	3.5750	2.5376
	FDR	0.1986	0.0721	0.0506
	(std)	0.0245	0.0258	0.0251
	$F\hat{D}R$	0.1630	0.0607	0.0408
	(std)	0.0060	0.0048	0.0033

Table 1.5: Additional simulations.

Method	multi-plat	1st ind.	2nd ind.
Scenario 1:	Extremely small sample size two measurements from each group		
PSR Mean	0.3022	0.2363	0.2179
PSR Var	0.0009	0.0006	0.0007
FDR Mean	0.3782	0.4436	0.4694
FDR Var	0.0023	0.0025	0.0027
Scenario 2:	Correlation among platforms set to 0.5 Disease and normal groups are independent		
PSR Mean	0.6689	0.5365	0.5578
PSR Var	0.0009	0.0008	0.0011
FDR Mean	0.2255	0.2690	0.2641
FDR Var	0.0008	0.0010	0.0010
Scenario 3:	Non-standardized version of t_m and t_p i.e. $t_m = \overline{x}_2 - \overline{x}_1$, $t_p = \overline{y}_2 - \overline{y}_1$		
PSR Mean	0.8142	0.5479	0.5992
PSR Var	0.0009	0.0005	0.0010
FDR Mean	0.1586	0.2358	0.2235
FDR Var	0.0006	0.0011	0.0010

are the same as in Table 1.1, scenario 1, right-sided test. Due to the high correlation among the platforms, the gain in power of the aggregated method is less pronounced than that of the independence case. This is because different platforms contribute overlapping information when they are highly correlated.

The proposed method allows different ways of constructing t_m and t_p as long as they provide summarized statistical evidence for that platform. The Student's *t*-statistic is adopted in the paper simply for illustration purpose. Alternatively, we can simply use the unstandardized differences: $t_m = \overline{x}_1 - \overline{x}_2$, and $t_p = \overline{y}_1 - \overline{y}_2$. Then we proceed with the randomization, obtain the estimated variances for t_m and t_p and form a weighted linear sum statistic. To compare the empirical performance of the standardized versus unstandardized versions, we conduct simulations under the setting 1 of Table 1.1 with right-sided test. The results are summarized in Table 5, scenario 3. The two versions have comparable performance in terms of PSR and FDR. The unstandardized version of t_m and t_p has a slightly higher PSR and a slightly lower FDR.

Our method can be extended to multivariate situation by taking covariance matrix into account. An alternative way of combining test statistics across different platforms is to form a multivariate quadratic statistic. Given two platforms, for example, we consider an alternative test statistic

$$t_Q = (t_m, t_p)^T \hat{\Sigma}^{-1} (t_m, t_p),$$

where $\hat{\Sigma}$ is the estimated covariance matrix of the vector (t_m, t_p) obtained from the empirical null distribution. When t_m and t_p are highly correlated, such multivariate statistic is good to use. This multivariate statistic can be used to test the overall null hypothesis against two-sided alternatives, while the weighted linear statistic that we propose can be used to test one-sided alternatives or two-sided alternatives. Thus, our method is more broadly applicable. We further conduct simulations to compare the multivariate quadratic form with our proposed weighted linear statistic for two-sided tests under the setting of scenario 2, Table 1.1, with results included in Table 1.6. For two-sided alternatives, the quadratic statistic has very similar performance to our proposed weighted linear statistic, with a slightly lower PSR and a slightly higher FDR.

Finally, we compare our method with the existing robust rank aggregation method (Kolde *et al.*, 2012) with results included in Table 1.7. The inference from rank aggregation method is based on the ranks of the test statistics. The ranking can in some degree reflect the significance of the test statistics. But the position of the rank does not always translate into the relatedness of the biomarker to the underlying biological mechanism. The rank aggregation method assigns p-values of the observed ranks under the null hypothesis that the normalized ranks of all biomarkers are uniformly distributed. But this is a null hypothesis which can correspond to two totally different situations: all the biomarkers are not related to the biological process or all of them are related with equal effect size.

Table 1.6: Comparison with the quadratic test statistic t_Q.

Method	multi-plat	Quadratic
$PSR\ Mean$	0.9377	0.9155
$PSR\ Var$	0.0003	0.0004
$FDR\ Mean$	0.1622	0.1804
$FDR\ Var$	0.0005	0.0005
Quadratic:	Exp1:	e = -1 for g_1 = 100; e = 1.5 for g_2 = 100
	Exp2:	e = 2 for g_1 = 100; e = -1 for g_2 = 100

This evaluation of p-values under such global null hypothesis has two implications. First of all, if all the biomarkers are related to the biological process with equal or similar effect sizes, the observed ranks will appear non-informative and thus the method will have little power to detect them. Secondly, the p-value of each observed rank is calculated under the global null hypothesis. Thus, the rank aggregation has a correct error control under the global null hypothesis but has no correct error control under other configurations of the individual hypotheses. In other words, it lack the strong control of the error rate under different configurations of the individual hypothesis (Hochberg and Tamhane, 1987). On the other hand, our method assigns p-values under the individual null hypotheses and thus have a strong control of the error rate. This means our method's actual false discovery rate and estimated false discovery rate will be in good agreement no matter how many of the genes belong to the null situation and how many belong to the alternative situation. While in contrast, the rank aggregation will tend to be very conservative if there are many biomarkers belonging to the alternative situation. To demonstrate this, we choose the number of significant markers ranging from 100, 200 to 400. It is shown in Table 1.7 that the rank aggregation behaves very conservatively in the presence of large number of significant markers. For instance, with five platforms and 200 significant biomarkers, our proposed method has a PSR of 0.9995 and a FDR of 0.1399, while the competing rank aggregation method has a much lower PSR of 0.4995 and FDR of 0.0823. This comparison further demonstrates the advantage of the proposed method. The rank aggregation method relies on the ranking of the test statistics. The higher ranking is, the more important biomarker is. Therefore, the rank aggregation method doesn't work well for some extreme cases. For example, if none of biomarkers are significant, it's hard to distinguish top biomarkers among all biomarkers. However, the rank aggregation method still rank test statistics in order to identify important biomarkers, even in fact all biomarkers are non-significant. Similarly, the rank aggregation method ranks test statistics to identify top biomarkers in the case of all biomarkers are significant.

Table 1.7: Comparison with Robust Rank Aggregation Method.

	Setting:	Method	multi-plat	RRA
1.	ρ = 0.5; g = $g_1 + g_2$ = 100			
	Exp1: e = 1.5 for g = 200	*PSR Mean*	1.000	0.7497
	Exp2: e = 1.5 for g_1 = 100; e = 1 for g_2 = 100	*PSR Var*	1.98e-6	0.0012
	Exp3: e = -0.5 for g_1 = 100; e = -2 for g_2 = 100	*FDR Mean*	0.2803	0.0912
	Exp4: e = -1 for g_1 = 100; e = 1.5 for g_2 = 100	*FDR Var*	0.0011	0.0003
	Exp5: e = 2 for g_1 = 100; e = -1 for g_2 = 100			
2.	ρ = 0.5; g = $g_1 + g_2$ = 200			
	Exp1: e = 1.5 for g = 100	*PSR Mean*	0.9995	0.4995
	Exp2: e = 1.5 for g_1 = 50; e = 1 for g_2 = 50	*PSR Var*	0.23e-06	0.0008
	Exp3: e = -0.5 for g_1 = 50; e = -2 for g_2 = 50	*FDR Mean*	0.1399	0.0823
	Exp4: e = -1 for g_1 = 50; e = 1.5 for g_2 = 50	*FDR Var*	0.0004	0.0004
	Exp5: e = 2 for g_1 = 50; e = -1 for g_2 = 50			
3.	ρ = 0.5; g = $g_1 + g_2$ = 400			
	Exp1: e = 1.5 for g = 100	*PSR Mean*	0.9992	0.1133
	Exp2: e = 1.5 for g_1 = 50; e = 1 for g_2 = 50	*PSR Var*	2.23e-6	0.0002
	Exp3: e = -0.5 for g_1 = 50; e = -2 for g_2 = 50	*FDR Mean*	0.0402	0.0796
	Exp4: e = -1 for g_1 = 50; e = 1.5 for g_2 = 50	*FDR Var*	0.0001	0.0015
	Exp5: e = 2 for g_1 = 50; e = -1 for g_2 = 50			

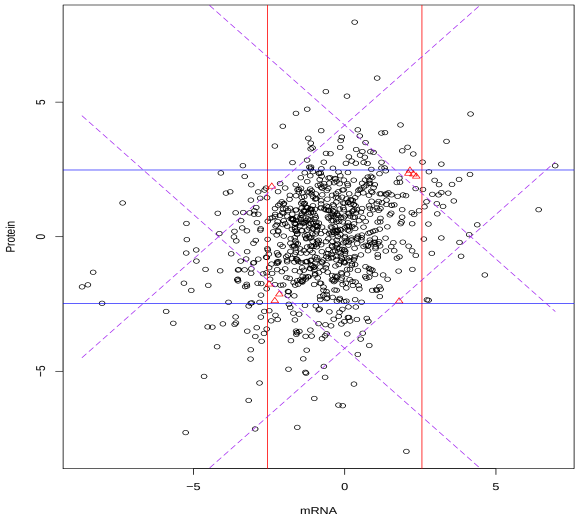
1.3.2 Results on Real Data

In this section, we apply our method to data from a study of growth and stationary phase adaption in *Streptomyces coelicolor* provided by Jayapal (2008). The data set contains both isobaric stable isotope labeled peptide (iTRAQ^{TM})-derived shotgun proteomic data and DNA microarray transcriptome data. To study different growth stages of *S. coelicolor* M145 cells, eight time point cell samples (7, 11, 14, 16, 22, 26, 34, and 38 h) were collected. Because the iTRQA ^{TM}system can only analyze four distinct samples in a single experiment, the eight protein samples were distributed across three runs of mass spectrometric (MS) analysis. The protein sample from 11h was run in three MS experiments, so it serves as a reference. Therefore, protein abundance ratios $r^i_{j/11hr,k}$ were obtained from experimental run k for protein i in sample $j$hr with respect to the 11 h reference. Protein identification and quantification were carried out by comparing the raw spectral data against a theoretical proteome of *S. coelicolor* using proteinPilot^{TM} software and the inbuilt Paragon^{TM} search engine. Only proteins identified with $\geq$ 99% confidence were considered for further analysis. Finally, all identified proteins were further processed to yield a protein abundance ratio with respect to the first time point (7 h) sample using $r^i_{j/7hr} = r^i_{j/11hr}/r^i_{7hr/11hr}$. Ultimately, only 886 proteins identified in the 7 h sample could be used for our analysis.

For microarray data, total mRNA from the same eight time point samples were isolated and a spotted DNA microarray experiment was conducted. Hybridization was performed using genomic DNA (gDNA) as a reference. The mRNA abundance was obtained using $\log_2$[cDNA/gDNA]. To be consistent with the protein data, mRNA abundance data from different samples were processed to calculate $\log_2$[cDNAi/cDNA$_{7hr}$] for each sample with respect to the first time point sample. Only gene expression values with protein values (894 genes) were analyzed. To deal with missing values, we deleted genes that had no values for mRNA at all or had at least five missing values in the protein data set. The rest of the missing values for genes were imputed by using R package MICE. In total, the number of genes suitable for the subsequent integrative analysis was 886. Based on the growth curve, time points were divided into two groups; those from 7, 11, 14 and 16 h represented the growth phase and those from 22, 26, 34 and 38 h represented the stationary phase.

The objective of our analysis is now to select the biomarkers that are differentially expressed between the two phases. We apply our multi-platform integration method to identify differentially expressed biomarkers. For the mRNA data, we formulate the null hypothesis as H_0: the mRNA expression level is the same between the two phases. Similarly, for protein data, the null hypothesis is formulated as H_0 : the protein ratio is the same between the two phases. For both mRNA data and protein data, two-sided alternatives are considered in the analysis. For each platform, we use Behrens-Fisher test statistics to summarize the statistical evidence, which are denoted as t_m and t_p. To obtain the multivariate

Figure 1.2: Decision lines.



Vertical lines use the mRNA data, horizontal lines use the protein data, and dashed lines use our multi-platform integration method.

null distribution, 100 permutations are conducted. The overall correlation between t_m and t_p is 0.2787. The variances of t_m and t_p are 3.0489 and 3.6411, respectively. Based on the decision line constructed at the significance level $\alpha = 0.05$, our method detects 172 differential expressed genes with an estimated $\hat{FP}$ equal to 44. Individual analysis on the mRNA data and the protein data detects 137 and 143 genes, respectively. Figure 1.2 depicts the decision lines for all three comparative analyses: the vertical lines using the mRNA data, the horizontal lines using the protein data, and the dashed lines using our multi-platform integration method.

Nine differentially expressed genes are identified by our method but not by the other two methods. Among these, we identify biosynthetic enzymes (SCO5080 actVA5, SCO5072 actVIORFI) involved in actinorhodin production. These genes are up-regulated only at late stages of the culture and produce antibiotics during the stationary phase. Expression of two genes encoding malate oxidoreductase (SCO2951) and translation elongation factor G (SCO4661) have been found to be depressed during the stationary phase compared with the growth phase (Manteca *et al.*, 2010). Table 1.8 summarizes the nine genes and the associated literature confirmations (Bentley *et al.*, 2002; Mehra *et al.*, 2006; Manteca *et al.*, 2010; Jayapal *et al.*, 2010; Jayapal *et al.*, 2008; Nieselt *et al.*, 2010).

1.4 Conclusion

With the advent of various types of genomic technologies, it is imperative to develop a method that can integrate different types of genomic data to solve biological questions. We develop a general framework for data integration across multiple data platforms. For each data set, a test statistic is formed to summarize the statistic evidence toward the specific null hypothesis tailored to the data platform. The types of test statistics can vary and their marginal distributions can be different. The observed test statistics can then be aggregated across different data platforms. The overall decision is based on the empirical distribution of the aggregated statistic obtained through random permutations. The

Table 1.8: SCO Summaries for the 9 genes which are identified by multi-platform integration method but not by individual platform analysis.

SCO	Sanger Abbreviation	Sanger Annotation	Sanger Category	Sanger Subcategory	TIGR Category	related paper*
SCO1958	uvrA	ABC excision nuclease subunit A	Macromolecule metabolism	DNA-replication, repair, restr./modific'n	excinuclease ABC, A subunit	[1]
SCO2940	other	putative oxidoreductase	Not classified (included putative assignments)	Not classified (included putative assignments)	xanthine dehydrogenase, putative	[1]
SCO2951	other	putative malate oxidoreductase	Central intermediary metabolisms	Other central intermediary metabolism	malate oxidoreductase	[1,3,4]
SCO3094	other	conserved hypothetical protein	hypothetical protein	Conserved in organism other than Escherichia coli	conserved hypothetical protein	[1]
SCO4661	fusA	elongation factor G	Macromolecule metabolism	Proteins - translation and modification	translation elongation factor G	[1,3,4]
SCO5072	actVIORF1	hydroxylacyl-CoA dehydrogenase	Secondary metabolism	PKS	hydroxylacyl-CoA dehydrogenase	[1,3,6]
SCO5080	actVA5	putative hydrolase	Secondary metabolism	PKS	putative hydrolase	[1,5]
SCO6219	Other	putative ATP/GTP binding protein, putative serine	Protein kinases	Serine/threonine		[1]
SCO6222	other	putative aminotransferase	Not classified (included putative assignments)	Not classified (included putative assignments)	aminotransferase, class I	[1, 2]

*1. Bentley et al.(2002). Complete genome sequence of the model actiononomycete Streptomyces coelicolor A3(2), *nature*, 414,141-147;
*2. Jayapal et al.(2008). Uncovering genes with divergent mRNA-Protein dynamics in Streptomyces coelicolor, *Plos One*, 3,e2097;
*3. Jayapal et al.(2010). Multiagging proteomic strategy to estimate protein turnover rates in dynamic systems,*J. Proteome Res.*, 9(5);
*4. Manteca et al.(2010). Quantitative proteomics analysis of Streptomyces coelicolor development demonstrates that onset of secondary metabolism coincides with hypha differentiation,*Mo Cell Proteomics*, 9(7):1423-36;
*5. Mehra et al.(2006). Aframe work to analyze multiple time series data: A case study with Streptomyces coelicolor,*J Ind Mirobio Biotechnol*, 33(2),189-72;
*6. Nieselt et al.(2010). The dynamic architecture of the metabolic switch in Streptomyces coelicolor, *BMC Genomics*,11:10;

symmetric correlation between measurements is required. Our method can accommodate different experimental designs and various data types across platforms. The optimal number of platforms depends on the effect size. The lager effect size is, the less platforms are required, and vice versa. Although including more platforms can increase the power of the method, the cost of the experiments will be increased as well. We need to balance the power and the cost.

2 Combining two chi-squared test statistics.

2.1 Introduction

Mixed data which contain both continuous and discrete data are abundant in scientific research especially in medical or biological studies. An effective clustering method for mixed data should partition a large complex data set into homogeneous subgroups that are manageable in statistical inference. Clustering methods thus have a wide range applications in almost all scientific studies including financial risk analysis, genetic analysis and medical studies. They are essential tools in analyzing large data sets.

Most of the clustering methods in the literature have been mainly focused on either continuous data or categorical data alone. K-means algorithm has been widely used in industrial applications for a long time. Detailed description and discussions can be found in Kaufman and Rousseeuw (2005). Non-Euclidean distances such as Manhattan distance or Mahoblis distance have also been used. Model-based clustering methods for continuous data have been proposed in the literature, see for example Banfield and Raftery (1993).

One of the most prominent methods in parametric clustering based on mixture model is proposed by Bradley *et al.* (1998). The number of clusters and outliers can be handled simultaneously by the mixture model. Fraley and Raftery (1998) propose to choose the number of clusters automatically using model-based clustering method. For clustering categorical data, there are far fewer reliable methods. K-modes algorithm has been proposed by Huang (1997) to extend the K-means to clustering categorical data. AutoClass method proposed by Cheeseman and Stutz (1995) is a well known method in clustering. AutoClass takes a data set containing both real and discrete valued attributes, and automatically computes the number of clusters and group memberships. This method has been used in NASA and helped to find infra-red stars in the IRAS Low Resolution Spectral catalogue and discovery of classes of proteins (Cheeseman and Stutz 1995).

In clustering mixed data, the main difficulty lies in the fact that continuous and categorical sample spaces are intrinsically different. Although both can be made into metric spaces, the continuous sample space resides on a differentiable manifold while the categorical one is defined entirely on a lattice. Attempts have been made in the literature to combine the two spaces by using a global and general distance function (Ahmad and Dey, 2007)). This naive approach ignores the fact that the two sample spaces are topologically incompatible. Alternatively, AutoClass combines information across probability spaces. However, the effectiveness of AutoClass depends on the validity of the assumed parametric model. Zhang *et al.* (2005) showed that both K-modes and AutoClass do not perform very well when applied to benchmark categorical data sets from UCI machine learning depository. Therefore, there is a need for a non-parametric clustering method for mixed data.

We extend the work by Zhang *et al.* (2005) to cluster mixed data by using adaptive quantization of the continuous sample space. The quantization process was developed in 1950's and it partitions the sample space through a discrete valued map (Gersho and Gray, 1992). For univariate case, the quantization is known as the vector quantization and it is the fundamental process for converting analog signals or information into digital forms (Gersho and Gray, 1992). It has been used in studying pricing in finance as well as engineering. Theoretical properties of quantization in probability distributions can be found in Graf and Luschgy (2000). The process of clustering mixed data is then performed on the quantized product space. The key idea is inspired by the fact that any manifold can be locally modelled by a Euclidian space. Therefore, each neighbourhood in the transformed product space can be locally characterized a fine grid endowed with a Hamming Distance. The Hamming Distance is widely used in information and coding theory (Roman,1992; Laboulias *et al.*, 2002). The statistical significance of a detected cluster is determined by a weighted local Chi-squared test. The advantage of our proposed method over AutoClass is demonstrated in simulations and by using two benchmark data sets from UCI machine learning depository.

This chapter of the dissertation is organized as follows. The method is proposed in Section 2.2. The clustering algorithm is presented in Section 2.3. Simulation results are provided in Section 2.4.

2.2 Clustering Methodology

In this section, we introduce quantization of the mixed sample space on which we adopt the Hamming Distance function to measure the relative positions of two data points. We also define a distance vector and an optimal separation point which are essential to measure spatial patterns as well as the size of any detected clusters. Separation points are introduced in order to extract detected cluster patterns.

2.2.1 Joint Sample Space of Mixed Data

Consider a general data structure for a mixed data set with p nominal categorical attributes and q continuous attributes. The categorical sample space is defined on $\Omega_p = R^p$ while the continuous one is defined on Ω_q. The product space for mixed data is then defined on the product space $\Omega_p \otimes \Omega_q$. The sample size is denoted by n.

The categorical part of mixed data is represented by $\mathbf{X} = (X_i^j)$, with $i = 1, 2, \ldots, n$ and $j = 1, \cdots, p$. Furthermore, row and column vectors in the categorical portion are denoted by $\mathbf{X}_i^{[\cdot]}$ and $\mathbf{X}_{[\cdot]}^{j}$. The j^{th} categorical attribute is categorized by m_j levels defined by set $A_j = (a_{j1}, \cdots, a_{jm_j}), j = 1, \cdots, p$.

We denote the continuous part of a mixed sample with size n by $\mathbf{Z} = (Z_i^k)$, with $i = 1, 2, \ldots, n$ and $k = 1, \cdots, q$. Furthermore, we denote the row and column vectors in the categorical portion by $\mathbf{Z}_i^{[\cdot]}$ and $\mathbf{Z}_{[\cdot]}^{k}$. The k^{th} attribute is a continuous random variable.

2.2.2 Quantization of Continuous Sample Space

Continuous data and discrete data are fundamentally different. Although the description provided by the continuous portion can be very detailed, they could carry excessive information that are not important for the clustering purpose. Furthermore, any pattern derived from the categorical part is based on a much coarse topology than the continuous counterpart. Since it is impossible to define a meaningful and objective manifold from a coarse data structure, the continuous one then must be mapped into a grid that is compatible with the relatively coarse topology from the categorical one.

The quantization is achieved in two steps. Firstly for observed realization z_i^j, continuous data are mapped onto the unit interval between 0 to 1 by applying the following formula:

$$\tilde{z}_i^k = \frac{z_i^k - z_{min}^k}{z_{max}^k - z_{min}^k}, \quad k = 1, ..., q; i = 1, ..., n$$

where z_{min}^k and z_{max}^k represent the minimum and maximum values of k column. Secondly, for the standardized observations, the continuous random variable is then mapped or quantized into a discrete random variable with M levels by following way:

$$Q(\tilde{z}_i^k) = m, \quad if \quad (m-1)/M \leq \tilde{z}_i^k < m/M$$

where $m = 1, 2, \cdots, M$, where M can be any positive integer value. Different numerical value of M could have impact on the quality of quantization and consequently the clustering result. Finer quantization grid might not be useful and could be more computationally intensive than a coarse one.

The number of levels M can be difficult to specify by a user with no prior information. Thus we propose to choose the level M adaptively by using F statistics based on the clustering results. For any fixed value of M that are reasonable, clustering memberships will then be used to perform ANOVA test by partioning the data into individual groups from which the F-statistic can be derived accordingly. The numerical value of a quantization which generates the largest value among calculated F-statistics is then selected as the appropriate number needed for quantization. Numerical results of quantization level will be illustrated in Section 2.4.1

2.2.3 Distance Vectors on Quantized Product Space

We use Hamming Distance (HD) to measure the relative separation of two categorical data points. To be more specific, for any two positions in the categorical sample space Ω_p,

$\mathbf{Q}_h^{[\cdot]} = (Q_h^{[1]}, \cdots, Q_h^{[p]})$ and $\mathbf{Q}_i^{[\cdot]} = (Q_i^{[1]}, \cdots, Q_i^{[p]})$, the HD between $Q_h^{[j]}$ and $Q_i^{[j]}$ on the jth attribute is

$$d(Q_h^j, Q_i^j) = \begin{cases} 0 \quad if \quad Q_h^j = Q_i^j, \\ \\ 1 \quad if \quad Q_h^j \neq Q_i^j; \end{cases}$$

Further, we define the distance between the two positions, that is, the summation of distance from each pair of the components. Therefore, we have the following:

$$HD(\mathbf{Q}_h^{[\cdot]}, \mathbf{Q}_i^{[\cdot]}) = \sum_{j=1}^{p} d(\mathbf{Q}_h^j, \mathbf{Q}_i^j).$$

After quantization, the new product space now resides on a high dimensional grid. Since for a grid, there is no natural origin. We can define a reference point $(\mathbf{S}, \mathbf{T})$ in the quantized product space with $\mathbf{S} = (s_1, \cdots, s_p) \in R^p$ and $\mathbf{T} = (t_1, \cdots, t_q) \in R^q$. For the categorical portion, $HD_C(\mathbf{X}_i, \mathbf{S})$ can take values ranging from 0 to p; and for quantized continuous data, we have $HD_Q(\mathbf{Z}_i, \mathbf{T})$ can take values ranging from 0 to q.

We then define the Distance Vector (DV) based on Hamming distance for the categorical and quantized continuous portion, respectively. We define two individual vectors to record the frequencies of each categorical and quantized continuous distance value accordingly, that is, a $(p + 1)$-element vector $DV_C(\mathbf{S})$ for categorical data and a $(q + 1)$-element vector $DV_Q(\mathbf{T})$ for quantized part. To be more specific, DV_C is defined as

$$DV_C(\mathbf{S}) = (DV_C^{[0]}(\mathbf{S}), DV_C^{[1]}(\mathbf{S}), \cdots, DV_C^{[p]}(\mathbf{S}))$$

and DV_Q is defined as

$$DV_Q(\mathbf{T}) = (DV_Q^{[0]}(\mathbf{T}), DV_Q^{[1]}(\mathbf{T}), \cdots, DV_Q^{[q]}(\mathbf{T})).$$

The j^{th} component in DV_C and h^{th} component in DV_Q are given as the following:

$$DV_C^{[j]}(S) = \sum_{i=1}^{n} \boldsymbol{I}\, [HD_C(\mathbf{X}_i^{[\cdot]}, \mathbf{S}) = j], \quad j = 0, 1, \cdots p;$$

$$DV_Q^{[h]}(T) = \sum_{i=1}^{n} \boldsymbol{I}\, [HD_Q(\mathbf{Q}_i^{[\cdot]}, \mathbf{T}) = h], \quad h = 0, 1, \cdots q;$$

where $\boldsymbol{I}(A)$ is the indicator function that takes value 1 when event A happens and 0 otherwise.

If there is no cluster pattern at all, we would expect a uniform distribution of all possible cases. Then it is equally likely for a randomly chosen data point to take any possible position in the joint sample space. The DV vectors under uniform distribution are referred as *uniform* distance vector (UDV). Thus, a UDV records the expected frequencies under the null hypothesis that there are no clustering patterns in data. Let $\mathbf{X}$ be a categorical portion of data and $\mathbf{Z}$ be a continuous portion of the data from a sample of size n, with each observation having an equal probability of locating at any position on space $\Omega_p \otimes \Omega_q$. The expected value of DV and DV associated with the null hypothesis are denoted by UDV_C, $\mathbf{U} = (U_0, \cdots, U_p)$ for categorical data and UDV_Q, $\mathbf{V} = (V_0, \cdots, V_q)$ for continuous data, respectively.

Zhang *et al.* (2005) provides the exact form of $UDV_C = \frac{n}{M_1}\mathbf{U}^*$, where $M_1 = \prod_{j=1}^{p} m_j$, $j = 1, 2, \cdots, p$; m_j is the number of states in set A_j for the j^{th} attribute; and $\mathbf{U}^* = (U_0^*, U_1^*, \cdots, U_p^*)$ with

$$
\begin{aligned}
&U_0^* = 1;\\
&U_1^* = (m_1 - 1) + (m_2 - 1) + \cdots + (m_p - 1);\\
&U_2^* = \textstyle\sum_{i<j}^{p}(m_i - 1)(m_j - 1);\\
&\vdots\\
&U_p^* = (m_1 - 1)(m_2 - 1)\cdots(m_p - 1).
\end{aligned}
$$

Similarly, we obtain the exact form of the UDV_Q for the quantized continuous part of data. $UDV_Q = \frac{N}{M_2}\mathbf{V}^*$, where $M_2 = \prod_{j=1}^{q} l_j$, $j = 1, 2, \cdots, q$; l_j is the the number of levels of quantization for the j^{th} continuous attribute; and $\mathbf{V}^* = (V_0^*, V_1^*, \cdots, V_q^*)$ with

$$
\begin{aligned}
&V_0^* = 1;\\
&V_1^* = (l_1 - 1) + (l_2 - 1) + \cdots + (l_q - 1);\\
&V_2^* = \textstyle\sum_{i<j}^{q}(l_i - 1)(l_j - 1);\\
&\vdots\\
&V_q^* = (l_1 - 1)(l_2 - 1)\cdots(l_p - 1).
\end{aligned}
$$

2.2.4 Optimal Separation Point

If the initial starting point is chosen to be the center of one particular cluster, then the frequency of HD should demonstrate a decreasing pattern in a local region as the HD function records the frequency of data points from the center of cluster and outwards. Small local bumps at the beginning part of the HD curve are expected if the initial starting point deviate slightly from the cluster center. The recorded frequencies might increase afterwards when the function begins to record distances from another cluster. Therefore, the valley area indicates a natural places to separate one cluster from the rest. Separation points are, therefore, defined for this identification purpose.

Assume that the categorical data **X** and quantized continuous data **Z** are not uniformly distributed in the sample space $\Omega_p \otimes \Omega_q$. Let $DV_C(\mathbf{S}) = (DV_C^{[0]}(\mathbf{S}), DV_C^{[1]}(\mathbf{S}), \cdots, DV_C^{[p]}(\mathbf{S}))^T$, $\mathbf{S} \in \Omega_p$ be the collection of all $(p+1)$-element DV_C in the space Ω_p and $DV_Q(\mathbf{T}) = (DV_Q^{[0]}(\mathbf{T}), DV_Q^{[1]}(\mathbf{T}), \cdots, DV_Q^{[q]}(\mathbf{T}))^T$, $\mathbf{T} \in \Omega_q$ be the collection of all $(q+1)$-element DV_Q in the space Ω_q, and let $\mathbf{U} = (U_0, U_1, \cdots, U_p)^T$ be the DV_C vector and $\mathbf{V} = (V_0, V_1, \cdots, V_q)^T$ be the DV_Q vector defined in the previous subsection. For a given distance value $j_C, j_C = 0, 1, \cdots, p$, for categorical distance values and $j_Q, j_Q = 0, 1, \cdots, q$, for quantized continuous distance values, there always exists at least one position **(S,T)** $\in \Omega_p \otimes \Omega_q$, such that the frequency at this distance value is lager than the corresponding component, U_j of the UDV_C vector and V_j of the UDV_Q vector.

In order to proceed to a comparison between DV_C and UDV_C and between DV_Q and UDV_C, we introduce a selection criterion for an optimal cut-off r^*. The categorical cut-off point was defined and proved by Zhang *et al.* (2005). Because our quantized continuous data behaves as categorical data, we extend that concept to quantized portion of the data. If the cluster structure is present, the early segment of an DV_C and DV_Q with respect to a data center should contain substantially larger frequencies than the corresponding frequencies of the UDV_C vector and UDV_Q vector. Therefore, the range corresponding frequencies of the UDV_V vector and UDV_Q vector that are consistently larger than the UDV_C vector and UDV_Q vector gives a reasonable indication of the r. This leads to an optimal r_C^* for categorical portion of data:

$$r_C^*(\mathbf{S}) = \min_{j_C>0}\{j_C | \frac{DV_C^{[j_C]}(\mathbf{S})}{U_{j_C}} < 1\} - 1, \mathbf{S} \in \Omega_p$$

Similarly, optimal r_Q^* for quantized portion of data be:

$$r_Q^*(\mathbf{T}) = \min_{j_Q>1}\{j_Q | \frac{DV_Q^{[j_Q]}(\mathbf{T})}{V_{j_Q}} < 1\} - 1, \mathbf{T} \in \Omega_q$$

2.3 Algorithm

There are two key parts of the algorithm. Firstly, we detect whether there exists any statistically significant clustering patterns. We propose a weighted local Chi-squared test to determine if the observed distance vectors differ significantly from the uniform distance vectors associated with no cluster pattern. Secondly, if the patterns are significant, we further extract the clusters based on the optimal separation strategies described in the previous section.

We consider the null hypothesis H_0: There is no clustering pattern in data set. The weighted local Chi-squared test statistic $\chi_w^{2*}(\mathbf{S,T})$ is defined as:

$$\chi_w^{2*}(\mathbf{S},\mathbf{T}) = \frac{pq}{p+q}\frac{1}{p}\chi_C^{2*}(\mathbf{S}) + \frac{pq}{p+q}\frac{1}{q}\chi_Q^{2*}(\mathbf{T}), \quad (\mathbf{S},\mathbf{T}) \in \Omega_{p\otimes q}$$

where the categorical part $\chi_C^{2*}(\mathbf{S})$ takes form as:

$$\chi_C^{2*}(\mathbf{S}) = \sum_{j=0}^{r_C^*} \frac{(DV_C^{[j]}(\mathbf{S}) - U_j)^2}{U_j} + \frac{(\sum_{j=0}^{r_C^*} DV_C^{[j]}(\mathbf{S}) - \sum_{j=0}^{r_C^*} U_j)^2}{\sum_{j=r_C^*+1}^{p} U_j} \tag{2.1}$$

and the quantized continuous part $\chi_Q^{2*}(\mathbf{T})$ takes the form:

$$\chi_Q^{2*}(\mathbf{T}) = \sum_{j=1}^{r_Q^*} \frac{(DV_Q^{[j]}(\mathbf{T}) - V_j)^2}{V_j} + \frac{(\sum_{j=1}^{r_Q^*} DV_Q^{[j]}(\mathbf{T}) - \sum_{j=1}^{r_Q^*} V_j)^2}{\sum_{j=r_Q^*+1}^{q} V_j}$$

where p and q are number of attributes from categorical and continuous data, respectively.

If the detected pattern passes a statistical test, we then proceed to extract a cluster by determining the cluster center $\boldsymbol{C}$ and estimate cluster radius $\boldsymbol{R}$ for mixed data. Therefore, a cluster center $\mathbf{C}$ is chosen where the χ_w^2 has the maximum value. It is chosen to be:

$$\mathbf{C} = \arg\max_{(\mathbf{S},\mathbf{T})} \chi_w^2$$

Zhang *et al.* (2005) gave the definition of radius which is the maximum distance of the data points in this cluster to its center. Radius is the distance at which the DV has its very first local minimum. Therefore, it is defined categorical Radius $R_C(\mathbf{C})$ as:

$$R_C(\mathbf{C}) = \underset{0<j<p_C}{\mathrm{mim}} \{j | DV_{C_j}(\mathbf{C}) < \mathrm{mim}(DV_{C_{j-1}}(\mathbf{C}), DV_{C_{j+1}}(\mathbf{C}))\} - 1;$$

For quantized continuous part of the data, the optimal cut-off point is used as quantized continuous radius $R_Q(\mathbf{C})$.

The step-by-step guide to our method is

Step 1. For each position $\boldsymbol{S}$, we calculate HD in the categorical data; further, we obtain DV_C.

Step 2. Standardize the continuous data and quantize the standardized data at a selected level. For each position calculate Hamming distance for quantized continuous data to obtain DV_Q.

Step 3. Compare DV_C, DV_Q with corresponding expected values UDV_C and UDV_Q;

Step 4. Determine cut-off point $r_C^*(\mathbf{S})$ and $r_Q^*(\mathbf{T})$ for categorical and quantized continuous data respectively; and further calculate the corresponding modified Chi-squared statistic $\chi_C^{2*}(\mathbf{S})$ and $\chi_Q^{2*}(\mathbf{T})$ and obtain the weighted local chi-square test statistic

$$\chi_w^{2*}(\mathbf{S,T}) = \frac{q}{p+q}\chi_C^{2*}(\mathbf{S}) + \frac{p}{p+q}\chi_Q^{2*}(\mathbf{T});$$

Step 5. Corresponding to the weighted local Chi-squared test, select the largest test statistic $\chi_w^{2*}(\mathbf{S,T})$; compare it with critical value $\chi_{(0.05)}^{2*}$ at right tail. If the $\max(\chi_w^{2*}(\mathbf{S,T}))$ is smaller than $\chi_{(0.05)}^{2*}$, stop the algorithm; otherwise, continue to step 6;

Step 6. Assign the position who has the largest test statistic $\chi_w^{2*}(\mathbf{S,T})$ as a center. Categorical data and continuous data share the same center position but with their own data points;

Step 7. Calculate categorical radius R_C and continuous radius R_Q; label all data points within radius in the cluster; recorder corresponding $\chi_C^{2*}(\mathbf{S})$ and $\chi_Q^{2*}(\mathbf{T})$; remove them from the current data set;

Step 8. Repeat Step 1 to 6 until no more significant clusters are detected.

Step 9. Prune the membership assignment by calculating the minimum distance from each data point to center positions; If the membership is assigned differently to categorical data and continuous data, we further compare their p-values which are calculated from $\chi_C^{2*}(S)$ and $\chi_Q^{2*}(S)$; Re-assign the membership to the one with the larger p-value by the one with the smaller p-value.

Step 10. Compute F test statistic to choose the best quantized level and corresponding clustering results as the final results.

2.4 Numerical Results

We conduct simulation studies and real data analysis to examine the performance of our proposed method. Classification rates and information gains are calculated to compare the performance from our proposed method with AutoClass.

2.4.1 Simulation Studies

In this section, we compare our method with AutoClass under various simulation settings. The simulation results are shown in Tables 2.1 - 2.4. All attributes are generated independently. The simulation setting is as the following:

1. Set the number of categorical attributes $p = 10$ and each attribute takes m_j levels which is randomly selected from the set $\{4, 5, 6\}$; Set the number of continuous attributes $q = 9$.

2. Set the number of clusters $K_C = K_Q = 3$ or $K_C = K_Q = 5$. The 3 cluster centers $\mathbf{C}_k$ are denoted as $C_k = (c_{k,1}, \cdots, c_{k,10})$, $k = 1, \cdots, 3$. The 5 cluster centers $\mathbf{C}_k$ are denoted as $C_k = (c_{k,1}, \cdots, c_{k,10})$, $k = 1, \cdots, 3$. For categorical centers, ensure the Hamming distance between any two of the centers are at least great than 5. For the continuous portion of data, choose a set of cluster mean as 2, 8, and 16 for 3 clusters, or 2, 8, 16, 20 and 35 for 5 clusters;

3. Set sample size $N = 200$ with cluster size $n_1 = 130$, $n_2 = 45$, and $n_3 = 25$; or set sample size $N = 100$ with the cluster size $n_1 = 40$, $n_2 = 25$, $n_3 = 15$, $n_4 = 10$, and $n_5 = 10$; or set sample size $N = 1000$ with the cluster size $n_1 = 500$, $n_2 = 200$, $n_3 = 100$, $n_4 = 100$, and $n_5 = 100$;

4. For categorical data, in the k^{th} cluster with center $\mathbf{C}_k$, generate n_k 10-attributes vec-

tors independently. More specifically, generate for each attribute from a multinomial distribution with center probability 0.7 and the rest probabilities are identically equal to $0.3/(m_j - 1)$; For continuous data, n_k 9-attributes vectors are 9 independent normal random variables with $\boldsymbol{\mu} = \mathbf{C}_k$ and σ^2 ranging from $0.25, 0.5$ and 1, respectively.

In our numerical results, average classification rate (CR) and information gain (IG) rate with their corresponding standard deviations are used to evaluate methods performance. The CR measures the accuracy of an algorithm to assign data points into correct clusters. With given K clusters, the CR is defined by

$$CR(K) = \sum_{k=1}^{K} \frac{\tilde{n}_k}{n},$$

where n is total number of data points and $\tilde{n}_k$ is the number of data points that have been correctly assigned to cluster k by an algorithm. Obviously, $0 \leq CR(K) \leq 1$, and a larger $CR(K)$ value indicates better performance of clustering. The information gain is an alternative criterion for assessing the performance of clustering algorithm. It is so-called cluster purity proposed by Bradley *et al.* (1998). Cluster purity essentially measures the information gain, which is the difference between the total entropy and weighted entropy for a given data partition, namely

$$information\ gain(IG(K)) = total\ entropy - weighted\ entropy(K),$$

where the weighted entropy is calculated by

$$weighted\ entropy(K) = \sum_{k=1}^{K} \frac{n_k}{n} \times cluster\ entropy(k),$$

with

$$cluster\ entropy = -\sum_{l=1}^{L} \frac{\tilde{n}_l^k}{n_k} \log_2 \left\{ \frac{\tilde{n}_l^k}{n_k} \right\},$$

where $\tilde{n}_l^k$ is the number of data points with true label l in cluster k, n_k is the number of data points known in cluster k, and L is the known number of classes. In this chapter, we take a ration of IG(K)/total entropy, named information gain rate (IGR), which is similar to the classification rate between 0 to 1. It is necessary to point out that in some situations, the information gain may lead to misleading. For example, in our simulation studies, IG may be equal to 1 which means perfect clustering. But, in fact, it splits each true cluster into two clusters which is obviously a wrong classification. This misleading situation happens in Table 2.2 and 2.2

Table 2.1 shows the selection of quantization levels for continuous portion of the data. As mentioned in section 2.2.2, we use the largest F values to choose the selected quantization level which gives the best classification rate. Table 2.2 to Table 2.4 provide results from simulated data with various settings of different sample sizes, number of clusters and cluster sizes. The number of replications is 500. Table 2.2 is obtained by analyzing simulated data with a sample size of 200 with 3 clusters of the sizes of 130, 45 and 25. Simulated data for Table 2.3 has sample size 1000 and number of clusters is 5, and each cluster size is 500, 200, 100, 100 and 100, respectively. Table 2.4 provides results from simulated data having sample size 10000 with 3 clusters and each cluster size 5500, 3000, and 1500, respectively.

As shown by Table 2.2 to 2.4, our proposed algorithm consistently has higher classification rate in comparison with that from AutoClass in all three different settings. For the three chosen settings, the mean classification rates and information gain rates of the two algorithms are getting closer to each other and could even be identical. Table 2.3 shows us that our algorithm has higher IG rates comparing to AutoClass. In Table 2.2 and 2.4, our algorithm has IG rates varying from 0.8923 to 0.93333. Although AutoClass could achieve one in some cases, this does not imply a perfect clustering due to the fact that AutoClass tends to split each true cluster into unnecessary more clusters. Hence, overall, all tables show us that our algorithm has better performance in terms of CR and IGR by comparing to AutoClass. The variances of classification rates and information rates of our algorithm decreases when the sample sizes increases. This is expected since the accuracy should increase with the sample size. The same pattern, however, is not observed for the AutoClass.

2.4.2 Real Data Analysis

We applied our method on to two real data sets. Both dat sets are Machine Learning Repository website. One is Heart Data Set and the other one is Australian Credit Approval Data Set. All these data sets are download form Machine Learning Depository at the University of California at Irvine. Heart data contains 7 categorical, 6 continuous attributes and 270 observations. The data provided the memberships for each observation. There are 2 clusters, absence or presence. The cluster sizes are 120 and 150, respectively. In Australian Credit Approval Data Set, there are 8 categorical attributes and 6 continuous attributes. The data set contains 2 clusters positive or negative with corresponding cluster size 307 and 383. We compared our method with AutoClass. Table 2.5 shows the results from these two real data sets. From the table, we can tell that our method correctly identified the number of clusters for both data sets, while, AutoClass couldn't detect correct cluster numbers. In addition, our method has higher classification rate comparing to AutoClass. Our method has classification rate 81.48% for Heart data and 73.62% for Credit data. But, AutoClass has 44.44% and 52.71%.

2.5 Conclusion

Mixed data are prolific in scientific research such as in business, engineering, life sciences and so on. It is imperative to develop a method that can cluster mixed data in order to discover true and significant underlying structures of a dataset and classify observations into different subsets. We propose a non-parametric method that uses a local weighted chi-squared statistic to determine underlying clusters. The proposed algorithm does not require any model assumption for attributes or any expensive numerical optimization procedures. Because the proposed algorithm extracts clusters sequentially with one cluster at each iteration, it does not need any convergence criterion. The algorithm is terminated when all data points have been used and no more cluster center can be detected. Consequently our algorithm automatically produce the number of clusters, and the resulting partition is unique. When compared with benchmark clustering algorithm for mixed data, AutoClass, we find that our algorithm out-performs AutoClass in various settings and produce similar accuracy in other settings.

Table 2.1: Quantization levels. The means of F statistics, CR and IG are obtained based on 500 replications.

Discretized Levels	Mean(F)	Mean(CR)	Mean(IGR)
5	630.1573	0.8302	0.7130
6	1523.4557	0.8455	0.7667
7	1722.3260	0.8227	0.6960
8	3223.9477	0.8635	0.7729
9	3916.3388	0.8816	0.7958
10	3708.5293	0.8682	0.7689
11	***6444.7055***	***0.9085***	***0.8573***
12	4778.9851	0.8893	0.8114
13	4912.8477	0.8907	0.8116
14	4262.3990	0.8907	0.8135
15	4000.3948	0.8879	0.8095
16	4234.9993	0.8863	0.7992
17	3549.8632	0.8787	0.7853
18	4042.0805	0.8785	0.7833
19	3657.4556	0.8768	0.7785
20	4303.8698	0.8872	0.8010

Table 2.2: Average CR and IGR with corresponding standard deviation for each method based on the simulated data of sample size 200 with 3 clusters; each cluster has size 130, 45 and 25, respectively. The mean values for each cluster are 2, 8 and 16, respectively. The number of replications is 500.

	AutoClass	Ours	AutoClass	Ours	AutoClass	Ours
	(Var=0.25)		(Var=0.5)		(Var=1)	
CR Mean	0.6424	0.9556	0.6335	0.9292	0.6325	0.9370
CR Std	0.0021	0.0035	0.0015	0.0069	0.0015	0.0060
IGR Mean	1.0000	0.8923	1.0000	0.9085	1.0000	0.9148
IGR Std	<0.0001	0.0148	<0.0001	0.0094	<0.0001	0.0070

Table 2.3: Average CR and IGR with corresponding standard deviation for each method based on the simulated data of sample size 1000 with 5 clusters; each cluster has size 500, 200, 100, 100 and 100, respectively. The mean values for each cluster are 2, 8, 16,20 and 35, respectively. The number of replications is 500.

	AutoClass	Our	AutoClass	Ours	AutoClass	Ours
	(Var=0.25)		(Var=0.5)		(Var=1)	
CR Mean	0.5638	0.8747	0.5598	0.8792	0.5615	0.8777
CR Std	0.0016	0.0185	0.0015	0.0179	0.0014	0.0189
IGR Mean	0.7337	0.9228	0.7338	0.9174	0.7338	0.9235
IGR Std	<0.0001	0.0021	<0.0001	0.0049	<0.0001	0.0037

Table 2.4: Average CR and IGR with corresponding standard deviation for each method based on the simulated data of sample size 10000 with 3 clusters; each cluster has size 5500, 3000 and 1500, respectively. Continuous data are from multivariate t distribution with degree freedom 5, 15 and 30, respectively. The number of replications is 100.

	AutoClass	Ours	AutoClass	Ours	AutoClass	Ours
	(Var=0.25)		(Var=0.5)		(Var=1)	
CR Mean	0.8120	0.9689	0.8231	0.9689	0.8202	0.9641
CR Std	0.0019	0.0031	0.0023	0.0031	0.0033	0.0034
IGR Mean	1.0000	0.9333	1.0000	0.9333	1.0000	0.9323
IGR Std	<0.0001	0.0067	<0.0001	0.0067	<0.0001	0.0048

Table 2.5: Two Real Data Results from two comparison methods. Heart data has 2 clusters with sample size 270 and Australian data has 2 clusters with sample size 690.

	Heart		Australian	
	AutoClass	Ours	AutoClass	Ours
CR	0.4444	0.8148	0.5217	0.7362
IGR	0.2754	0.6975	0.2761	0.8314
Number of clusters	5	2	7	2

3 Weighted integrative AICs criterion for model selection

3.1 Introduction and Literature Review

Models are essential in statistical analysis. Once a model has been established, various forms of inference, such as information extraction, model validation, risk assessment and prediction can be performed. Due to model uncertainty, a true model is often out of reach. Therefore, we have to choose an approximate model in order to conduct statistical inference. How to choose a suitable approximate model among a class of competing models by suitable model evaluation criteria become a crucial issue. Akaike's entropy information criterion (AIC) (Akaike, 1973) is one of the commonly used model evaluation criteria. AIC selects the best model based on information containing one single data set. However, information could come from multiple data sets that are too different to be merged into one. How to effectively perform model selection by integrating information from different data sets is main focus of this chapter.

We propose a weighted integrative AICs as a model evaluation criterion. Our proposed method combines AICs across multiple data sets with different weights that minimize the variance of the integrative AICs. Simulation studies show that, in the context in variable selection, our proposed method has the lowest false negative numbers and false detected numbers comparing with individual test and equal weights combining test.

3.1.1 Kullback-Leibler(K-L) divergence and AIC

We first review Kullback-Leibler (K-L) (Kullback and Leibler, 1951) divergence and AIC.

Let $X_1, \cdots, X_n$ be identically independent distributed from unknown true probability distribution function f and denote $X = (X_1, \cdots, X_n)$. Let $g(x;\theta)$ be a specified model with parameter θ. The validity of an assumed model must be assessed in term of its closeness to true probability distribution f. The best model is then chosen to be the probability density function that minimizes a chosen divergence function defined in the functional space of probability density functions. K-L divergence is widely used in model selection and it is defined as the following

$$\begin{aligned} I(f,g) &= E_X\left[\log\left\{\frac{f(X)}{g(X;\theta)}\right\}\right] \\ &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{f(x)}{g(x;\theta)}\right)\mathrm{d}x \\ &= E_X\log(f(X)) - E_X\log(g(X;\theta)), \end{aligned}$$

where E_X is the expectation with respect to the true probability distribution function f. The first expectation is constant for any given $f(x)$, and the second expectation determines the goodness-of-fit of $g(x;\theta)$ with respect to $f(x)$. We can re-write the above equation to

$$I(f,g) - \text{constant} = -E_X\left[\log(g(X;\theta))\right].$$

Properties of $I(f;g)$ include:

(i). $I(f,g) \geq 0$;

(ii). $I(f,g) = 0$, if and only if $f(x) = g(x)$;

(iii). if $X_1, \cdots, X_n$ are independent and identically distributed random variables, then the K-L divergence is additive for the whole sample i.e. $I_n(f,g) = nI(f,g)$.

Akaike (1973) proposes a model selection criterion based on K-L divergence theory. In reality, K-L divergence is not directly observable or estimable because it depends on the true distribution and consequently on the unknown true parameter. But the expected K-L divergence $E_X[I(f, g(X'|\hat{\theta}(X)))]$ can be estimated, where X and X' are both generated from $f(x)$, i.e. X' is a future copy of current data X, and $\hat{\theta}(X)$ estimates θ based on X. Note that the expectation is taken with respect to the true probability density function f of observations X'. Let $\theta_* = \arg\min_\theta E_X[\log(g(X;\theta))]$ and $\hat{\theta}$ be the maximum likelihood estimator (MLE) using likelihood defined by $g(X;\theta)$. Let $L(\hat{\theta};X)$ denote likelihood based on $g(X;\theta)$. An asymptotically unbiased estimator of expected K-L divergence is $\log(L(\hat{\theta}; X)) + \mathrm{tr}(J(\theta_*)H(\theta_*)^{-1})$, where $J(\theta_*) = E_X\left[\left(\frac{\partial \log(g(X;\theta_*))}{\partial\theta_*}\right)\left(\frac{\partial \log(g(X;\theta_*))}{\partial\theta_*}\right)'\right]$ and $H(\theta_*) = E_X\left[\frac{\partial^2 \log(g(X;\theta_*))}{\partial\theta_*\partial\theta_*'}\right]$ (Takeuchi, 1976). When the model is correctly specified, i.e. $g \equiv f$, then $J(\theta_*) = -H(\theta_*)$, $\mathrm{tr}(J(\theta_*)H(\theta_*)^{-1}) = -p$, where p is the number of estimable parameters in the model g. Akaike (1973) then defined an information criterion, named AIC, multiplying the estimated expected K-L divergence by -2,

$$AIC = -2\log(L(\hat{\theta}|X)) + 2p.$$

AIC model selection procedure selects the model with the smallest AIC value because this model is estimated to be closest to the unknown true model.

In more general cases, the equality of J matrix and H matrix doesn't hold, i.e. $J(\theta_*) \neq H(\theta_*)$. Takeuchi (1976) proposed a robust AIC, which is known as Takeuchi Information Criterion (TIC):

$$TIC = -2\log(L(\hat{\theta}|X)) + 2\mathrm{tr}(\hat{J}(\theta)\hat{H}(\theta)^{-1}),$$

where $\hat{J}(\theta)$ and $\hat{H}(\theta)$ are consistent estimators for$J(\theta_*)$ and $H(\theta_*)$, respectively. Stone (1977) and Shibata (1989) showed that the TIC is asymptotically equivalent to the cross-validation.

Based on K-L divergence and AIC, Varin and Vidoni (2005) introduced an information criterion for model selection based on composite likelihood, which is the extension of TIC. Varin's composite likelihood information criterion selects the model maximising $\log(L_C(\hat{\theta}_{MCL}|X)) + \mathrm{tr}(\hat{J}(\theta)\hat{H}(\theta)^{-1})$, where $L_C(\hat{\theta}_{MCL}|X)$ is composite likelihood, $\hat{\theta}_{MCL}$ is defined as a solution to the composite likelihood equation, $\hat{J}(\theta)$ and $\hat{H}(\theta)$ are consistent, first-order unbiased estimator for $J(\theta_*)$ and $H(\theta_*)$.

There are many variants information criteria based on AIC, such as AICc (Hurvich and Tsai, 1989), which is a modified AIC with a second order correction for small sample sizes; QAIC and QAICc, Quasi-likelihood modification to AIC and AICc (Lebreton *et al.*, 1992) and so on. However, all these existing information criteria compare models by using single data set. In reality, data can come from several different data sets. How to efficiently combine information criteria to perform the model selection procedures across different data sets is of interest to us.

We propose a model evaluation criterion based on weighted integrative AICs. Our proposed method combines AICs with different weights across multiple data sets. The weights are chosen to minimize the variance of the integrative AICs. Our simulation studies show that, comparing with individual test and equal weights combining criterion, our criterion has better performance in terms of false detected numbers and false negative numbers.

The next section illustrates the developed method. Simulation results are shown in Section 3.3.

3.2 Method

The aim of our weighted integrative AICs method is to select the best model among competing models across multiple data sets. If there exists Q different data sets, our proposed information criterion integrates a set of Q AICs across multiple data sets with the weights chosen to minimize the variance of the integrated AICs. For simplicity, in the following, we demonstrate our method by considering two independent data sets that were generated using the same model with different numeric values for the coefficients.

3.2.1 Integrative AICs

Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be *i.i.d.* from same family of unknown true density functions $f(X;\phi_1)$ and $f(Y;\phi_2)$, respectively, and $\phi_1 \neq \phi_2$. Denote $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$. We consider a family of density functions $g(\cdot;\theta)$ to approximate the true density $f(\cdot;\phi)$. Let the approximated density function for X be $g(X;\theta_1)$ and for Y be $g(Y;\theta_2)$. We would like to choose the model which offers the most satisfactory predictive description of the observed data X and Y. To be more precise, if X' and Y' are future random variables, defined as independent copies of X and Y, we are interested in the choice of best model for forecasting X' and Y', as a realization of X and Y. As usual for an information criterion, model selection can be approached on the basis of the expected K-L divergence.

Consider a weighted linear combination of K-L divergence from Q data sets

$$I_Q(f,g) = \sum_{q=1}^{Q} w_q I_q(f, g(\cdot;\theta_q)),$$

where w_q denotes assigned weights. In our illustration example with two data sets X and Y, the weighted integrative K-L divergence is written as follows,

$$\begin{aligned} I(f,g) &= \sum_{q=1}^{2} w_q I_q(f,g) \\ &= \text{Constant} - \{w_1 E_X \log(g(X|\theta_1)) + w_2 E_Y \log(g(Y|\theta_2)\}. \end{aligned}$$

The $I(f,g)$ is not available because g has to rely estimate based on current data X and Y. As usual for an information criterion, model selection can be approached on the basis of the expected K-L divergence between the true densities $f(X)$ and $f(Y)$ and estimated densities $g(X', \hat{\theta}(X))$ and $g(Y', \hat{\theta}(Y))$. Let $\varphi_X = E_X E_{X'} \log(g(X'|\hat{\theta}(X)))$, $\varphi_Y = E_Y E_{Y'} \log(g(Y'|\hat{\theta}(Y)))$, and $\varphi(f,g) = w_1\varphi_X + w_2\varphi_Y$. We select the model with minimise $w_1 E_X[I(f, g(\hat{\theta}(X))] + w_2 E_Y[I(f, g(\hat{\theta}(Y))]$ or, equivalently, which maximises

$$\begin{aligned} \varphi(f,g) &= w_1\varphi_X + w_2\varphi_Y \\ &= w_1 E_X E_{X'} \log(g(X'|\hat{\theta}_1)) + w_2 E_Y E_{Y'} \log(g(Y'|\hat{\theta}_2)). \end{aligned}$$

The above equation defines a theoretical criterion to select the best predictive model. However, it requires the knowledge of the unknown true densities. In practice we should maximize a selection statistic $\hat{\varphi}(f,g)$, defined as a suitable estimator for $\varphi(f,g)$ based on X and

Y. Denote $\ell(\hat{\theta}_1; X) = \log L(\hat{\theta}_1; X)$ and $\ell(\hat{\theta}_2; Y) = \log L(\hat{\theta}_2; Y)$. We look for estimators that are unbiased. A natural estimator is

$$\ell(\hat{\theta}_1, \hat{\theta}_2; X, Y) = w_1 \ell(\hat{\theta}_1; X) + w_2 \ell(\hat{\theta}_2; Y)$$

In the following Lemmas, we show that $\ell(\hat{\theta}_1, \hat{\theta}_2; X, Y)$ is biased and we introduce a modification with corrects the bias.

We state several regularity assumptions as follows:

Assumption 3.2.1. *The parameter space Θ_1 and Θ_2 are compact subsets of $\mathbb{R}^p (p \geq 1)$ and, for every fixed x and y, $L(\theta_1; x)$ and $L(\theta_2, y)$ are twice differentiable with respect to θ_1 and θ_2, respectively.*

Assumption 3.2.2. *The maximum likelihood estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are defined as solutions to the likelihood equations and there exists vector $\theta_{1*}, \theta_{2*} \in \text{int}(\Theta)$, such that, exactly or with an error term that is negligible as n goes to infinity, $E_X \left[\frac{\partial \ell(\theta_{1*}; X)}{\partial \theta_{1*}} \right] = 0$ and $E_Y \left[\frac{\partial \ell(\theta_{2*}; Y)}{\partial \theta_{2*}} \right] = 0$*

Assumption 3.2.3. *The estimator $\hat{\theta}_1$ and $\hat{\theta}_2$ converge in probability to θ_{1*} and θ_{2*} respectively as n goes to infinity.*

Assumption 3.2.4. *when $n \to +\infty$, the distribution of $\sqrt{n}(\hat{\theta} - \theta_*)$ with respect to the maximum likelihood estimator $\hat{\theta}$ converges in law to the normal distribution with mean* *vector* 0 *and the variance covariance matrix* $H(\theta_*)^{-1}J(\theta_*)H(\theta_*)^{-1}$, *i.e. as* $n \to +\infty$, *the following holds:*

$$\sqrt{n}(\hat{\theta} - \theta_*) \to N(0, H(\theta_*)^{-1}J(\theta_*)H(\theta_*)^{-1}).$$

Lemma 3.2.1. *Under Assumption* 3.2.1 - 3.2.3*, we have*

$$\begin{aligned} \varphi(f, g) &= w_1\varphi_X + w_2\varphi_Y \\ &= w_1\left\{E_X[\ell(\theta_{1*}; X)] + \frac{1}{2}\text{tr}[J(\theta_{1*})H^{-1}(\theta_{1*})]\right\} + \\ &\quad w_2\left\{E_Y[\ell(\theta_{2*}; Y)] + \frac{1}{2}\text{tr}[J(\theta_{2*})H^{-1}(\theta_{2*})]\right\} + o(1) \end{aligned}$$

with $J(\theta_{1*}) = \text{var}\left[\frac{\partial\ell(\theta_{1*};X)}{\partial\theta_{1*}}\right]$, $J(\theta_{2*}) = \text{var}\left[\frac{\partial\ell(\theta_{2*};Y)}{\partial\theta_{2*}}\right]$, $H(\theta_{1*}) = E\left[\frac{\partial\ell^2(\theta_{1*};X)}{\partial\theta_{1*}\partial\theta'_{1*}}\right]$, *and* $H(\theta_{2*}) = E\left[\frac{\partial\ell^2(\theta_{2*};Y)}{\partial\theta_{2*}\partial\theta'_{2*}}\right]$.

When $J(\theta_*) = -H(\theta_*)$, Lemma 3.2.1 is reduced to

$$\varphi(f, g) = w_1\left\{E_X\ell(\theta_{1*}; X) - \frac{1}{2}p_X\right\} + w_2\left\{E_Y\ell(\theta_{2*}; Y) - \frac{1}{2}p_Y\right\} + o(1).$$

Let $\zeta_X = E_X[\ell(\hat{\theta}(X); X)]$, $\zeta_Y = E_Y[\ell(\hat{\theta}(Y); Y)]$ and $\zeta = w_1\zeta_X + w_2\zeta_Y$. The ζ is weighted linear combination of expectations of log likelihoods at MLE's.

Lemma 3.2.2. *Under Assumptions* 3.2.1 - 3.2.4*, we have that*

$$\begin{aligned} \zeta(f, g) &= w_1\left\{E_X[\ell(\theta_{1*}; X)] - \frac{1}{2}\text{tr}[J(\theta_{1*})H^{-1}(\theta_{1*})]\right\} \\ &\quad + w_2\left\{E_Y[\ell(\theta_{2*}; Y)] - \frac{1}{2}\text{tr}[J(\theta_{2*})H^{-1}(\theta_{2*})]\right\} + o(1). \end{aligned}$$

When $J(\theta_*) = -H(\theta_*)$, Lemma 3.2.2 is reduced to

$$\begin{aligned} \zeta &= w_1\zeta_X + w_2\zeta_Y \\ &= w_1\left\{E_X[\ell(\theta_{1*};X)] + \frac{1}{2}p_X\right\} + w_2\left\{E_Y[\ell(\theta_{2*};Y)] + \frac{1}{2}p_Y\right\} + o(1). \end{aligned}$$

The proof of the above Lemmas take similar approach as Varin and Vidoni (2005) and is shown in Appendix A.

From the Lemmas we can immediately see that $\ell(\hat{\theta}_1, \hat{\theta}_2; X, Y)$ is biased and that, under the standard regularity conditions, the following defined information criterion is a first-order unbiased estimators for $\varphi(g, f)$, and selects the model that maximizes the information criterion

$$w_1\left[\ell(\hat{\theta}_1; X) + \mathrm{tr}(\hat{J}(X)\hat{H}(X)^{-1})\right] + w_2\left[\ell(\hat{\theta}_2; Y) + \mathrm{tr}(\hat{J}(Y)\hat{H}(Y)^{-1})\right],$$

where $\hat{J}(X)$, $\hat{H}(X)$, $\hat{J}(Y)$, and $\hat{H}(Y)$ are consistent, first-order unbiased estimators for $J_X(\theta_{1*})$, $H_X(\theta_{1*})$, $J_Y(\theta_{2*})$ and $H_Y(\theta_{2*})$, respectively. It is equivalent to minimize

$$w_1 AIC_X + w_2 AIC_Y.$$

In general, we are able to write our criterion for the case with Q data sets as the follows

$$\sum_{q=1}^{Q} w_q AIC_q, \tag{3.1}$$

where $q = 1, \cdots, Q$ indicates the number of data sets.

3.2.2 Weighted integrative AICs

In practice, data may come from variety sources which have different sizes, formats and qualities. The variability of AIC comes from these differences. One may wish to take into account these differences. Therefore, we assign different weights to different data sets. Our objective is to define the weights to minimize the variance of the weighted integrative AICs. The proof of the following Lemma takes similar approach as Fraser (1976) and is shown in Appendix A. Consider the case of two data sets with two independent test statistic t_1 and t_2.

Lemma 3.2.3. *Given two test statistics t_1 and t_2, a weighted sum of the test statistics is*

$$wt_1 + (1 - w)t_2,$$

where the sum of weights is 1. The weight to minimize the variance is

$$w = \frac{\mathrm{var}(t_2)}{\mathrm{var}(t_1) + \mathrm{var}(t_2)}. \tag{3.2}$$

In more general cases, if there exists Q data sets, the q^{th} weight w_q is the variance of the q^{th} test statistic proportion to the total variance of the test statistics from the all Q data sets. In other words, we can write equation (3.2) as

$$w_q = \frac{\mathrm{var}(t_q)}{\sum_{q=1}^{Q} \mathrm{var}(t_q)}, \quad q = 1, \cdots, Q.$$

Consequently, our information criterion in equation (3.1) can be rewritten as minimizing the following criterion

$$\sum_{q=1}^{Q} w_q AIC_q, \tag{3.3}$$

where $w_q = \frac{\text{var}(AIC_q)}{\sum_{q=1}^{Q} \text{var}(AIC_q)}$. Since AIC is based on models, first we need to find an approximated true model under $w \equiv 1$. Based on the approximated true model, we estimate the variance of AIC and further choose our weights. In practice, in order to obtain the weights, we need to estimate the variance of AICs by applying bootstrap. Let AIC_1 be obtained from data X and AIC_{1B} be obtained from bootstrap. First, we resample the data set with replacement by N times. N should be large enough. Second, compute $AIC_{1B} = (AIC_1^{[1]}, AIC_1^{[2]}, \cdots, AIC_1^{[N]})$ by using the same computing formula as the one used for AIC_1, i.e. $AIC_1^{[k]}, k = 1, 2, \cdots, N$, are bootstrap copies of AIC_1. Third, we are able to calculate the variance of AIC_1 according to AIC_{1B}. We repeat the same procedures for Q data sets in order to obtain var(AIC) for Q data sets.

3.3 Simulation Data Results

In this section, we perform simulation experiments to compare our proposed weighted integrative AIC method with individual AICs and equal weights integrated AIC method. The result shows our method has the best performance in terms of false detected numbers (FD) and false negative numbers (FN). We simulate two different scenarios. In the first scenario, we simulate true variables with large coefficients and all rest coefficients are set to be zeros. In the second scenario, the true model contains several covariates with very small coefficients. In this way, we can find our whether our method can correctly detect minor effected covariates.

Consider two data sets share the same regression model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$$

In the first scenario, the sample sizes for two data sets are 100 and 200. The number of variables p is set 80 for both data sets. The true model contains first 10 independent variables in the model. We generate X from normal distribution $N(3, 10)$ for the first data set and $N(5, 6)$ for the second data set. For the first data set, there are 10 coefficient βs generated from uniform $U(-2, 5)$, and the rest of βs are set to be 0. For the second data set, 10 βs are generated from $U(-2, 1)$, and the rest of βs are set to be 0. The error ε is i.i.d from $N(0, 2)$ for the first data set and $N(0, 1)$ for the second data set. The simulation results are shown in Table 3.1. In the second scenario, the sample size for two data sets are 100 and 200. The number of variables p is set 80 for both data sets. The true model contains first 50 independent variables in the model. We generate X from normal distribution $N(3, 5)$ for the first data set and $N(-5, 6)$ for the second data set. For 1^{st} data set, there are 47 βs generated from $U(10, 30)$, 3 βs from $U(0.05, 0.5)$, and the rest of βs are set to be zero. For 2^{nd} data set, there are 47 βs generated from uniform $U(5, 25)$, 3 βs from $U(0.02, 0.3)$, and the rest of βs are set to be zero. The error ε is i.i.d from $N(0, 1)$ for the first data set and $N(0, 2)$ for the second data set. The simulation results are shown in Table 3.2.

The first step in our simulation experiment is applying least absolute shrinkage and selection operator (LASSO) regression method to selected sub-sets. The LASSO proposed by Tibshirani (1996) is a promising variable selection technique, which is a penalized least squares method, imposing a constraint on the L1 norm of the regression coefficients. We used the statistical R package ”lars” to obtain regression coefficients $\hat{\beta}_{lasso}$, $\hat{\beta}_{lasso} = \arg\min_{\beta} \left| Y - \sum_{j=1}^{p} X_j \beta_j \right|^2 = \lambda \sum_{j=1}^{p} |\beta_j|$, where λ takes non-negative values. After LASSO selected subsets, we calculate AICs and apply our proposed weighted integrative AICs to select the best model among all competing models. The prediction model is fitted by multiple linear regression model. We compare our proposed weighted integrative AICs method with individual AIC method and equal weights AIC method, respectively. Let AIC_1 and AIC_2 denote AIC values from data set 1 and data set 2. We are minimizing the following information criteria, respectively, i.e. individual AIC, equal weights AIC and weighted integrative AIC:

$$
\begin{aligned}
AIC_1 &= -2\ell(\hat{\beta}_1; X_1) + p_1, \\
AIC_2 &= -2\ell(\hat{\beta}_2; X_2) + p_2, \\
AIC_{eq} &= -2\ell(\hat{\beta}_1; X_1) + p_1 + -2\ell(\hat{\beta}_2; X_2) + p_2, \\
AIC_w &= wAIC_1 + (1 - w)AIC_2,
\end{aligned}
$$

where p is the number of parameters in the subset models, β is the unpenalized regression coefficient based on the subset models and $w = \frac{var(AIC_1)}{var(AIC_1+AIC_2)}$.

In the weighted integrative AICs method, the variance of AIC is obtained by bootstrap. We resample 100 times for two data sets respectively with replacement. For each resampled data, we include LASSO selected variables and fit model by multiple linear regression. Then, we calculate AIC_1 and AIC_2 according to the above formulas. Accordingly, we can further obtain the variance of AIC_1 and AIC_2.

Table 3.1 and Table 3.2 show our simulation results based on 100 simulations. In the tables, we compare three methods, individual AIC (AIC_1, AIC_2), equal weights integrative AICs (AIC_{eq}) and weighted integrative AICs (AIC_w). We report, on average, how many number of variables are false detected(FD) and false negative (FN) selected by the selected best model among the competing models according to three methods, respectively. The corresponding standard deviation values are reported in the table as well. The better the method performs, the smaller values should be shown in the corresponding method. We can see that among these three methods, in the first scenario, our proposed method has the lowest mean of FD with small variance. In the second scenario, for FD, our method falsely detects the number of true variables is 18 less than individual AIC_1, 2 less than individual AIC_2 and 14 less than AIC_{eq} on average. For FN, our method has only 0.11 higher FN detected true variable numbers than AIC_1 and 0.1 higher than AIC_{eq}. Comparing to the

FD, the FN is much smaller. Hence, overall, our method out-performs the individual methods and equal weights method.

Table 3.1: Simulation Results: Mean and standard deviation of FD and FN for 1^{st} scenario.

	AIC_1	AIC_2	AIC_{eq}	AIC_w
FD Mean	54.57	15.45	19.07	6.21
FD std	27.31	5.55	21.38	6.90
FN Mean	0	0	0	0
FN std	0	0	0	0

Table 3.2: Simulation Results: Mean and standard deviation of FD and FN for 2^{nd} scenario.

	AIC_1	AIC_2	AIC_{eq}	AIC_w
FD Mean	26.43	10.60	21.77	8.15
FD std	6.03	4.28	7.14	6.23
FN Mean	0.01	0.23	0.02	0.12
FN std	0.10	0.49	0.14	0.38

3.4 Conclusion

When a true statistical model can not be specified, we propose to choose an approximate model in order to conduct statistical inference. Akaike's information criterion (AIC) is one of the commonly used model evaluation criteria based on one single data set. When information come from multiple different data sets, we propose a weighted integrative AICs method for a model evaluation criterion. Our proposed method combines AICs across multiple data sets with different weights that minimize the variance of the integrative AICs. In the simulation studies, the proposed method provides better performance than individual method or equal weights method in terms of false detected and false negative selected of true variable numbers. The disadvantage of the method is that weights are not easy to compute. The possible application could be that the data sets contain same observations and measurements, but measurements are measured at different time points. One may wish to find a common predict model across multiple data sets.

4 Model Comparison Test

4.1 Introduction

Model selection is an important topic in statistical inference. When there exist a class of competing models, we are interested in choosing the best model by a suitable model evaluation criterion. Many methods are developed in statistical literature, such as Mallows Cp, stepwise, backward and forward selection procedures, Akaike information criterion (AIC), Bayesian information criterion (BIC), cross-validation, and so on. Model comparison is usually performed by comparing some information criteria like AIC or BIC. AIC and BIC compare a collection of models. But, neither AIC nor BIC gives p-value or reflects the sampling variance. Hypotheses test is able to take into account sampling variance and report p-values when two models are compared.

For fixed alternative hypothesis under which the distance between two models is independent of sample size, Linhart (1988) proposed a test of whether AICs differ significantly associated with two candidate models for non-nested model. Contrary to fixed alternatives, the local alternatives means under which the distance between the alternative and the null can decrease when sample size goes large. For example, as sample size getting larger, the collection of predictors (which is not the true model) can predict the response better. Shimodaira (1998) proposed a modification of Linhart's test statistic by adding a second order term for local alternatives and developed corresponding asymptotic theory. Both Linhart test statistic and Shirmodaria's modified test statistic are based on full likelihood function. However, the full likelihood function can be difficult to specify in high dimension. Therefore, composite likelihood is useful in these situations.

We extend Linhart's and Shimodaira's test statistic by using composite likelihood function for correlated data sets. In our proposed method, we aim to improve the accuracy for estimating the variance of difference of two AICs and perform model comparison test. Indeed, the second order term in our proposed test offers improvement in both variance estimation and test error probability especially for small samples. In our simulation, we compare our proposed methods with Linhart's method and Shimodaira's method. In variance estimation of the difference of two AICs, second order method has better variance estimation than first order method by taking bootstrapped variance as threshold. In assessing error probability for model comparison test, our method has lower error probability in fixed and local alternatives.

In the next section, we review Linhart's test statistic, Shimodaira's modification of

Linhart's test statistic and composite likelihood function. In Section 4.3, we illustrate our proposed methods in details. Our simulation experiments settings and results are shown in Section 4.4. Appendix B lists terminologies and notations for this chapter. Proofs for fix alternative scenario are shown in Appendix C. Lemmas and proofs for local alternative scenario are shown in Appendix D.

4.2 Literature Reviews

4.2.1 Linhart's Test Statistic and Modification of Linhart's Test Statistic

Linhart (1988) considered a test of whether two AICs differ significantly. The test statistic is a standardized difference of AIC between the two models. It asymptotically converges to a standard normal distribution $N(0, 1)$, as the sample size n goes to infinity under the null hypothesis that the two expected discrepancies are equal. In Linhart's test, f is the true distribution function and g is an approximating distribution function. The test statistic is based on K-L divergence defined in Chapter 3. Under certain regularity conditions and misspecification for both models, Linhart's hypothesis about the expected discrepancies for model 1 and model 2 is stated as the following:

$$E_X\left[E_{X'}(\log g^{(1)}(X'|\hat{\theta}^{(1)}(X)))\right] \leq E_X\left[E_{X'}(\log g^{(2)}(X'|\hat{\theta}^{(2)}(X)))\right],$$

where $g^{(1)}(X'|\hat{\theta}^{(1)}(X))$ and $g^{(2)}(X'|\hat{\theta}^{(2)}(X))$ are two competing models. This is equivalent to say that model 1 fits better than model 2. Linhart proposes the test statistic

$$Z = \frac{\sqrt{n}(AIC^{(1)} - AIC^{(2)})}{\sqrt{\hat{\lambda}^{(1,1)} + \hat{\lambda}^{(2,2)} - 2\hat{\lambda}^{(1,2)}}},$$

where the elements $\lambda^{(i,j)}$ and $\hat{\lambda}^{(i,j)}, i, j = 1, 2$ are defind, respectively, as the following:

$$\lambda^{(i,j)} = E_X\left[\log g^{(i)}(X;\hat{\theta}^{(i)})\log g^{(j)}(X;\hat{\theta}^{(j)})\right] - E_X\left[\log g^{(i)}(X;\hat{\theta}^{(i)})\right]E_X\left[\log g^{(j)}(X;\hat{\theta}^{(j)})\right],$$

$$\hat{\lambda}^{(i,j)} = n^{-1}\sum_{t=1}^{n}\log g^{(i)}(X_t;\hat{\theta}^{(i)})\log g^{(j)}(X_t;\hat{\theta}^{(j)}) - n^{-2}\sum_{t=1}^{n}\log g^{(i)}(X_t;\hat{\theta}^{(1)})\sum_{t=1}^{n}\log g^{(j)}(X_t;\hat{\theta}^{(2)})$$

The test statistic converges to $N(0, 1)$, when sample size n goes to infinity.

Shimodaira (1997) considered a sequence of densities converging to $O(1/\sqrt{n})$ so that the test statistic will be bounded in probability even if n goes to infinity. In Shimodaira's test statistic, the second order term is added to the variance estimator of the difference between the two AIC's. The proposed estimator of the var$(AIC^{(1)} - AIC^{(2)})$ takes form as $(V^{(1,2)}/n + \nu^{(1,2)}/n^2)$, and $V^{(1,2)}$ and $\nu^{(1,2)}$ are described below

$$\begin{aligned} V^{(1,2)} &= n^{-1}\sum_{t=1}^{n}\left(\log g^{(1)}(X_t;\hat{\theta}^{(1)}) - \log g^{(2)}(X_t;\hat{\theta}^{(2)})\right)^2 \\ &- \left(n^{-1}\sum_{t=1}^{n}\log g^{(1)}(X_t;\hat{\theta}^{(1)}) - n^{-1}\sum_{t=1}^{n}\log g^{(2)}(X_t;\hat{\theta}^{(2)})\right)^2, \end{aligned}$$

and

$$\nu^{(1,2)} = (p^{(1)} + p^{(2)})/2 - tr(G^{(1,2)}G^{(2,2)^{-1}}G^{(2,1)}G^{(1,1)^{-1}}),$$

where $p^{(1)}$ and $p^{(2)}$ are the numbers of parameters in model 1 and model 2, and

$$G^{(i,j)} \;=\; n^{-1}\sum_{t=1}^{n}\left\{\frac{\partial \log g^{(i)}(x_t;\hat{\theta}^{(i)})}{\partial\theta^{(i)}}\cdot\frac{\partial \log g^{(j)}(x_t;\hat{\theta}^{(j)})}{\partial\theta^{(j)}}\right\}, i, j = 1, 2.$$

The modification of Linhart test statistic by Shimodaria is defined as

$$T = \frac{AIC^{(1)} - AIC^{(2)}}{\sqrt{V^{(1,2)}/n + v^{(1,2)}/n^2}}.$$

Shimodaira's test statistic improves the variance estimation and model comparison test, especially, for small sample size and local alternative situations.

4.2.2 Composite Likelihood

Composite likelihood methods are extensions of the Fisherian likelihood theory, one of the most influential approaches in statistics. Such extensions are generally motivated by the issue of computational feasibility arising in the application of the likelihood method in high-dimensional data analysis. It is methodologically appealing in projecting high-dimensional complicated likelihood functions to low-dimensional computationally feasible likelihood objects. Composite likelihood inherits many of the good properties of inference based on the full likelihood function, but is more easily implemented with high-dimensional data sets.

In general formulation of composite likelihood, we can group composite likelihoods into two main groups. The first includes subsetting method, which is pseudo-likelihood constructed from lower dimensional densities. For example, the pairwise likelihood (Cox and Reid, 2004), which is based on marginal events related to pairs of observations. Similarly, we may define the tripletwise likelihood and so on. The other class is based on omission method, which the composite likelihoods are obtained by omitting some components in the full likelihood to simplify the evaluation. Examples include m^{th}-order likelihood for stationary processes (Azzalini, 1983), partial likelihood (Cox, 1975), pseudo likelihood (Besag, 1974), and so on.

In our research, we focus on the subsetting method. Let Y be a p-dimensional random vector with probability density function $f(y;\theta)$, where $\theta \in \Theta$ is a d-dimension parameter vector of interest. Suppose $\{A_1, \cdots, A_K\}$ is a set of events with associated likelihood function $L_k(\theta; y) \propto f(y \in A_k; \theta),\ k = 1, 2, \cdots, K.$ Following Lindsay (1988), the composite likelihood function is defined as

$$\mathrm{CL}(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k},$$

where $\{w_k\}$ is a set of positive weights assigned to each component in order to improve estimation efficiency.

There are two general types of composite likelihood: conditional and marginal composite likelihood. The conditional type of composite likelihood method was first proposed by Besag (1974). The idea is to specify the joint probability distribution by conditional probability functions,

$$CCL(\theta; y) = \prod_{i=1}^{p} f(y_i | y_{-i}; \theta)^{w_i},$$

where y_{-i} denotes the random vector y_i deleted. In the type of marginal composite likelihood, the simplest composite likelihood is the one constructed under the independence assumption:

$$L_{ind}(\theta; Y) = \prod_{i=1}^{p} f(y_i; \theta)^{w_i}.$$

The most popular form in the current literature is pairwise composite likelihood. It contains the minimal modeling blocks of marginal and dependence parameters, essential for correlated data analysis (Cox and Reid, 2004; Varin, 2008),

$$L_{pair}(\theta; y) = \prod_{i=1}^{p-1} \prod_{j=r+1}^{p} f(y_i, y_j; \theta)^{w_{ij}}.$$

There are many other designed composite likelihoods such as tripletwise likelihood, blockwise likelihood, pairwise differences likelihood and so on. One may also combine composite conditional likelihoods and composite marginal likelihoods (Cox and Reid, 2004).

With a sample of independent observations $\mathbf{y} = (y^{(1)}, \cdots, y^{(n)})$, the overall composite log likelihood function is

$$c\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} c\ell(\theta; y^{(i)}) = \sum_{i=1}^{n} \log CL(\theta; y^{(i)}).$$

The maximum composite likelihood estimator (MCLE) is defined by

$$\hat{\theta}_{CL} = \arg\max_{\theta} c\ell(\theta; \mathbf{y}).$$

We suppose the random vector Y has distribution function $F(y)$; the marginal distribution function for a sub-vector $Y_k \subset Y$ is $F_k(y_k)$ and the corresponding density function is $f_k(y_k), k = 1, \cdots, K$. Now consider the family of modelled distributions for Y_k, with common support and family of density functions $\{g_k(y_k;\theta); \theta \in \Omega\}$. We are interesting in the weighted composite marginal likelihood and its corresponding log liklihood function:

$$CL(\theta; y) = \prod_{k=1}^{K} g_k(y_k;\theta)^{w_k},$$

and

$$\begin{aligned} c\ell(\theta; y) &= \sum_{k=1}^{K} c\ell_k(\theta; y_k) \\ &= \sum_{k=1}^{K} w_k \log(g_k(y_k;\theta)). \end{aligned} \tag{4.1}$$

For misspecified composite likelihood, θ^* is a parameter point which minimizes the composite K-L distance (Varin and Vidoni, 2005):

$$\theta^* = \arg\min_\theta \mathrm{E}_Y \left\{ \log \frac{\prod_{k=1}^{K} f_k(Y_k)}{CL(\theta; Y)} \right\} = \arg\min_\theta \sum_{k=1}^{K} \mathrm{E}_Y \left\{ \log \frac{f_k(Y_k)}{g_k(Y_k;\theta)} \right\}.$$

The maximum composite likelihood estimator (MCLE) $\hat{\theta}$ solves the composite likelihood score function $u(\theta; Y_k)$, which is defined as

$$u(\theta; Y_k) = \sum_{k=1}^{K} \frac{\partial c\ell_k(\theta; Y_k)}{\partial\theta} = \sum_{k=1}^{K} w_k \frac{\partial \log g_k(Y_k;\theta))}{\partial\theta}.$$

We solve it at

$$\sum_{k=1}^{K} w_k \frac{\partial \log g_k(Y_k;\theta)}{\partial\theta} = 0.$$

Although the composite likelihood is not a real likelihood, the maximum composite likelihood estimate is still consistent for θ^*. This is because the composite score function is a linear combination of several valid likelihood score functions. Under the usual regularity conditions, it is still unbiased (Gao and Song, 2011). The asymptotic covariance matrix of maximum composite likelihood estimator takes the form of the inverse of the Godambe information (Godambe, 1960):

$$G(\theta) = H(\theta)J^{-1}(\theta)H(\theta), \tag{4.2}$$

where $H(\theta) = E\left[-\frac{\sum_{k=1}^{K} \partial^2 c\ell_k(\theta;Y_k)}{\partial\theta\partial\theta'}\right]$ is the sensitivity matrix, and $J(\theta) = \text{var}\left[\sum_{k=1}^{K} \frac{\partial c\ell_k(\theta;Y_k)}{\partial\theta}\right]$ is the variability matrix. In the full likelihood, the Godambe information becomes Fisher information since $H(\theta) = J^{-1}(\theta)$. However, when using composite likelihood methods, we have to consider likelihood theory under misspecification even if the true model for the data is taken into account. As the result, the identity of $H(\theta)$ and $J(\theta)$ doesn't hold, i.e. $H(\theta) \neq J^{-1}(\theta)$, leading to the loss of efficiency compared to the maximum likelihood estimation (Song, 2007, Chapter3).

4.3 Method

This section illustrates our approaches to develop variance estimators of the difference between AICs and model comparison test statistics under full or composite likelihood with local alternative or fixed alternative setting for correlated data sets. Our theoretical proofs take similar approaches as Linhart's (1988) and Shimodaira's (1997). Section 4.3.1 extends Linhart method to composite likelihood with fixed alternative setting. The corresponding proofs are shown in Appendix C. Section 4.3.2 extends Shimodaira's method to composite likelihood with local alternative setting. The corresponding proofs are shown in Appendix D. The related notations and terminologies are listed in Appendix B.

4.3.1 Composite Likelihood with Fixed Alternative Setting

Consider a parametric family of densities of random variable Y, $f(\cdot) = f(Y;\phi)$ where $\phi \in \Phi \subset \mathcal{R}^d$ is the parameter value. Let $Y = (Y^{(1)}, \cdots, Y^{(n)})$ be independently and identically distributed with unknown true distribution function $f(Y;\phi)$. Let the approximated density function for Y under model α be $g(Y;\theta_\alpha)$ and under model β be $g(Y;\theta_\beta)$, respectively. As defined in Section 4.2.2, the overall composite log likelihood function is

$$\begin{aligned} c\ell^{(n)}(\phi;Y) &= \sum_{i=1}^{n} c\ell(\phi;Y^{(i)}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_k \log g_k(Y_k^{(i)};\phi). \end{aligned}$$

The sensitivity matrix H and variability matrix J for each data point are written as $H_{ij}(\phi) = E\left[-\frac{\sum_{k=1}^{K} \partial^2 c\ell_k(\phi;Y_k)}{\partial\phi^i \partial\phi^j}\right]$ and $J_{ij}(\phi) = \left[\sum_{k=1}^{K}\left(\frac{\partial c\ell_k(\phi;Y_k)}{\partial\phi^i}\right)\left(\frac{\partial c\ell_k(\phi;Y_k)}{\partial\phi^j}\right)'\right]$, respectively. Assume $\hat{\phi}$ is defined as a solution to the composite likelihood equation, i.e. $\hat{\phi} = \arg\sup_\phi c\ell^{(n)}(\phi;Y)$. Let $AIC_{c\ell} = -2c\ell^{(n)}(\hat{\phi};Y) + 2\mathrm{tr}(J^* H^{*-1})$ denotes composite likelihood information criterion as

Varin(2005). Under model α and β and dividing $AIC_{c\ell}$ by $2n$, we have

$$C^{(n)}_\alpha = -c\ell^{(n)}_\alpha(\hat{\theta}_\alpha; Y)/n + \mathrm{tr}(J^*_\alpha H^{-1*}_\alpha)/n$$

and

$$C^{(n)}_\beta = -c\ell^{(n)}_\beta(\hat{\theta}_\beta; Y)/n + \mathrm{tr}(J^*_\beta H^{*-1}_\beta)/n.$$

For analytical proofs, there are several regularity assumptions need to be introduced. First, we borrow the assumptions 1-8 from Xu and Reid(2011). We also assume that, for every fixed $y \in Y$, $c\ell^{(n)}(\phi; Y)$ is twice differentiable with continuity with respect to ϕ. Let $p$lim denotes the convergence in probability. Under model α and β, our null hypothesis is

$$H_0 : E_Y\left[\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y)\right] \leq E_Y\left[\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y)\right].$$

Theorem 4.3.1. *For model $\alpha, \beta \in \mathcal{M}$, the estimation of the variance and test statistic are as the following:*

$$\underset{n\to\infty}{p\mathrm{lim}}\, nV_{\alpha\beta} = \underset{n\to\infty}{p\mathrm{lim}}\, \mathrm{var}\left(C^{(n)}_\alpha - C^{(n)}_\beta\right), \tag{4.3a}$$

$$T_{\alpha\beta} = \frac{C^{(n)}_\alpha - C^{(n)}_\beta}{\sqrt{V^{(n)}_{\alpha\beta}/n}}, \tag{4.3b}$$

where

$$V^{(n)}_{\alpha\beta} = n^{-1}\sum_{i=1}^{n}\left[c\ell(\hat{\theta}_\alpha; Y^{(i)}) - c\ell(\hat{\theta}_\beta; Y^{(i)})\right]^2 - \left[c\ell^{(n)}_\alpha(\hat{\theta}_\alpha; Y^{(i)})/n - c\ell^{(n)}_\beta(\hat{\theta}_\beta; Y^{(i)})/n\right]^2.$$

4.3.2 Composite Likelihood with Local Alternative Setting

Let $Y = (Y^{(1)}, \cdots, Y^{(n)})$ be independently and identically distributed with unknown true distribution function $f(Y; \phi^{(n)})$, where $\phi^{(n)} \in \Phi$ is true parameter value. We consider $\phi^{(n)}$ depends on the number of observations n, and it converges to ϕ^*, an interior of Φ. The rate of the convergence is of order $O(1/\sqrt{n})$, that is $\lim_{n\to\infty} \sqrt{n}(\phi^{(n)} - \phi^*) = \phi^\diamond \in \mathcal{R}^d$. Let Φ^* denote a generic neighborhood of $\phi^* \in \Phi$, whose scale is magnified by $\sqrt{n}$ times. Later, we will see the space of distribution in Φ^*, whose scale is magnified by $\sqrt{n}$ times, reduces asymptotically to a linear space as $n \to \infty$. The composite likelihood function, H and J matrices are defined as in previous section. Except assumptions stated in Section 4.3.1, there are two additional assumptions as follows:

Assumption 4.3.1. *Assume $\sqrt{n}(\hat{\phi}^{(n)} - \phi^{(n)}) \overset{d}{\sim} N(0, G^{-1})$, where G is Godambe information matrix defined in equation* (4.2).

Assumption 4.3.2. *Assume $\sqrt{n}(\phi^{(n)} - \phi^*) = \phi^\diamond$, and assume CMLE is asymptotically bounded in probability. That is, $p\lim_{n\to\infty} \sqrt{n}(\hat{\phi}^{(n)} - \phi^*) = \hat{\phi}^\diamond = O_p(1)$.*

Let $\mathcal{M}$ be the set of α's for the candidate models, where α indexes models. Under model α, consider a parametric family of density functions $f_\alpha(Y; \theta_\alpha)$. We assume $f_\alpha(\cdot)$ is a subset of $f(\cdot)$ and $f(Y; \phi^*)$ is interior to $f_\alpha(\cdot)$. For each $\alpha \in \mathcal{M}$, we consider a composite log likelihood function $c\ell_\alpha(\theta_\alpha; Y), \theta_\alpha \in \Theta_\alpha$. Using a function $\phi_\alpha : \Theta_\alpha \to \Phi$, for notation convenience, we write $c\ell_\alpha(\theta_\alpha; Y) = c\ell(\phi_\alpha(\theta_\alpha); Y)$, $\hat{\phi}_\alpha^{(n)} = \phi_\alpha(\hat{\theta}_\alpha^{(n)})$ and under the local alternative setting $\phi^* = \phi_\alpha(\theta_\alpha^*)$ for some $\theta_\alpha^* \in \Theta_\alpha$, where θ_α^* is interior to Θ_α.

To estimate $\mathrm{var}\left(C_\alpha^{(n)} - C_\beta^{(n)}\right)$, we are going to investigate an estimate which takes form

$$V_{\alpha\beta}^{(n)}/n + \nu_{\alpha\beta}^{(n)}/n^2,$$

where two terms $V_{\alpha\beta}^{(n)}$ and $\nu_{\alpha\beta}^{(n)}$ are:

$$V_{\alpha\beta}^{(n)} = n^{-1} \sum_{i=1}^{n} \left[c\ell(\hat{\phi}_\alpha^{(n)}; Y) - c\ell(\hat{\phi}_\beta^{(n)}; Y^{(i)})\right]^2 - \left[c\ell_\alpha^{(n)}(\hat{\phi}_\alpha^{(n)}; Y)/n - c\ell_\beta^{(n)}(\hat{\phi}_\beta^{(n)}; Y)/n\right]^2, \quad (4.4)$$

and

$$\begin{aligned} \nu_{\alpha\beta}^{(n)} &= \mathrm{tr}\left(H_{\alpha\alpha}^{(n)-1} J_{\alpha\alpha}^{(n)} H_{\alpha\alpha}^{(n)-1} J_{\alpha\alpha}^{(n)} + H_{\beta\beta}^{(n)-1} J_{\beta\beta}^{(n)} H_{\beta\beta}^{(n)-1} J_{\beta\beta}^{(n)}\right)/2 \qquad (4.5) \\ &\quad -\mathrm{tr}(H_{\alpha\alpha}^{(n)-1} J_{\alpha\beta}^{(n)} H_{\beta\beta}^{(n)-1} J_{\beta\alpha}^{(n)}). \end{aligned}$$

Each element in $H^{(n)}_{\alpha\alpha}$, $H^{(n)}_{\beta\beta}$, $J^{(n)}_{\alpha\alpha}$, $J^{(n)}_{\beta\beta}$ and $J^{(n)}_{\alpha\beta}$ is as the following

$$\begin{aligned}
(H^{(n)}_{\alpha\alpha})_{ij} &= \frac{\partial^2 c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)}{\partial\theta^i_\alpha \partial\theta^{j\prime}_\alpha},\\
(H^{(n)}_{\beta\beta})_{ij} &= \frac{\partial^2 c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)}{\partial\theta^i_\beta \partial\theta^{j\prime}_\beta},\\
(J^{(n)}_{\alpha\alpha})_{ij} &= \left(\frac{\partial c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)}{\partial\theta^i_\alpha}\right)\left(\frac{\partial c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)}{\partial\theta^j_\alpha}\right)',\\
(J^{(n)}_{\beta\beta})_{ij} &= \left(\frac{\partial c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)}{\partial\theta^i_\beta}\right)\left(\frac{\partial c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)}{\partial\theta^j_\beta}\right)',\\
(J^{(n)}_{\alpha\beta})_{ij} &= \left(\frac{\partial c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)}{\partial\theta^i_\alpha}\right)\left(\frac{\partial c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)}{\partial\theta^j_\beta}\right)',\\
(J^{(n)}_{\beta\alpha})_{ij} &= \left(\frac{\partial c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)}{\partial\theta^i_\beta}\right)\left(\frac{\partial c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)}{\partial\theta^j_\alpha}\right)'.
\end{aligned}$$

Under model α and β, our null hypothesis is

$$E_Y\left[\frac{1}{n}c\ell^{(n)}(\phi_\alpha(\hat\theta^{(n)}_\alpha); Y)\right] \le E_Y\left[\frac{1}{n}c\ell^{(n)}(\phi_\beta(\hat\theta^{(n)}_\beta); Y)\right].$$

Theorem 4.3.2. *For model $\alpha, \beta \in \mathcal{M}$, the estimation of the variance and test statistic are as the following:*

$$\underset{n\to\infty}{p\mathrm{lim}}(nV_{\alpha\beta} + \nu_{\alpha\beta}) = \underset{n\to\infty}{p\mathrm{lim}}\,\mathrm{var}\left(C^{(n)}_\alpha - C^{(n)}_\beta\right), \tag{4.7a}$$

$$T_{\alpha\beta} = \frac{C^{(n)}_\alpha - C^{(n)}_\beta}{\sqrt{V^{(n)}_{\alpha\beta}/n + \nu^{(n)}_{\alpha\beta}/n^2}}. \tag{4.7b}$$

Note, in the full likelihood function, the J matrix is equal to the H matrix which is known as Fisher's expected information matrix. We denote it as $\mathcal{I}$. Therefore, our second term can be simplified as

$$\nu_{\alpha\beta} = \mathrm{tr}(m_\alpha + m_\beta)/2 + \mathrm{tr}(\mathcal{I}_{\alpha\alpha}^{(n)-1}\mathcal{I}_{\alpha\beta}^{(n)}\mathcal{I}_{\beta\beta}^{(n)-1}\mathcal{I}_{\beta\alpha}^{(n)}), \tag{4.8}$$

where m_α and m_β are the number of parameters in the model α and β respectively. The model comparison test is derived under assumptions that (4.7b) is normally distributed with unit variance.

4.4 Simulation Results

In this section we present our main results from simulation studies. Firstly, we evaluate the accuracy of estimation of the variance of the difference of two AICs, and then we assess the error probability for model comparison test. We compare our second order method with first order method under various simulation settings, such as, independent and correlate case under fix or local alternatives.

4.4.1 Data Generation

(i). **Data Generation with Independent Case**

Consider regression model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where ϵ is *i.i.d.* from $N(0, \sigma^2)$. We set $\sigma^2 = 1$. The number of covariates p is set to be 30. The true model contains 15 covariates. All covariates are generated from normal distribution with mean 0 and variance 1. The sample sizes n = 40, 50, 60, 70, 80, 100, 300 and 500 are considered.

(ii). **Data Generation with Multivariate Correlated Case**

Denote the numbers of families by n and members in each family by s. The response vector of measurements for the i^{th} family is denoted by $Y_i = (y_{i1}, \cdots, y_{im})'$. Associated is a set of covariates at individual level, $X_i = (x_{i1}, \cdots, x_{ik})'$, with $x_{ij} = (x_{ij1}, ..., x_{ijp})'$, representing the p covariates observed for the j^{th} individual in the i^{th} family. The response vector for i^{th} family, Y_i, follows a multivariate normal distribution $MVN_s(\mu_i, \Sigma)$, where the mean vector is governed by a linear model, $\mu_i = X_i\beta$, with $\beta = (\beta_1, \cdots, \beta_p)'$. The covariance matrix Σ is specified according to an exchangeable dependence structure, $\sigma_{j,j'} = \rho$.

We set $p = 30$ and $s = 8$. The within-family correlation $\rho = 0.8$. The covariates are generated from $N(0, 1)$. The 15 regression coefficients of the true marginal model are set $\beta_{true} \sim N(2, 5)$, with the other 15 coefficients set to zero. The number of families are set to be $n = 10, 15, 25, 30, 50, 100, 300, 500$, respectively.

4.4.2 Variance Estimation of the Difference of AICs.

For estimating the variance of the difference of two AICs, we compare Linhart estimation, Shimodaria estimation and our extended methods to bootstrap values. The comparison are under independent case and correlated case with fixed or local alternatives.

(i). **Fixed Alternatives:**

Consider model a and model b denoted as $\mathcal{M}_a$ and $\mathcal{M}_b$. Model $\mathcal{M}_a$ is an overfitted model that contains 15 true covariates and 8 wrong covariates. Model $\mathcal{M}_b$ is a competing model that contains partial true covariates and partial wrong covarites. Two models are as follows

$$\mathcal{M}_a : E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{23} x_{23},$$

$$\mathcal{M}_b : E(Y) = \beta_0 + \beta_1 x_3 + \beta_2 x_4 + \cdots + \beta_{21} x_{23}.$$

In bootstrap method, we generate error term from normal distribution with mean 0 and variance 1 for each simulation. We replicate error by generating random error term 1000 times to obtain variance of the difference for two AICs.

The comparing methods are as follows:

- In the independent case, the comparing methods are

  (a) Bootstrap method,

(b) Linhart method,

(c) extended Shimodaria method.

- In the correlated case, the comparing methods are

  (a) Bootstrap method,

  (b) extended Linhart method,

  (c) extended Shimodaria method,

Figure 4.1 shows the variance estimation results with fixed alternatives. In the left panel of the figure, we empirically verify that in the independent case our extended Shimodaria method has better estimation for the variance of the difference two of AICs than Linhart method because its curve is closer to the bootstrap curve than Linhart curve. The right panel of the figure shows that in the correlated case our Shimodaira method also has better estimation comparing to extended Linhart method.

(ii). **Local Alternatives:**

Consider local alternatives setting for $\mathcal{M}_a$ and $\mathcal{M}_b$, where $\mathcal{M}_a$ is true model and $\mathcal{M}_b$ is competing model,

$$\mathcal{M}_a : E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{15} x_{15},$$

$$\mathcal{M}_b : E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{15} x_{15}^{\triangle}.$$

We set $x_{15}^{\triangle} = x_{15} + \frac{c}{\sqrt{n}} * z$, where n takes value as number of clusters, c is a constant set as 2 and z is from uniform distribution $z \sim U(0.5, 1)$.

The comparing methods are as follows:

- In the independent case, the comparing methods are
  (a) Bootstrap method,
  (b) extended Linhart method,
  (c) Shimordaira method.
- In the correlated case, the comparing methods are
  (a) Boostrap method,
  (b) extended Linhart method,
  (c) extended Shimodaira method.

Left panel of Figure 4.2 shows the results in the independent case where the bootsrap method, extended Linhart method and extended Shimodaira method are compared. Akin to fix alternative, extended Shimodaira method method has better estimation because its estimation curve is the closest one to the bootstrap curve. In addition, when the sample size is small, the extended Shimodaira method improves the variance estimation obviously since its curve is much closer to the bootstrap curve than the extend Linhart curve. This is because the second order term is added to the extended Shimordaria method but not in extend Linhart method.

4.4.3 Error probability assessment

In this section, we assess the error probability for model comparison test. Our null hypothesis is H_0: $\mathcal{M}_a$ is better than $\mathcal{M}_b$. We calculate the rate for rejection H_0 over 1000 simulations. The smaller reject rate means the less error and higher power in model comparison test. Same as variance estimate, we evaluate our test statistics under full likelihood function, composite likelihood function, fixed alternatives or local alternatives. The test statistic is proposed as equation (4.7b),

$$T = \frac{AIC_a/2n - AIC_b/2n}{\sqrt{V_{ab}/n + v_{ab}/n^2}}.$$

(i). **Fixed Alternatives:**

Consider two models under fixed alternative setting. Model $\mathcal{M}_a$ is true model containing true 15 covariates. For competing model $\mathcal{M}_b$, we randomly drop one true covariate, which means $\mathcal{M}_b$ containing 14 true covariates.

The comparing methods are as follows:

- In the independent case, the comparing methods are

  (a) Linhart test,

  (b) extended Linhart test.

- In the correlated case, the comparing methods are

  (a) extended Linhart test,

  (b) extended Shimodaira test,

Figure 4.3 shows the simulation results for assessing the error of the test. The left panel of the figure shows the results under independent case. We empirically verify that extended Shimodaira test statistic's rejection rate doesn't go over 0.05 and it has lower rejection rate than Linhart has. Hence, extended Shimodaira test has higher testing power comparing to Linhart method. The right panel presents the results under correlated scenario. The extended Shimodaira method has lower error probability and higher testing power comparing to extended Linhart test. Both error rates from these two tests go down to zero when sample size getting larger as expected.

(ii). **Local alternatives:**

We are using same local alternatives setting as we did in variance estimation. The simulation results of the following comparison methods are obtained:

- In the independent case, the comparing methods are

  (a) extend Linhart test,

  (b) Shimodaira test.

- In the correlated case, the comparing methods are

(a) extended Linhart test,

(b) extend Shimodaira test,

Figure 4.4 shows the error probability under local alternatives has lower error rate and high test power. Under independent case, extended Shimodaira method has lower error probably and reaches to reject rate 0.05 faster than extended Linhart method as shown in the left panel of the figure. The right panel shows the results under correlated case, again, extended Shimodaira method has higher testing power comparing to extended Linhart method.

4.5 Conclusion

We extend Linhart's and Shimodaira's test statistics by using composite likelihood function for correlate data. The asymptotic variance estimation and error probability of Linhart's and Shimodaira's model selection test are evaluated. We examine two cases, one is where the expected discrepancies of the candidate models from the true model are fixed when sample size goes large. The other case is where the expected discrepancies of the candidate models from the true model remains unchange when sample size goes large. In the first case the fixed alternatives method is applied in the limiting operation of the asymptotic evaluation. In the second case, the local alternatives method is employed in

Figure 4.1: Plots for variance estimation under fixed alternatives.

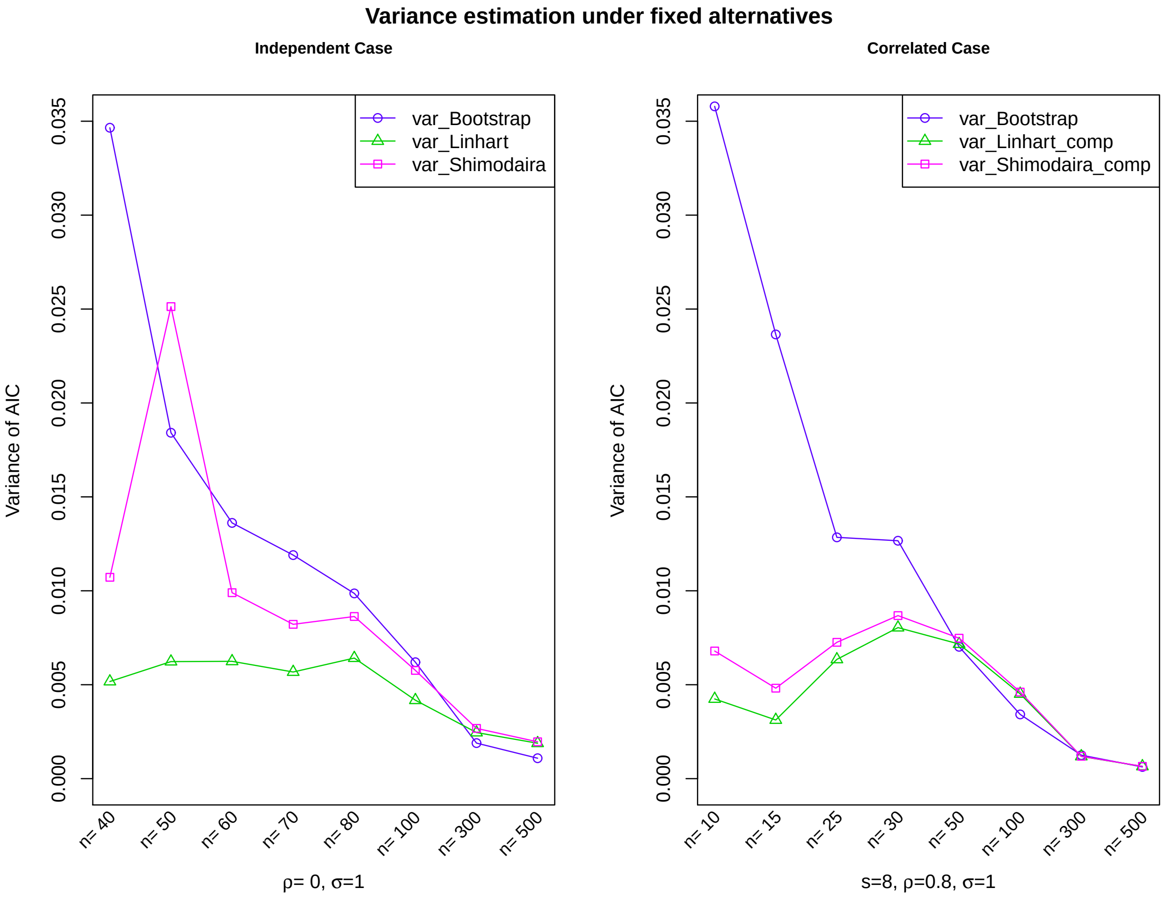


Figure 4.2: Plots for variance estimation under local alternatives.

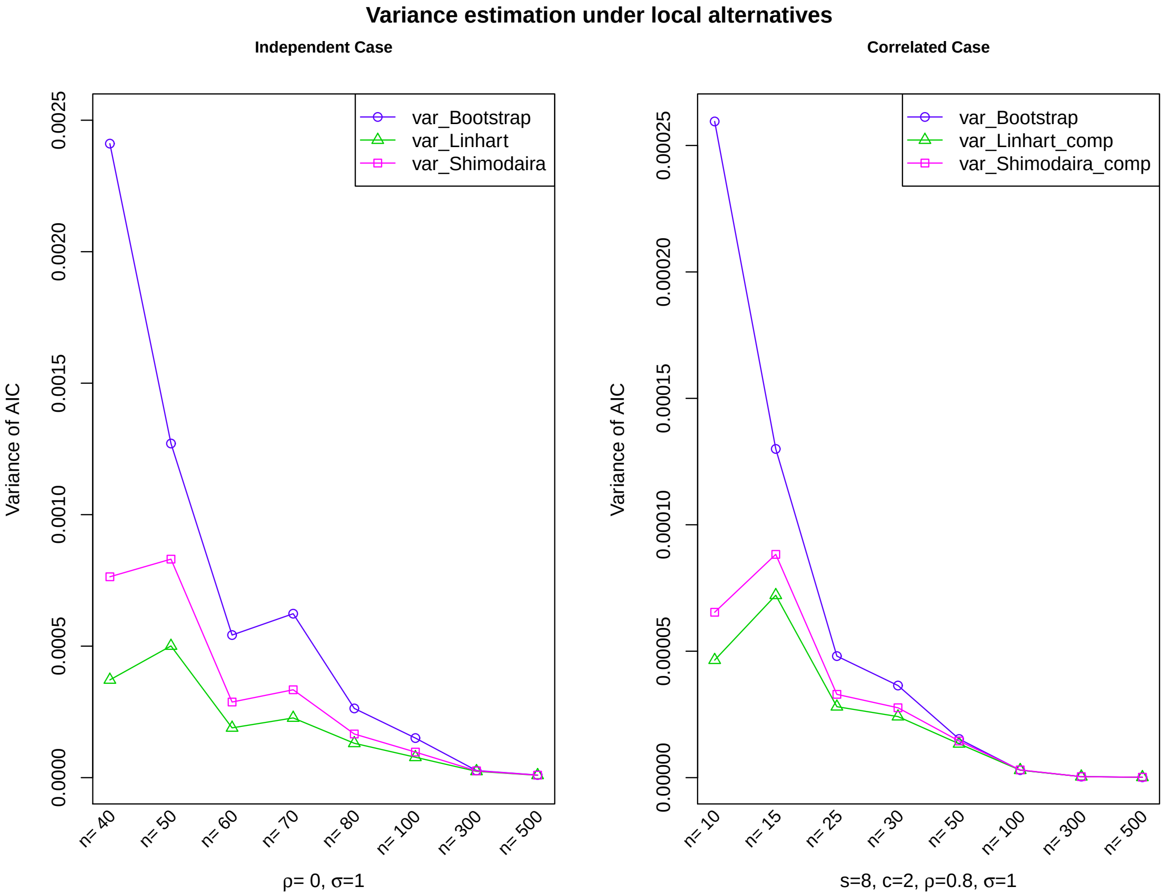


Figure 4.3: Plots for for assessing error probability under fixed alternatives.

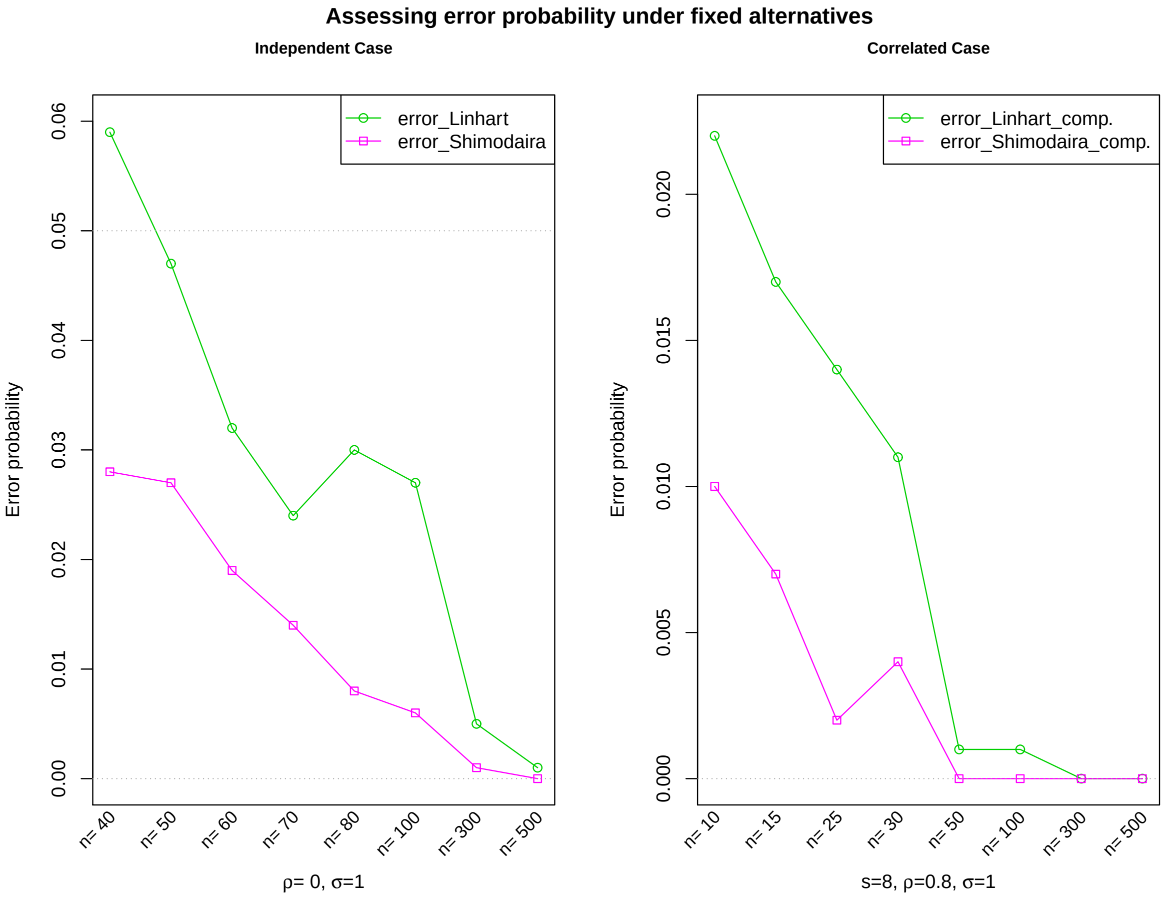
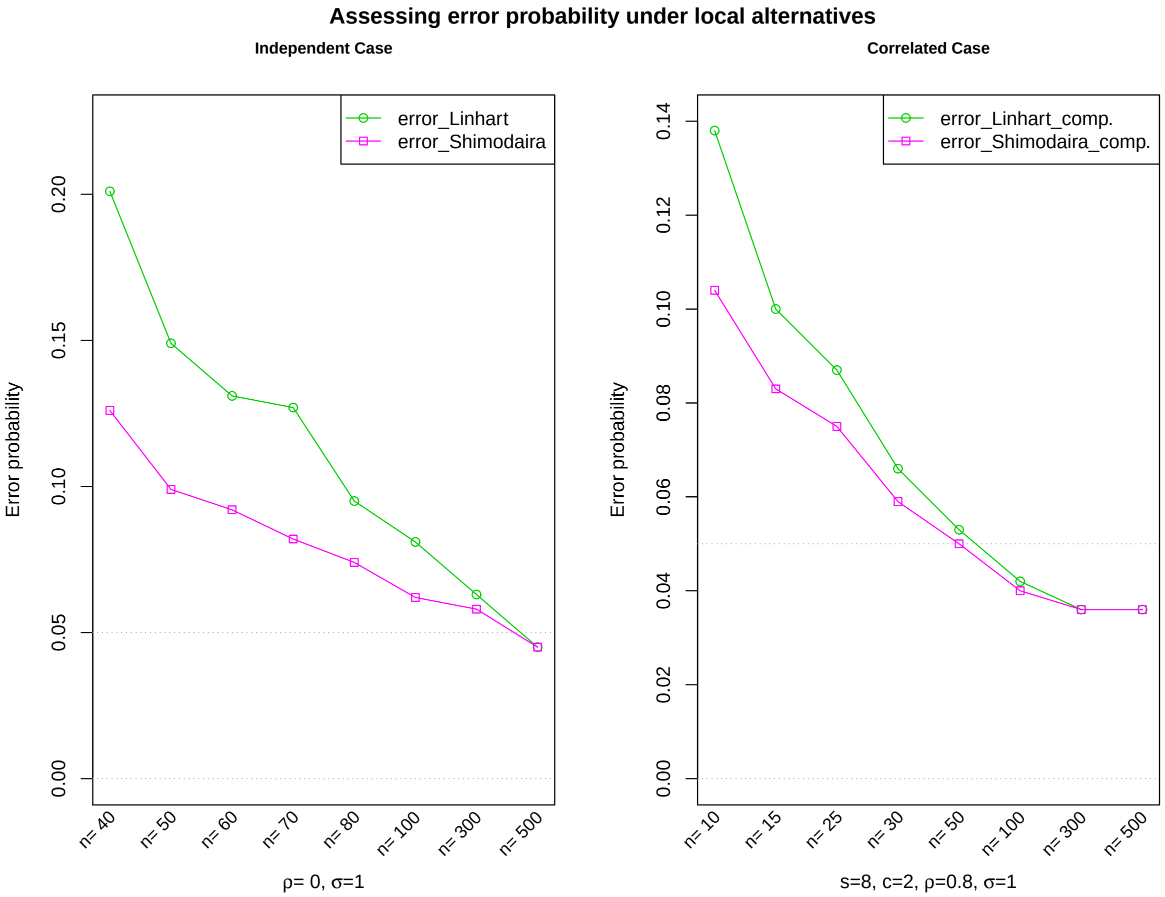


Figure 4.4: Plots for assessing error probability under local alternatives.



the limiting operation of the asymptotic evaluation. The second order term is added to the variance estimator of the difference between the two AICs, whereas only the first order term is used in Linhart's test statistic. This modification improves the variance estimation and model selection test to a considerable extent, especially for small sample size data set. The effectiveness of proposed variance estimation and model selection test are confirmed through analysis and numerical simulation.

5 Discussion and Future work

In the Multiple-platform data integration project, we proposed integration method involving aggregated test statistics which is weighted sum of the test statistics across different platforms with the weights being the inverse of the empirical standard deviation of each statistic. However, this linear combination of test statistic can be sensitive to the direction of data sets, but it cannot capture the geometric shape. Our interest is to obtain optimized aggregated test. With exploration the methods of combine test statistics, we raise an interesting topic which is, in the case of two independent models for investigating the same population characteristics, we are going to combine log likelihood which is locally defined canonical parameter (an ingredient for 3rd order asymptotic) when the model is a one parameter model with no nuisance parameter. Furthermore, we will generalize it to nuisance parameter case.

In the clustering mixed data project, we have proposed a nonparametric clustering method for finding group in mixed data. Numerical results show that the proposed method outperforms the AutoClass algorithm based on examinations of classification rate and entropy measure. In the future work, in order to increase computation efficiency, instead of using weighted local Chi-squared test to approximate the distribution of the weighted Chi-squares, the saddle point method will be applied. We will also extend the proposed method to cluster spatial and temporal data.

Regarding to model selection criteria, there are extensive model selection literatures, but many of them focus on the analysis of univariate data set. Relatively limited work has been done for multiple data sets. AIC is one of the widely used promising model selection criteria. It is based on the likelihood and asymptotic properties of the maximum likelihood estimator. The AIC method can only be applied when a full likelihood function is available. However, if the full likelihood cannot be defined for the data set such as multiple data sets. Our proposed weighted integrative AICs criterion can perform model selection across multiple data sets. Simulation studies show us the proposed method has good performance in terms of lower false negative numbers numbers and false detected numbers by comparing to individual test and equal weights combining test. In the future work, we may extend our method to highly correlated longitudinal data or clustered data. Furthermore, we will extend the proposed method to data sets with large number of independent variables but small number of observations.

In model comparison project, we have extended Linhart's and Shirmodaria's test statistics by using composite likelihood function with local and fixed alternatives for correlated data sets to perform model comparison test. In the future work, we will perform multiple comparison to model selection. Rather than choosing a single model, we consider a confidence set of models meaning constructed a set of good models.

Bibliography

[1] Adourian A, Jennings E, Balasubramanian R, Hines W, Damian D, Plasterer T, Clish C, Stroobant P, McBurney R, Verheij E, Bobeldijk I, van der Greef J, Lindberg J, Kenne K, Andersson U, Hellmold H, Nilsson K, Salter H, Schuppe-Koistinen I (2003) **Correlation network analysis for data integration and biomarker selection.** *The Royal Society of Chemistry*, **4**:249–259.

[2] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006), **Gene prioritization through genomic data fusion.** *Nature Biotechnology*, **24**:537–544.

[3] Ahmad A and Dey L (2007), **A k-mean clustering algorithm for mixed numeric and categorical data.** *Data & Knowledge Engineering*, **63**: 503527.

[4] Akaike H (1973), **Information theory as an extension of the maximum likelihood principle**. *Second International Symposium on Information Theory* (B. N. Petrov, and F. Csaki, Eds.) Akademiai Kiado, Budapest.

[5] Azzalini A (1983), **Maximum likelihood of order m for stationary stochastic processes.** *Biometrika*, **70**, 381-397

[6] Banfield J, and Raftery A (1993), **Model-Based Gaussian and Non-Gaussian Clustering**. *Biometrics*, **49**, 803-821

[7] Bentley S, Chater K, Cerdeno-Tarraga A, Challis G, Thomson N, James K, Harris D, Quail M, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen C, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream M, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell B, Parkhill J, Hopwood D (2002), **Complete genome sequence of the model actionomycete Streptomyces coelicolor A3(2).** *Nature*, **417**:141–147.

[8] Besag J (1974), **Spatial interaction and the statistical analysis of lattice systems.** *Journal of the Royal Statistical Society*, Series B, **36**, 192-236.

[9] Bradley P, Fayyad U, Reina C (1998), **Scaling Clustering Algorithms to Large Databases.** *in Proceedings of the Fourth International conference on Knowledge Discovery and Data Mining, New York, August 1998, CA: AAAI Press*, pp.9-15.

[10] Buness A, Ruschhaupt M, Kuner R, Tresch A (2009) **Classification across gene expression microarrray studies**. *Bioinformatics*, **10**:453.

[11] Bussey K, Chin K, Lababidi S, Reimers M, Reinhold W, Kuo W, Gwadry F, Ajay, Kouros-Mehr H, Fridlyand J, Jain A, Collins C, Nishizuka S, Tonon G, Roschke A, Gehlhaus K, Kirsch I, Scudiero D, Gray J, Weinstein J (2006) **Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel.** *Molecular Cancer Therapeutics*, **5**:853–867.

[12] Cheeseman P, Stutz J (1995), **Bayesian classification (AUTOCLASS): Theory and Results.** *in Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, Menlo Park, CA: AAAI Press*, pp 153-180.

[13] Cox D R (1975), **Partial likelihood.** *Biometrika* **62**, 269-276.

[14] Cox D R and Reid N (2004), **A note on pseudolikelihood constructed from marginal densities.** *Biometrika* **91**, 729-737.

[15] Daemen A, Gevaert O, De Bie T, Debucquoy A, Machiels J, De Moor B, Haustermans K (2008), **Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer.** *Pacific Symposium on Biocomputing*, **13**:166–177.

[16] Fraley C, and Raftery A (1998), **How Many Clusters? Which Clustering Method? Answer via Model-Based Cluster Analysis.** *The Computer Journal*, **41**, 578-587.

[17] Fraser D A S (1976), **Probability and Statistics: Theory and Applications** *DAI, University of Toronto Bookstore* pp.382

[18] Gao X (2006), **Construction of null statistics in permutation based multiple testing for multi-factorial microarray experiments.** *Bioinformatics*, **22**:1486–1494.

[19] Gao X and Song P (2011), **Composite Likelihood EM Algorithm with Applications to Multivariate Hidden Markov Model.** *Statistica Sinica*, **21**, 165-186.

[20] Gersho, A and Gray, R M (1992), **Vector Quantization and Signal Compression.** *Kluwer Academic Publishers, Boston*.

[21] Godambe V P (1960), **An optimum property of regular maximum likelihood equation.** *Annals of Mathematical Statistics*, **31**, 1208-1211.

[22] Graf, S and Luschgy, H (2000), **Foundations of quantization for probability distributions. Lecture Notes in Mathematics**, No. 1730, p230, *Springer*.

[23] Hamid J, Hu P, Roslin M, Ling V, Greenwood C, Beyene J (2009), **Data integration in genetics and genomics: Methods and challenges**. *Human Genomics and Proteomics*, **9**:869093.

[24] Hochberg Y, Tamhane A (1987), *Multiple comparison procedures*. New Jersey: Wiley.

[25] Hu P, Greenwood C, Beyene J (2006), **Statistical methods for meta-analysis of microarray data: A comparative study.** *Information Systems Frontiers*, **8**:9–20.

[26] Huang Z X (1997), **Extensions to the K-Means Algorithm for Clustering large Data Sets With Categorical Values**. *Data Mining and Knowledge discovery*, **2**, 283-304.

[27] Hurvich C M and Tsai C L (1989), **Regression and Time Series Model Selection in Small Samples.** *Biometrika*, **76**, 297-307.

[28] Jayapal K, Philp R, Kok Y, Yap M, Sherman D, Griffin T, Hu W (2008), **Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor.** *PLoS One*, **3**:e2097.

[29] Jayapal K, Sui S, Philp R, Kok Y, Yap M, Griffin T, Hu W (2010), **Multitagging proteomic strategy to estimate protein turnover rates in dynamic systems.** *Journal of Proteome Research*, **9**:2087–2097.

[30] Jennrich R I (1969), **Asymptotic properties of non-linear least squares estimators.** *The Annals of Mathematical Statistics*, **40**, 633-643.

[31] Kaufman L, and Rousseeuw P J (2005), **Finding Groups in Data: An Introduction to Cluster Analysis** *New York: Wiley.*

[32] Kolde R, Laur S, Adler P, Vilo J (2012), **Robust rank aggregation for gene list integration and meta-analysis.** *Bioinformatics*, **4**:573–580.

[33] Kullback S, and Leibler R (1951), **On information and sufficiency.** *Annals of Mathematical Statistics*, **22**, 79-86.

[34] Laboulais C, Ouali M, Le Bret M, and Gabarro-Arpa J (2002), **Hamming Distance Geometry of a Protein Conformational Space. Application to the Clustering of Molecular Dynamics Trajectory of the HIV-1 Intergrase Catalytic Core.** *Proteins: Structure, Function and Genetics*, **47**, 169-179

[35] Lanckriet G, Bie T, Cristianini N, Jordan M, Noble S (2004), **A statistical framework for genomic data fusion.** *Bioinformatics*, **20**:2626–2635.

[36] Lebreton J-D, Burnham K P, Clobert J, Anderson D R (1992), **Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies.** *Ecological Monograph*, **62** 67-118

[37] Linhart H.(1988) **A test whether two AIC's differ significantly.** *South African Statist. J.* **22**,153-161.

[38] Lindsay B (1988), **Composite likelihood methods.** *Statistical inference from stochastic process*, Ed. N. U. Prabhu, pp.221-39. Providence, RI: American Mathematical Society.

[39] Ma Y, Ding Z, Qian Y, Wan Y, Tosun K, Shi X, Castranova V, Harner E, Guo N (2009), **An integrative genomic and proteomic approach to chemosensitivity prediction.** *International Journal of Oncology*, **34**:107–115.

[40] Manteca A, Sanchez J, Jung H, Schwamle V, Jensen O (2010), **Quantitative proteomics analysis of Streptomyces coelicolor development demonstrates that onset of secondary metabolism coincides with hypha differentiation.** *Molecular Cellular Proteomics.*, **9(7)**:1423–36.

[41] Mehra S, Lian W, Jayapal K, Charaniya S, Sherman D, Hu W (2006), **A framework to analyze multiple time series data: A case study with Streptomyces coelicolor.** *Journal of Industrial Microbiology Biotechnology*, **33(2)**:159–72.

[42] Nieselt K, Battke F, Herbig A, Bruheim P, Wentzel A, Jakobsen O, Sletta H, Alam M, Merlo M, Moore J, Omara W, Morrissey E, Juarez-Hermosillo M, Rodriguez-Garcia A, Nentwich M, Thomas L, Iqbal M, Legaie R, Gaze G WH andChallis, Jansen R, Dijkhuizen L, Rand D, Wild D, Bonin M, Reuther J, Wohlleben W, Smith M, Burroughs N, Martin J, Hodgson D, Takano E, Breitling R, Ellingsen T, Wellington

E (2010), **The dynamic architecture of the metabolic switch in Streptomyces coelicolor.** *BMC Genomics*, **11**:10.

[43] Reif D, White B, Moore J (2004), **Integrated analysis of genetic, genomic and proteomic data.** *Expert Review of Proteomics*, **1**:67–75.

[44] Rhodes D, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan A (2004), **Large-scale meta analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.**, **25**:9309–9314.

[45] Roman S (1992), **Coding and Information Theory**. *Springer-Verlag, New York.*

[46] Shibata R (1989), **Statistical Aspects of Model Selection** in *From Data to Model*, ed. by J. C. Willems, pp. 215240. Springer-Verlag New York.

[47] Shimodaira H (1997), **Assessing the error probability of the model selection test.** *Ann. Inst. Statist. Math.* **49**,395-410.

[48] Shimodaira H (1998), **An application of multiple comparison techniques to model selection.** *Ann. Inst. Statist. Math.* **50**,1-13.

[49] Song P X (2007), **Correlated Data Analysis: Modeling, Analytics, and Applications.** *New York: Springer*.

[50] Stone M (1977), **An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion.** *Journal of the Royal Statistical Society. Series B*, **39**, 44-47.

[51] Takeuchi K (1976), **Distribution of informational statistics and a criterion of model fitting.** *Suri-Kagaku Mathematical Sciences* (in Japanese) **153**: 12-18.

[52] Tian Q, Stepaniants S, Mao M, Weng L, Feetham M, Doyle M, Yi E, Dai H, Thorsson V, Eng J, Goodlett D, Berger J, Gunter B, Linseley P, Stoughton R, Aebersold R, Collins S, Hanlon W, Hood L (2004), **Integrated genomic and proteomic analyses of gene expression in mammalian cells.** *Molecular Cellular Proteomics*, **3**:960–969.

[53] Tibshirani R (1996), **Regression Skrinkage and Selection via the Lasso**. *JR Statist Soc B* ;**58**:267-88.

[54] Varin C and Vidoni P (2005), **A Note on Composite Likelihood Inference and Model Selection** *Biometrika*, **3**:519–528.

[55] Varin C (2008), **On composite marginal likelihoods.** *ASTA Advances in Statistical Analysis*, **92**, 1-28

[56] Varin C, Reid N, Firth D (2011), **An Overview of Composite Likelihood Methods** *Statistica Sinica*, Vol. 21, pp. 5-42

[57] Xu X, Reid N, (2011) **On the robustness of maximum composite likelihood estimate**. *J. Statist. Plann. Inference*, doi:10.1016/j.jspi.2011.03.026.

[58] Zhang P, Wang X and Song P X (2005), **Clustering Categorical Data Based on Distance Vectors.** *The Journal of the American Statistical Association*. **473** 355-367.

A Appendix: Proofs for Section 3.2 Weighted integrative AICs criterion

Proof. Lemma 3.2.1:

Let's start from φ_x. Consider the Taylor Expansion for $\ell(\hat{\theta}_1(X); X')$ with respect to $\hat{\theta}_1(X)$ around θ_{1*}, we have

$$\begin{aligned}
\ell(\hat{\theta}_1(X); X') \quad &= \quad \ell(\theta_{1*}; X') + (\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial \ell(\theta_{1*}; X')}{\partial \theta_{1*}} \right) + \\
&\qquad \frac{1}{2}(\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X')}{\partial \theta_{1*} \partial \theta'_{1*}} \right) (\hat{\theta}_1(X) - \theta_{1*}) + o_p(1),
\end{aligned}$$

Take expectation with respect to the true distribution of X'. Note X and X' are *i.i.d.* and Assumption 3.2.3 holds, we have

$$\begin{aligned}
E_{X'}[\ell(\hat{\theta}_1(X); X')] \quad &= \quad E_{X'}[\ell(\theta_{1*}; X)] + (\hat{\theta}_1(X) - \theta_{1*})^T E_{X'} \left[\left(\frac{\ell(\partial \theta_{1*}; X)}{\partial \theta_{1*}} \right) \right] \\
&\qquad + \frac{1}{2}(\hat{\theta}_1(X) - \theta_{1*})^T E_{X'} \left[\frac{\partial^2 \ell(\theta_{1*}; X')}{\partial \theta_{1*} \partial \theta'_{1*}} \right] (\hat{\theta}_1(X) - \theta_{1*}) + o_p(1) \\
&= \quad E_{X'}[\ell(\theta_{1*}; X)] + \frac{1}{2}(\hat{\theta}_1(X) - \theta_{1*})^T \\
&\qquad E_{X'} \left[\frac{\partial^2 \ell(\theta_{1*}; X')}{\partial \theta_{1*} \partial \theta'_{1*}} \right] (\hat{\theta}_1(X) - \theta_{1*}) + o_p(1).
\end{aligned}$$

Moreover, take expectation with respect to X. We have

$$\begin{aligned}
\varphi_X &= E_X\left[E_{X'}[\ell(\theta_{1*};X)]\right] + E_X\left\{\frac{1}{2}(\hat{\theta}_1(X)-\theta_{1*})^T E_{X'}\left[\frac{\partial^2\ell(\theta_{1*};X')}{\partial\theta_{1*}\partial\theta'_{1*}}\right](\hat{\theta}_1(X)-\theta_{1*})\right\} \\
&= E_X[\ell(\theta_{1*};X) + \frac{1}{2}\mathrm{tr}\left\{E_X\left[\frac{\partial^2\ell(\theta_{1*};X)}{\partial\theta_{1*}\partial\theta'_{1*}}\right]E_X\left[(\hat{\theta}_1(X)-\theta_{1*})(\hat{\theta}_1(X)-\theta_{1*})^T\right]\right\}.
\end{aligned}$$

In the above equation, we know $E_X\left[\frac{\partial^2\ell(\theta_{1*};X)}{\partial\theta_{1*}\partial\theta'_{1*}}\right]$ is actually H matrix. Now let's investigate the term $E_X\left[(\hat{\theta}_1(X)-\theta_{1*})(\hat{\theta}_1(X)-\theta_{1*})^T\right]$. We denote it as $V(\theta_{1*})$. By the assumption 3.2.4, we know asymptotically $\sqrt{n}(\hat{\theta}_1-\theta_{1*})$ follows normal distribution with mean vector 0 and the variance covariance matrix $H(\theta_{1*})^{-1}J(\theta_{1*})H(\theta_{1*})^{-1}$. Therefore, we are able to obtain

$$V(\theta_{1*}) = H(\theta_{1*})^{-1}J(\theta_{1*})H(\theta_{1*})^{-1} + o(n). \tag{A.1}$$

By all above, we are able to obtain

$$\varphi_X = E_X[\ell(\theta_{1*};X)] + \frac{1}{2}\mathrm{tr}[J(\theta_{1*})H(\theta_{1*})^{-1}] + o(1).$$

By applying the same approaches and arguments, we can derive

$$\varphi_Y = E_Y[\ell(\theta_{2*};Y)] + \frac{1}{2}\mathrm{tr}[J(\theta_{2*})H(\theta_{2*})^{-1}] + o(1).$$

Hence,

$$\begin{aligned}
\varphi(f,g) &= w_1\varphi_X + w_2\varphi_Y \\
&= w_1\left\{E_X[\ell(\theta_{1*};X)] + \frac{1}{2}\mathrm{tr}[J(\theta_{1*})H(\theta_{2*})^{-1}]\right\} \\
&\quad +w_2\left\{E_Y[\ell(\theta_{2*};Y)] + \frac{1}{2}\mathrm{tr}[J(\theta_{2*})H(\theta_{2*})^{-1}]\right\} + o(1).
\end{aligned}$$

We complete the proof for Lemma 3.2.1. □

Proof. Lemma 3.2.2:

We start with ζ_X. Take Taylor expansion to $\ell(\hat{\theta}(X); X)$ with respect to $\hat{\theta}_1(X)$ around θ_{1*}, we have

$$\begin{aligned}\ell(\hat{\theta}_1(X); X) &= \ell(\theta_{1*}; X) + (\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial \ell(\theta_{1*}; X)}{\partial \theta_{1*}}\right) \\ &\quad + \frac{1}{2}(\hat{\theta}_1(x) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X)}{\partial \theta_{1*} \partial \theta_{1*}}\right)(\hat{\theta}_1(X) - \theta_{1*}) + o_p(1).\end{aligned}$$

Since asymptotically $\frac{\partial \ell(\theta_{1*};X)}{\partial \theta_{1*}} \approx -(\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial \ell^2(\theta_{1*};X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)$, therefore the above equation becomes

$$\begin{aligned}\ell(\hat{\theta}_1(X); X) &= \ell(\theta_{1*}; X) - (\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)(\hat{\theta}_1(X) - \theta_{1*}) \\ &\quad + \frac{1}{2}(\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)(\hat{\theta}_1(X) - \theta_{1*}) + o_p(1) \\ &= \ell(\theta_{1*}; X) - \frac{1}{2}(\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)(\hat{\theta}_1(X) - \theta_{1*}) + o_p(1).\end{aligned}$$

Take the expectation with respect to the true distribution of X:

$$\begin{aligned}E_X[\ell(\hat{\theta}_1(X); X] &= E_X[\ell(\theta_{1*}; X)] \\ &\quad - \frac{1}{2}E_X\left[(\hat{\theta}_1(X) - \theta_{1*})^T \left(\frac{\partial^2 \ell(\theta_{1*}; X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)(\hat{\theta}_1(X) - \theta_{1*})\right] + o(1).\end{aligned}$$

Note $\frac{\partial^2 \ell(\theta_{1*};X)}{\partial \theta_{1*} \partial \theta'_{1*}} = E_X\left[\left(\frac{\partial^2 \ell(\theta_{1*};X)}{\partial \theta_{1*} \partial \theta'_{1*}}\right)\right] + o(1)$ and apply equation (A.1), we derive

$$\zeta_X = E_X[\ell(\hat{\theta}_1(X); X] = E_X[\ell(\theta_{1*}; X)] - \frac{1}{2}\mathrm{tr}[J(\theta_{1*})H^{-1}(\theta_{1*})] + o(1).$$

By applying similarly arguments and approaches, we are able to derive

$$\zeta_Y = E_Y[\ell(\hat{\theta}_2(Y); Y] = E_Y[\ell(\theta_{2*}; Y)] - \frac{1}{2}\mathrm{tr}[J(\theta_{2*})H^{-1}(\theta_{2*})] + o(1).$$

Therefore, we have

$$\begin{aligned}\zeta(f,g) &= w_1\zeta_X + w_2\zeta_Y \\ &= w_1\left\{E_X[\ell(\theta_{1*};X)] - \frac{1}{2}\mathrm{tr}[J(\theta_{1*})H^{-1}(\theta_{1*})]\right\} \\ &\quad + w_2\left\{E_Y[\ell(\theta_{2*};Y)] - \frac{1}{2}\mathrm{tr}[J(\theta_{2*})H^{-1}(\theta_{2*})]\right\} + o(1).\end{aligned}$$

This is completed prove of Lemma 3.2.2. □

Proof. Lemma 3.2.3:

Take variance of $wt_1 + (1-w)t_2$. We have

$$\begin{aligned}V &= \mathrm{var}\left[wt_1 + (1-w)t_2\right] \\ &= w^2\mathrm{var}(t_1) + (1-w)^2\mathrm{var}(t_2).\end{aligned}$$

Taking the derivative with respect to w and setting equation equal to zero gives

$$\begin{aligned}\frac{\partial V}{\partial w} = 2w\mathrm{var}(t_1) &- 2(1-w)\mathrm{var}(t_2) = 0 \\ 2w\mathrm{var}(t_1) &= 2(1-w)\mathrm{var}(t_2) \\ \frac{w}{1-w} &= \frac{\mathrm{var}(t_2)}{\mathrm{var}(t_1)}\end{aligned}$$

$$w = \frac{\mathrm{var}(t_2)}{\mathrm{var}(t_1) + \mathrm{var}(t_2)}.$$

The equation (3.2) is proved. □

B Appendix: Terminologies and notations for Chapter 4

The general terminologies and notations are used throughout Chapter 4 and corresponding Theorem and Lemma proofs in Appendix C and D.

- $\phi \in \Phi \subset \mathcal{R}^d$ is parameter space;
- $Y = (Y^{(1)}, \cdots, Y^{(n)})$ are i.i.d with unknown true distribution function $f(Y;\phi)$, and the approximating density function is $g(Y;\phi)$;
- $\phi^{(n)} \in \Phi$ is true parameter value depending on sample size n and convergence to ϕ^*;
- $\lim_{n\to\infty} \sqrt{n}(\phi^{(n)} - \phi^*) = \phi^\diamond$;
- $c\ell(\phi; Y) = \sum_{k=1}^{K} w_k \log g_k(Y_k|\phi)$ is composite log likelihood function for a sub-vector;
- $c\ell^{(n)}(\phi; Y) = \sum_{i=1}^{n} c\ell_k(\phi; Y^{(i)})$ is over all composite log likelihood function;
- $\hat{\phi}^{(n)}$ is the MLE of the above composite likelihood function;
- $H_{ij}(\phi) = E\left[-\frac{\sum_{k=1}^{K} \partial^2 c\ell_k(\theta; Y_k)}{\partial\phi^i \partial\phi^j}\right]$ is the sensitivity matrix for each data point

- $J_{ij}(\phi) = \left[\sum_{k=1}^{K}\left(\frac{\partial c\ell_k(\phi;Y_k)}{\partial\phi^i}\right)\left(\frac{\partial c\ell_k(\phi;Y_k)}{\partial\phi^j}\right)'\right]$ is the variability matrix for each data;

- $\lim_{n\to\infty} H(\phi^{(n)}) = H^*$ and $\lim_{n\to\infty} J(\phi^{(n)}) = J^*$;

- $\sqrt{n}(\phi^{(n)} - \phi^*) = \phi^{\diamond}$.

The following notations are used under model α

- $\theta_\alpha : \Theta_\alpha \to \Xi$ is a function;

- $\theta_\alpha = \phi_\alpha(\theta_\alpha)$, $\theta_\alpha^* \in \Theta_\alpha$;

- $B^i_{\alpha j}(\theta_\alpha) = \partial\phi^i_\alpha/\partial\theta^j_\alpha$ is a Jacobian matrix;

- $\theta^{(n)}_\alpha = \arg\sup_{\theta_\alpha\in\Theta_\alpha} \mathrm{KL}(\phi^{(n)}, \phi_\alpha(\theta_\alpha))$; KL denotes K-L divergence;

- $\lim_{n\to+\infty} \sqrt{n}(\theta^{(n)}_\alpha - \theta^*_\alpha) = \theta^{\diamond}_\alpha$;

- $D_{KL}(\phi_1, \phi_2) = D(\phi_1, \phi_1) - D(\phi_1, \phi_2)$ is K-L distance;

- $D(\phi_1, \phi_2) = E_{\phi_1}\left[\sum_{k=1}^{K} w_k \log g(y_k|\phi_2)\right]$;

- $\hat{\theta}^{(n)}_\alpha$ is the MLE under model α

- $p\lim_{n\to\infty} \sqrt{n}(\hat{\theta}^{(n)}_\alpha - \theta^*_\alpha) = \hat{\theta}^{\diamond}_\alpha = O_p(1)$

- $\hat{\phi}^{(n)}_\alpha = \phi_\alpha(\hat{\theta}^{(n)}_\alpha)$

- $\hat{\phi}^{\diamond}_\alpha = p\lim_{n\to\infty} \sqrt{n}(\hat{\phi}^{(n)}_\alpha - \phi^*)$;

- $B^*_{\alpha} = B_{\alpha}(\theta^*_{\alpha})$;

- $\phi^{(n)}_{\alpha} = \phi_{\alpha}(\theta^{(n)}_{\alpha})$;

- $\phi^{\diamond}_{\alpha} = \lim_{n \to +\infty} \sqrt{n}(\phi^{(n)}_{\alpha} - \phi^*)$;

- $\phi^{\dagger}_{\alpha} = H^{*1/2}\phi^{\diamond}_{\alpha}$;

- $\hat{\phi}^{\dagger}_{\alpha} = H^{*1/2}\hat{\phi}^{\diamond}_{\alpha}$;

- $B^{\dagger}_{\alpha} = H^{*1/2}B^*_{\alpha}$;

- $P^{\dagger}_{\alpha} = B^{\dagger}_{\alpha}(B^{\dagger\prime}_{\alpha}B^{\dagger}_{\alpha})^{-1}B^{\dagger\prime}_{\alpha}$ is the projection operator of $I_m B^{\dagger}_{\alpha}$;

- $\hat{c\ell}^{(n)}_{\alpha} = c\ell^{(n)}(\hat{\phi}^{(n)}_{\alpha}; Y)$;

- $C^{(n)}_{\alpha} = \frac{AIC_{c\ell}}{2n}$;

- $C^{(n)}_{\alpha} = -\hat{c\ell}^{(n)}_{\alpha}/n + \mathrm{tr}(J^*_{\alpha}H^{*-1}_{\alpha})/n$.

C Appendix: Proofs for Section 4.3.1 composite likelihood with fixed alternative setting

Proof. For this we need at first the joint asymptotic distribution of

$$\sqrt{n}\left(\tfrac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y) - E_f[c\ell^{(n)}(\theta^*_\alpha; Y)]\right) \tag{C.1}$$

and

$$\sqrt{n}\left(\tfrac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y) - E_f[c\ell^{(n)}(\theta^*_\beta; Y)]\right).$$

Consider under model α the equation (C.1) can be write as

$$\sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y) - \frac{1}{n}c\ell^{(n)}(\theta^*_\alpha; Y)\right) + \sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\theta^*_\alpha; Y) - E_f[c\ell^{(n)}(\theta^*_\alpha; Y)]\right).$$

We want to show the first part of the above equation is negligible,i.e., $\sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y) - \frac{1}{n}c\ell^{(n)}(\theta^*_\alpha; Y)\right) = 0$. Consider

$$\sqrt{n}(\hat{\theta}_\alpha - \theta^*_\alpha)\left(\frac{1}{n}\frac{\partial c\ell^{(n)}(\bar{\theta}_\alpha; Y^{(i)})}{\partial\bar{\theta}_\alpha}\right),$$

where $\bar{\theta}_\alpha$ is neighborhood of θ^*_α and $\hat{\theta}_\alpha$. We already know that $\sqrt{n}(\hat{\theta}_\alpha - \theta^*_\alpha) = o_p(1)$ and $\frac{1}{n}\frac{\partial c\ell^{(n)}(\theta^*_\alpha;Y)}{\partial\theta^*_\alpha} = 0$. Note $\frac{1}{n}\frac{\partial c\ell^{(n)}(\theta_\alpha;Y)}{\partial\theta_\alpha} \xrightarrow{a.s} \frac{\partial E_f[c\ell^{(n)}(\theta_\alpha;Y)]}{\partial\theta_\alpha}$, uniformly in θ (Jennrich, 1969, Theorem

2), thus $\left(\frac{1}{n}\frac{\partial c\ell^{(n)}(\hat{\theta}_\alpha;Y)}{\partial\hat{\theta}_\alpha}\right) \xrightarrow{a.s} \frac{\partial E_f[c\ell^{(n)}(\theta^*_\alpha;Y)]}{\partial\theta^*_\alpha} = 0$. Therefore, it's sufficient to consider

$$\sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\theta^*_\alpha;Y) - E_f[c\ell^{(n)}(\theta^*_\alpha;Y)]\right)$$

By similar approach, under model β, it's sufficient to consider

$$\sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\theta^*_\beta;Y) - E_f[c\ell^{(n)}(\theta^*_\beta;Y)]\right).$$

By the central limit theorem, we are able to show

$$\begin{pmatrix} \sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\theta^*_\alpha;Y) - E_f[c\ell^{(n)}(\theta^*_\alpha;Y)]\right) \\ \sqrt{n}\left(\frac{1}{n}c\ell^{(n)}(\theta^*_\beta;Y) - E_f[c\ell^{(n)}(\theta^*_\beta;Y)]\right) \end{pmatrix} \overset{d}{\sim} N(0,\Lambda),$$

where the elements in Λ are

$$\begin{aligned}
\lambda_{\alpha\alpha} &= E_f[c\ell^{(n)}(\theta^*_\alpha;Y)^2] - (E[c\ell^{(n)}(\theta^*_\alpha;Y)])^2, \\
\lambda_{\alpha\beta} &= E_f[c\ell^{(n)}(\theta^*_\alpha;Y)c\ell^{(n)}(\theta^*_\beta;Y)] - E[c\ell^{(n)}(\theta^*_\alpha;Y)]E[c\ell^{(n)}(\theta^*_\beta;Y)], \\
\lambda_{\beta\alpha} &= E_f[c\ell^{(n)}(\theta^*_\beta;Y)c\ell^{(n)}(\theta^*_\alpha;Y)] - E[c\ell^{(n)}(\theta^*_\beta;Y)]E[c\ell^{(n)}(\theta^*_\alpha;Y)], \\
\lambda_{\beta\beta} &= E_f[c\ell^{(n)}(\theta^*_\beta;Y)^2] - (E[c\ell^{(n)}(\theta^*_\beta;Y)])^2.
\end{aligned}$$

Thus,

$$\frac{\frac{1}{(n)}c\ell^{n}(\hat{\theta}_\alpha;Y) - \frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta;Y) - E[c\ell^{(n)}(\theta^*_\alpha;Y)] + E[c\ell^{(n)}(\theta^*_\beta;Y)]}{\sqrt{(\lambda_{\alpha\alpha} + \lambda_{\beta\beta} - 2\lambda_{\alpha\beta})/n}} \overset{d}{\sim} N(0,1).$$

Note the expected K-L divergences for model α and model β are approximately equal to $E_f[c\ell^{(n)}(\theta^*_\alpha;Y)] + \frac{\mathrm{tr}J^*_\alpha H^{*-1}_\alpha}{2n}$ and $E_f[c\ell^{(n)}(\hat{\theta}^*_\beta;Y)] + \frac{\mathrm{tr}J^*_\beta H^{*-1}_\beta}{2n}$. Under the hypothesis that the two expected K-L information are equal

$$\frac{\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y) - \frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y) - \frac{\mathrm{tr} J^*_\alpha H^{*-1}_\alpha}{2n} + \frac{\mathrm{tr} J^*_\beta H^{*-1}_\beta}{2n}}{\sqrt{(\lambda_{\alpha\alpha} + \lambda_{\beta\beta} - 2\lambda_{\alpha\beta})/n}} \overset{d}{\sim} N(0, 1).$$

This is,

$$\frac{C^{(n)}_\alpha - C^{(n)}_\beta}{\sqrt{(\lambda_{\alpha\alpha} + \lambda_{\beta\beta} - 2\lambda_{\alpha\beta})/n}} \overset{d}{\sim} N(0, 1).$$

Consistent estimators $\hat{\Lambda}$ of Λ are obtained if in the expression for Λ the parameters θ^*_α, θ^*_β are replaced by $\hat{\theta}^*_\alpha$ and $\hat{\theta}^*_\beta$, respectively, and true density function $f(\cdot)$ by the empirical density function $g(\cdot)$. The asymptotic distribution remains unchanged if Λ is replaced by $\hat{\Lambda}$. Each element in $\hat{\Lambda}$ is

$$\begin{aligned}
\hat{\lambda}_{\alpha\alpha} &= \frac{1}{n}\sum_{i=1}^{n} c\ell(\hat{\theta}_\alpha; Y^{(i)})^2 - \left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y)\right)^2, \\
\hat{\lambda}_{\alpha\beta} &= \frac{1}{n}\sum_{i=1}^{n} c\ell(\hat{\theta}_\alpha; Y^{(i)})c\ell(\hat{\theta}_\beta; Y^{(i)}) - \left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y)\right)\left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y)\right), \\
\hat{\lambda}_{\beta\alpha} &= \frac{1}{n}\sum_{i=1}^{n} c\ell(\hat{\theta}_\beta; Y^{(i)})c\ell(\hat{\theta}_\alpha; Y^{(i)}) - \left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y)\right)\left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\alpha; Y)\right), \\
\hat{\lambda}_{\beta\beta} &= \frac{1}{n}\sum_{i=1}^{n} c\ell(\hat{\theta}_\beta; Y^{(i)})^2 - \left(\frac{1}{n}c\ell^{(n)}(\hat{\theta}_\beta; Y)\right)^2.
\end{aligned}$$

because $\hat{\Lambda} \xrightarrow{a.s} \Lambda$. This leads our variance estimator for $(C^{(n)}_\alpha - C^{(n)}_\beta)$ to be

$$\begin{aligned}
\mathrm{var}(C^{(n)}_\alpha - C^{(n)}_\beta) &= \left[\frac{1}{n}\sum_{i=1}^{n}\left(c\ell(\hat{\theta}_\alpha; Y^{(i)}) - c\ell(\hat{\theta}_\beta; Y^{(i)})\right)^2 - \left(\frac{c\ell^{(n)}(\hat{\theta}_\alpha; Y)}{n} - \frac{c\ell^{(n)}(\hat{\theta}_\beta; Y)}{n}\right)^2\right]/n \\
&= V^{(n)}_{\alpha\beta}/n,
\end{aligned}$$

and the test statistic to be

$$T \quad = \quad \frac{C_{\alpha}^{(n)} - C_{\beta}^{(n)}}{\sqrt{V_{\alpha\beta}^{(n)}/n}}.$$

□

D Appendix: Proofs for Section 4.3.2 composite likelihood with local alternative setting

In a neighborhood of θ^*_α, there is a Jacobian matrix for mapping of θ_α. we denote it as $B^i_{\alpha j}(\theta_\alpha) = \partial\phi^i_\alpha / \partial\theta^j_\alpha$, where $\partial/\partial\theta^j_\alpha$ denotes the partial differentiation with respect to the j-th element of θ_α. The Jacobian matrix is of rank m_α. Let $\theta^{(n)}_\alpha = \arg\sup_{\theta_\alpha \in \Theta_\alpha} \mathrm{KL}(\phi^{(n)}, \phi_\alpha(\theta_\alpha))$ and assume $\lim_{n\to+\infty} \sqrt{n}(\theta^{(n)}_\alpha - \theta^*_\alpha) = \theta^\diamond_\alpha$ exists. Let $\hat{\theta}^{(n)}_\alpha = \arg\sup_{\theta_\alpha \in \Theta_\alpha} c\ell^{(n)}(\phi_\alpha(\theta_\alpha))$ denote the MLE under model α. Assume $\mathrm{p}\lim_{n\to\infty} \sqrt{n}(\hat{\theta}^{(n)}_\alpha - \theta^*_\alpha) = \hat{\theta}^\diamond_\alpha = O_p(1)$.

For notational simplicity, we write $\hat{\phi}^{(n)}_\alpha = \phi_\alpha(\hat{\theta}^{(n)}_\alpha)$, $\hat{\phi}^\diamond_\alpha = \mathrm{p}\lim_{n\to\infty} \sqrt{n}(\hat{\phi}^{(n)}_\alpha - \phi^*)$, $B^*_\alpha = B_\alpha(\theta^*_\alpha)$, $\phi^{(n)}_\alpha = \phi_\alpha(\theta^{(n)}_\alpha)$ and $\phi^\diamond_\alpha = \lim_{n\to+\infty} \sqrt{n}(\phi^{(n)}_\alpha - \phi^*)$. The following lemmas motivate the main result of this paper. The arguments involve expansions that are standard in full likelihood considerations and are similar to those leading the Linhart's test statistics discussed previously. However, the results are presented here in more general context using composite likelihood.

Lemma D.0.1. *Let $\phi^{(n)}_\alpha \in \Xi, \alpha = 1, 2$ be the sequences converging to ϕ^*, such that*

$\lim_{n\to\infty}\sqrt{n}(\phi_\alpha^{(n)} - \phi^*) = \phi_\alpha^\diamond$, $\alpha = 1, 2$, *exist. then we have*

$$\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_2^{(n))}) = (\phi_1^\diamond - \phi_2^\diamond)' H^*(\phi_1^\diamond - \phi_2^\diamond)/2, \tag{D.1}$$

where $H^* = H(\phi^*)$.

Proof. Lemma D.0.1:

Consider the definition of $D_{KL}(\xi_1, \xi_2)$ and $D(\phi_1, \phi_2)$, we can write

$$\begin{aligned} D_{KL}(\phi_1, \phi_2) &= D(\phi_1, \phi_1) - D(\phi_1, \phi_2) \\ &= E_{\phi_1}\left[\sum_{k=1}^K w_k \log g(Y_k|\phi_1)\right] - E_{\phi_1}\left[\sum_{k=1}^K w_k \log g(Y_k|\phi_2)\right] \end{aligned}$$

Take Taylor expansion to $D(\phi_1, \phi_2)$ with respect to ϕ_2 around ϕ_1,we have

$$\begin{aligned} D_{KL}(\phi_1, \phi_2) &= E_{\phi_1}\left[\sum_{k=1}^K w_k \log g(Y_k|\phi_1)\right] - E_{\phi_1}\left[\sum_{k=1}^K w_k \log g(Y_k|\phi_1)\right] \\ &\quad -(\phi_2 - \phi_1)E_{\phi_1}\left[\sum_{k=1}^K w_k \frac{\partial \log g(Y_k|\phi_1)}{\partial \phi_1}\right] \\ &\quad -\frac{1}{2}E\left[\sum_{k=1}^K w_k \frac{\partial^2 \log g(Y_k|\phi_1)}{\partial \phi_1 \partial \phi_1'}\right](\phi_1 - \phi_2)(\phi_1 - \phi_2)' \\ &\quad +o\|\phi_1 - \phi_2\|^2. \end{aligned}$$

Note $E\left[\sum_{k=1}^K w_k \frac{\partial \log g(Y_k|\phi_1)}{\partial \phi_1}\right] = 0$, and $E\left[\sum_{k=1}^K w_k \frac{\partial^2 \log g(Y_k|\phi_1)}{\partial \phi_1 \partial \phi_1'}\right] = H(\phi_1)$. Times n to the both side of the above equation, we get

$$nD_{KL}(\phi_1, \phi_2) = H(\phi_1)\left(\sqrt{n}(\phi_1 - \phi_2)\right)\left(\sqrt{n}(\phi_1 - \phi_2)'\right)/2 + o\|\sqrt{n}(\phi_1 - \phi_2)\|^2.$$

Because $\lim_{n\to\infty} \sqrt{n}(\phi_1^{(n)} - \phi_2^{(n)}) = \phi_1^\diamond - \phi_2^\diamond$, and $\lim_{n\to\infty} H(\phi_1^{(n)}) = H^*$, we are able to derive

$$\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_2^{(n)}) = (\phi_1^\diamond - \phi_2^\diamond)' H^* (\phi_1^\diamond - \phi_2^\diamond)/2.$$

We complete proof for Lemma D.0.1 equation (D.1). □

Lemma D.0.2. *The asymptotic limit $\phi_\alpha^\diamond$ in model-α satisfies*

$$\phi_\alpha^\diamond = B_\alpha^* \theta_\alpha^\diamond, \tag{D.2a}$$

$$B_\alpha^{*\prime} H^* (\phi_\alpha^\diamond - \phi^\diamond) = 0, \tag{D.2b}$$

$$\theta_\alpha^\diamond = (B_\alpha^{*\prime} H^* B_\alpha^*)^{-1} B_\alpha^{*\prime} H^* \phi^\diamond. \tag{D.2c}$$

Note that $\phi_\alpha^\diamond$ is the projection of $\phi^\diamond$ onto $I_m B_\alpha^*$, the linear space spanned by the column vectors of B_α^*. Using H^* as the metric.

Proof. Lemma D.0.2 (D.2a):

Expand $\phi_\alpha(\theta_\alpha^{(n)})$ at $\phi_\alpha(\theta_\alpha^*)$, we have

$$\phi_\alpha(\theta_\alpha^{(n)}) = \phi_\alpha(\theta_\alpha^*) + \frac{\partial \phi_\alpha}{\partial \theta_\alpha}(\theta_\alpha^{(n)} - \theta_\alpha^*) + o\|\theta_\alpha^{(n)} - \theta_\alpha^*\|,$$

Since $B_\alpha(\theta^*_\alpha) = \frac{\partial \phi_\alpha}{\partial \theta_\alpha}$, we have

$$\phi_\alpha(\theta^{(n)}_\alpha) = \phi_\alpha(\theta^*_\alpha) + B_\alpha(\theta^*_\alpha)(\theta^{(n)}_\alpha - \theta^*_\alpha) + o(||\theta^{(n)}_\alpha - \theta^*_\alpha||),$$

$$\sqrt{n}\phi_\alpha(\theta^{(n)}_\alpha) = \sqrt{n}\phi_\alpha(\theta^*_\alpha) + \sqrt{n}B_\alpha(\theta^*_\alpha)(\theta^{(n)}_\alpha - \theta^*_\alpha) + o(||\theta^{(n)}_\alpha - \theta^*_\alpha||),$$

$$\sqrt{n}\left[\phi_\alpha(\theta^{(n)}_\alpha) - \phi_\alpha(\theta^*_\alpha)\right] = \sqrt{n}B_\alpha(\theta^*_\alpha)(\theta^{(n)}_\alpha - \theta^*_\alpha) + o(||\theta^{(n)}_\alpha - \theta^*_\alpha||),$$

$$\lim_{n\to\infty} \sqrt{n}\left[\phi_\alpha(\theta^{(n)}_\alpha) - \phi_\alpha(\theta^*_\alpha)\right] = \lim_{n\to\infty} \sqrt{n}B_\alpha(\theta^*_\alpha)(\theta^{(n)}_\alpha - \theta^*_\alpha),$$

$$\lim_{n\to\infty} \sqrt{n}(\phi_\alpha - \phi) = \lim_{n\to\infty} \sqrt{n}B_\alpha(\theta^*_\alpha)(\theta^{(n)}_\alpha - \theta^*_\alpha),$$

$$\phi^\diamond_\alpha = B^*_\alpha \lim_{n\to\infty} \sqrt{n}(\theta^{(n)}_\alpha - \theta^*_\alpha) = B^*_\alpha\theta^\diamond_\alpha.$$

We complete the proof for the first equation (D.2a). □

Proof. Lemma D.0.2 (D.2b):

We have the definition $\theta^\diamond_\alpha = \lim_{n\to\infty} \arg\inf_{u\in\mathcal{R}^{\Updownarrow_\alpha}} nD_{KL}(\phi^{(n)}, \phi_\alpha(\theta^*_\alpha + u/\sqrt{n}))$. Because of the smoothness of the K-L discrepancy and $||\theta^\diamond_\alpha|| < \infty$, the *limit* and *arginf* can be exchanged, that is,

$$\theta^\diamond_\alpha = \arg\inf_{u\in\mathcal{R}^{\Updownarrow_\alpha}} \lim_{n\to\infty} nD_{KL}(\phi^{(n)}, \phi_\alpha(\theta^*_\alpha + u/\sqrt{n})).$$

Follow equation (D.1), we obtain

$$\begin{aligned}\theta^\diamond_\alpha &= \arg\inf_{u\in\mathcal{R}^{m_\alpha}}(\phi^\diamond - B^*_\alpha u)'H^*(\phi^\diamond - B^*_\alpha u)\\ &= \arg\inf_{u\in\mathcal{R}^{m_\alpha}}(\phi^{\diamond\prime}H^*\phi^\diamond - \phi^{\diamond\prime}H^*B^*_\alpha u - u'B^{*\prime}_\alpha H^*\phi^\diamond + u'B^{*\prime}_\alpha H^*B^*_\alpha u).\end{aligned}$$

Take differential of it at $u = \theta^{\diamond}$, we get

$$-2B_{\alpha}^{*\prime}H^{*}\phi^{\diamond} + 2B_{\alpha}^{*\prime}H^{*}B_{\alpha}^{*}u = 0,$$

$$B_{\alpha}^{*\prime}H^{*}(-\phi^{\diamond} + B_{\alpha}^{*}\theta_{\alpha}^{\diamond}) = 0.$$

We known $\phi_{\alpha}^{\diamond} = B_{\alpha}^{*}\theta_{\alpha}^{\diamond}$, hence

$$B_{\alpha}^{*\prime}H^{*}(\phi_{\alpha}^{\diamond} - \phi^{\diamond}) = 0.$$

Equation(D.2b) is proved completely . □

Proof. Lemma D.0.2 (D.2c):

From above, we know $B_{\alpha}^{*\prime}H^{*}(-\phi^{\diamond} + B_{\alpha}^{*}\theta_{\alpha}^{\diamond}) = 0$, so

$$B_{\alpha}^{*\prime}H^{*}B_{\alpha}^{*}\theta_{\alpha}^{\diamond} = B_{\alpha}^{*\prime}H^{*}\phi^{\diamond},$$

$$\theta^{\diamond} = (B_{\alpha}^{*\prime}H^{*}B_{\alpha}^{*})^{-1}B_{\alpha}^{*\prime}H^{*}\phi^{\diamond}.$$

We complete proof for the equation (D.2c). □

Lemma D.0.3. *Let $\hat{\phi}^{(n)}$ be the MLE of $\phi^{(n)}$ for $g(Y;\phi), \phi \in \Xi$. The asymptotic distribution of the MLE is normal,*

$$\hat{\phi}^{\diamond} \sim N(\phi^{\diamond}, G^{*-1}), \tag{D.3}$$

where $G^{*} = H^{*}J^{*-1}H^{*}$.

Proof. Lemma D.0.3:

Note when n is sufficiently large, we have $\frac{\partial c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}} = 0$, Expand $\frac{\partial c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}}$ around $\phi^{(n)}$:

$$0 = \frac{\partial c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}} + \frac{\partial^2 c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}\partial\phi^{(n)\prime}}(\hat{\phi}^{(n)} - \phi^{(n)}) + o_P(1).$$

We can write above equation as:

$$\sqrt{n}(\hat{\phi}^{(n)} - \phi^{(n)}) = \left(\frac{1}{\sqrt{n}}\frac{\partial c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}}\right) / \left(-\frac{1}{n}\frac{\partial^2 c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}\partial\phi^{(n)\prime}}\right)$$

By the assumption 4.3.1, we know $\left(\frac{1}{\sqrt{n}}\frac{\partial c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}}\right) \sim N(0, J^*)$. By the definition, we know $\left(-\frac{1}{n}\frac{\partial^2 c\ell^{(n)}(\phi^{(n)})}{\partial\phi^{(n)}\partial\phi^{(n)\prime}}\right) \xrightarrow{P} H^*$. From the assumption 4.3.2, we know $\operatorname{p\,lim}_{n\to\infty} \sqrt{n}(\hat{\phi}^{(n)} - \phi^*) = \hat{\phi}^\diamond = O_p(1)$. Therefore, $\phi^\diamond \sim N(\phi^\diamond, H^{*-1}J^*H^{*-1})$, which is $N(\phi^\diamond, G^{*-1})$. In deed, G is Godambe information as shown in equation (4.2). Hence, Lemma D.0.3 is proved. □

Lemma D.0.4. *The asymptotic limit $\hat{\phi}^\diamond_\alpha$ of the MLE for model-α satisfies*

$$\hat{\phi}^\diamond_\alpha = B^*_\alpha\hat{\theta}^\diamond_\alpha, \tag{D.4a}$$

$$B^{*\prime}_\alpha H^*(\hat{\phi}^\diamond_\alpha - \hat{\phi}^\diamond) = 0, \tag{D.4b}$$

$$\hat{\theta}^\diamond_\alpha = (B^{*\prime}_\alpha H^* B^*_\alpha)^{-1} B^{*\prime}_\alpha H^* \hat{\phi}^\diamond. \tag{D.4c}$$

Proof. Lemma D.0.4 (D.4a):

This proof is similar to (D.2a) ∼ (D.2c). Expand $\phi_\alpha(\hat{\theta}_\alpha^{(n)})$ at $\phi_\alpha(\hat{\theta}_\alpha^*)$, we have

$$\phi_\alpha(\hat{\theta}_\alpha^{(n)}) = \phi_\alpha(\hat{\theta}_\alpha^*) + \tfrac{\partial\phi_\alpha}{\partial\theta_\alpha}|_{\theta_\alpha^*(\hat{\theta}_\alpha^{(n)}-\hat{\theta}_\alpha^*)} + o||\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*||,$$

$$\phi_\alpha(\hat{\theta}_\alpha^{(n)}) = \phi_\alpha(\hat{\theta}_\alpha^*) + B_\alpha^*(\hat{\theta}_\alpha)(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*) + o(||\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*||),$$

$$\sqrt{n}\phi_\alpha(\hat{\theta}_\alpha^{(n)}) = \sqrt{n}\phi_\alpha(\hat{\theta}_\alpha^*) + \sqrt{n}B_\alpha^*(\hat{\theta}_\alpha)(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*) + o(\sqrt{n}||\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*||),$$

$$\sqrt{n}[\phi_\alpha(\hat{\theta}_\alpha^{(n)}) - \phi_\alpha(\hat{\theta}_\alpha^*)] = \sqrt{n}B_\alpha^*(\hat{\theta}_\alpha)(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*) + o_p(1),$$

$$\lim_{n\to\infty}\sqrt{n}[\phi_\alpha(\hat{\theta}_\alpha^{(n)}) - \phi_\alpha(\hat{\theta}_\alpha^*)] = \lim_{n\to\infty}\sqrt{n}B_\alpha^*(\hat{\theta}_\alpha)(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*),$$

$$\lim_{n\to\infty}\sqrt{n}(\hat{\phi}_\alpha - \hat{\phi}_\alpha^*) = \lim_{n\to\infty}\sqrt{n}B_\alpha^*(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*),$$

$$\hat{\phi}_\alpha^\diamond = B_\alpha^* \lim_{n\to\infty}\sqrt{n}(\hat{\theta}_\alpha^{(n)} - \hat{\theta}_\alpha^*)$$

$$\hat{\phi}_\alpha^\diamond = B_\alpha^*\hat{\theta}_\alpha^\diamond.$$

We complete the proof for the first equation (D.4a). □

Proof. Lemma D.0.4 (D.4b):

Consider $\frac{\partial\phi_\alpha^i}{\partial\theta_\alpha^j}c\ell^{(n)}(\phi_\alpha(\hat{\theta}_\alpha^{(n)})) = \sum_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)})\frac{\partial_j c\ell^{(n)}(\hat{\phi}_\alpha^{(n)})}{\partial\hat{\phi}_{\alpha j}^{(n)}} = 0$ for sufficiently large n. Time it with $\sqrt{n}$ and expand it with respect to $\hat{\phi}_\alpha^{(n)}$ around $\hat{\phi}^{(n)}$ to obtain

$$\begin{aligned}
&\sqrt{n}\sum_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)})\frac{\partial_j c\ell^{(n)}(\hat{\phi}_\alpha^{(n)})}{\partial\hat{\phi}_{\alpha j}^{(n)}} \\
=&\sum_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)})\left(\sqrt{n}\frac{\partial_j c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}}\right) + \sum_j \partial_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)})\frac{\partial_j^2 c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}\partial\hat{\phi}^{(n)\prime}}\sqrt{n}(\hat{\phi}_\alpha^{(n)} - \hat{\phi}^{(n)})^j + o_p(1) \\
=&0.
\end{aligned}$$

Note $\frac{\partial_j c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}} = 0$, $E\left(\frac{\partial_j^2 c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial\hat{\phi}^{(n)}\partial\hat{\phi}^{(n)\prime}}\right) = H^*$, $\hat{\phi}_\alpha^\diamond = \text{p}\lim_{n\to\infty} \sqrt{n}(\hat{\phi}_\alpha^{(n)} - \phi_\alpha^*)$, and $\hat{\phi}^\diamond = \text{p}\lim_{n\to\infty} \sqrt{n}(\hat{\phi}^{(n)} - \phi^*)$. Therefor, we get

$$\text{p}\lim_{n\to\infty} \sum_j B_{\alpha i}^j(\hat{\theta}_\alpha^{(n)}) \frac{\partial_j c\ell^{(n)}(\hat{\phi}_\alpha^{(n)})}{\partial\hat{\phi}_j^{(n)}} = B_\alpha^{*\prime} H^*(\hat{\phi}_\alpha^\diamond - \hat{\phi}^\diamond) = 0.$$

We proved (D.4b). □

Proof. Lemma D.0.4 (D.4c):

From the above equation $B_\alpha^{*\prime} H^*(\hat{\phi}_\alpha^\diamond - \hat{\phi}^\diamond) = 0$ and $\hat{\phi}_\alpha^\diamond = B_\alpha^* \theta_\alpha^\diamond$, we know

$$B_\alpha^{*\prime} H^* \hat{\phi}_\alpha^\diamond = B_\alpha^{*\prime} H^* \hat{\phi}^\diamond,$$

$$B_\alpha^{*\prime} H^* B_\alpha^* \hat{\theta}_\alpha^\diamond = B_\alpha^{*\prime} H^* \hat{\phi}^\diamond,$$

$$\hat{\theta}_\alpha^\diamond = (B_\alpha^{*\prime} H^* B_\alpha^*)^{-1} B_\alpha^{*\prime} H^* \hat{\phi}^\diamond.$$

We proved equation(D.4c). □

Let $\phi_\alpha^\dagger = H^{*1/2}\phi_\alpha^\diamond$, $\hat{\phi}_\alpha^\dagger = H^{*1/2}\hat{\phi}_\alpha^\diamond$ and $B_\alpha^\dagger = H^{*1/2}B_\alpha^*$. Note that H^* is a square root decomposition, i.e. $H^* = (H^{*1/2})' H^{*1/2}$.

Proposition D.0.1. *Define* $P_\alpha^\dagger = B_\alpha^\dagger(B_\alpha^{\dagger\prime} B_\alpha^\dagger)^{-1} B_\alpha^{\dagger\prime}$ *the projection operator of* $I_m B_\alpha^\dagger$*. Letting* $\phi_\alpha^\dagger = H^{*1/2}\phi_\alpha^\diamond$*, equation* (D.1) *in Lemma* D.0.1*, equation* (D.2a) *in Lemma* D.0.2*, equa-*

tion(D.4a) *in Lemma* D.0.4*, and equation*(D.3) *in lemma* D.0.3 *will be*

$$\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_2^{(n))}) = \|\phi_1^\dagger - \phi_2^\dagger\|^2/2, \tag{D.5a}$$

$$\phi_\alpha^\dagger = P_\alpha^\dagger \phi^\dagger, \tag{D.5b}$$

$$\hat{\phi}_\alpha^\dagger = P_\alpha^\dagger \hat{\phi}^\dagger, \tag{D.5c}$$

$$\hat{\phi}^\dagger \sim N(\phi^\dagger, W^*). \tag{D.5d}$$

where $W = H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}}$

Proof. Proposition D.0.1 (D.5a):

From equation (D.1) $\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_2^{(n)}) = (\phi_1^\diamond - \phi_2^\diamond)' H^* (\phi_1^\diamond - \phi_2^\diamond)/2$, and because $\phi^\dagger = H^{*\frac{1}{2}} \phi^\diamond$ and $\phi^\diamond = H^{*-\frac{1}{2}} \phi^\dagger$, we can have

$$\begin{aligned}
\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_2^{(n)}) &= (H^{*-\frac{1}{2}}\phi_1^\dagger - H^{*-\frac{1}{2}}\phi_2^\dagger)' H^{*\frac{1}{2}} H^{*\frac{1}{2}} (H^{*-\frac{1}{2}}\phi_1^\dagger - H^{*-\frac{1}{2}}\phi_2^\dagger)/2 \\
&= (\phi_1^\dagger - \phi_2^\dagger)' H^{*-\frac{1}{2}\prime} H^{*1/2} H^{*1/2} H^{*-\frac{1}{2}} (\phi_1^\dagger - \phi_2^\dagger)/2 \\
&= (\phi_1^\dagger - \phi_2^\dagger)'(\phi_1^\dagger - \phi_2^\dagger)/2 \\
&= \|\phi_1^\dagger - \phi_2^\dagger\|^2/2
\end{aligned}$$

We proved equation (D.5a). □

Proof. Proposition D.0.1 (D.5b):

We know $\phi_\alpha^\dagger = H^{*\frac{1}{2}} B_\alpha^* \theta_\alpha^\diamond$ because we have $\phi_\alpha^\dagger = H^{*\frac{1}{2}} \phi_\alpha^\diamond$, and $\phi_\alpha^\diamond = B_\alpha^* \theta_\alpha^\diamond$. Also, because $\theta_\alpha^\diamond = (B_\alpha^{*\prime} H^* B_\alpha^*)^{-1} B_\alpha^{*\prime} H^* \phi^\diamond$, we obtain $\phi_\alpha^\dagger = H^{*\frac{1}{2}} B_\alpha^* (B_\alpha^{*\prime} H^* B_\alpha^*)^{-1} B_\alpha^{*\prime} H^* \phi^\diamond$. In addition, note

$B^{\dagger}_{\alpha} = H^{*\frac{1}{2}} B^{*}_{\alpha}$ and $B^{*}_{\alpha} = H^{*\frac{1}{2}} B^{\dagger}_{\alpha}$. Therefore, we get

$$\begin{aligned}
\phi^{\dagger}_{\alpha} &= H^{*\frac{1}{2}} H^{*-\frac{1}{2}} B^{\dagger}_{\alpha} [(H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})' H^{*} (H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})]^{-1} (H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})' H^{*} \phi^{\diamond} \\
&= B^{\dagger}_{\alpha} [B^{\dagger\prime}_{\alpha} H^{*-\frac{1}{2}\prime} H^{*} H^{*-\frac{1}{2}} B^{\dagger}_{\alpha}]^{-1} B^{\dagger\prime}_{\alpha} H^{*-\frac{1}{2}\prime} H^{*} \phi^{\diamond} \\
&= B^{\dagger}_{\alpha} [B^{\dagger\prime}_{\alpha} B^{\dagger}_{\alpha}]^{-1} B^{\dagger\prime}_{\alpha} H^{*\frac{1}{2}} \phi^{\diamond} \\
&= B^{\dagger}_{\alpha} [B^{\dagger\prime}_{\alpha} B^{\dagger}_{\alpha}]^{-1} B^{\dagger\prime}_{\alpha} H^{*\frac{1}{2}\prime} (H^{*-\frac{1}{2}} \phi^{\dagger}) \\
&= B^{\dagger}_{\alpha} [B^{\dagger\prime}_{\alpha} B^{\dagger}_{\alpha}]^{-1} B^{\dagger\prime}_{\alpha} \phi^{\dagger} \\
&= P^{\dagger}_{\alpha} \phi^{\dagger}
\end{aligned}$$

We proved equation (D.5b). □

Proof. Proposition D.0.1 (D.5c):

Similar to the proof of equation (D.5b), we know $\hat{\phi}^{\dagger}_{\alpha} = H^{*\frac{1}{2}} \hat{\phi}^{\diamond}_{\alpha}$, $\hat{\phi}^{\diamond} = B^{*}_{\alpha} \hat{\theta}^{\diamond}_{\alpha}$, $\hat{\theta}^{\diamond}_{\alpha} = (B^{*\prime}_{\alpha} H^{*} B^{*}_{\alpha})^{-1} B^{*\prime}_{\alpha} H^{*} \hat{\phi}^{\diamond}$, $B^{*}_{\alpha} = H^{-\frac{1}{2}} B^{\dagger}_{\alpha}$, and $\hat{\phi}^{\diamond} = H^{*-\frac{1}{2}} \phi^{\dagger}$. We can get

$$\begin{aligned}
\hat{\phi}^{\dagger}_{\alpha} &= H^{*\frac{1}{2}} (B^{*}_{\alpha} \hat{\theta}^{\diamond}_{\alpha}) \\
&= H^{*\frac{1}{2}} B^{*}_{\alpha} (B^{*\prime}_{\alpha} H^{*} B_{\alpha})^{-1} B^{*\prime}_{\alpha} H^{*} \hat{\phi}^{\diamond} \\
&= H^{*\frac{1}{2}} (H^{*-\frac{1}{2}} B^{\dagger}_{\alpha}) [(H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})' H^{*} (H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})]^{-1} (H^{*-\frac{1}{2}} B^{\dagger}_{\alpha})' H^{*} H^{*-\frac{1}{2}} \hat{\phi}^{\dagger} \\
&= B^{\dagger}_{\alpha} [B^{\dagger\prime}_{\alpha} H^{*-\frac{1}{2}\prime} H^{*} H^{*-\frac{1}{2}} B^{\dagger}_{\alpha}]^{-1} B^{\dagger\prime}_{\alpha} H^{*-\frac{1}{2}\prime} H^{*} H^{*-\frac{1}{2}} \hat{\phi}^{\dagger} \\
&= B^{\dagger}_{\alpha} (B^{\dagger\prime}_{\alpha} B^{\dagger}_{\alpha})^{-1} B^{\dagger\prime}_{\alpha} \hat{\phi}^{\dagger} \\
&= P^{\dagger}_{\alpha} \hat{\phi}^{\dagger}
\end{aligned}$$

Equation (D.5c) is proved. □

Proof. Proposition D.0.1 (D.5d):

From equation (D.3), we know $\hat{\phi}^{\diamond} \sim N(\phi^{\dagger}, H^{*-1}J^{*}H^{*-1})$, which means $\text{var}(\hat{\phi}^{\diamond}) = H^{*-1}J^{*}H^{*-1}$

We also know $\hat{\phi}^{\dagger} = H^{*\frac{1}{2}}\hat{\phi}^{\diamond}$. Hence,

$$\begin{aligned} \text{var}(\hat{\phi}^{\dagger}) &= \text{var}(H^{*\frac{1}{2}}\hat{\phi}^{\diamond}) \\ &= H^{*\frac{1}{2}}H^{-1}J^{*}H^{*-1}H^{*\frac{1}{2}} \\ &= H^{*-\frac{1}{2}}J^{*}H^{*-\frac{1}{2}} \end{aligned}$$

Finally, we obtain $\hat{\phi}^{\dagger} \sim N(\phi^{\dagger}, W^{*})$, where $W^{*} = H^{*-\frac{1}{2}}J^{*}H^{*-\frac{1}{2}}$. Equation (D.5d) is proved.

□

Lemma D.0.5. *For* $\alpha, \beta \in \mathcal{M}$

$$\text{p}\lim_{n\to\infty}\left(\hat{c\ell}_{\alpha}^{(n)} - c\ell^{(n)}(\phi^{*})\right) = \|\hat{\phi}_{\alpha}^{\dagger}\|^{2}/2, \tag{D.6a}$$

$$\text{p}\lim_{n\to\infty}\left(\hat{c\ell}_{\beta}^{(n)} - c\ell^{(n)}(\phi^{*})\right) = \|\hat{\phi}_{\beta}^{\dagger}\|^{2}/2, \tag{D.6b}$$

$$\text{p}\lim_{n\to\infty}(\hat{c\ell}_{\alpha}^{(n)} - \hat{c\ell}_{\beta}^{(n)}) = \hat{\phi}^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\hat{\phi}^{\dagger}/2, \tag{D.6c}$$

$$\text{p}\lim_{n\to\infty} n(C_{\alpha}^{(n)} - C_{\beta}^{(n)}) = -\hat{\phi}^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\hat{\phi}^{\dagger}/2 + \text{tr}[J_{\alpha}^{*}H_{\alpha}^{*-1} - J_{\beta}^{*}H_{\beta}^{*-1}], \tag{D.6d}$$

$$c\ell E(nC_{\alpha}^{(n)} - nC_{\beta}^{(n)}) = -\left[\phi^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\phi^{\dagger}\right]/2 + \text{tr}\left[(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})W^{*})\right]/2, \tag{D.6e}$$

$$c\ell V(nC_{\alpha}^{(n)} - nC_{\beta}^{(n)}) = \phi^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})'W^{*}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\phi^{\dagger} + \text{tr}[(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})W^{*}]^{2}/2. \tag{D.6f}$$

where $W^* = H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}}$, $P^\dagger_\alpha = B^\dagger_\alpha (B^{\dagger\prime}_\alpha B^\dagger_\alpha)^{-1} B^{\dagger\prime}_\alpha$, $P^\dagger_\beta = B^\dagger_\beta (B^{\dagger\prime}_\beta B^\dagger_\beta)^{-1} B^{\dagger\prime}_\beta$. *The asymptotic expectation and variance are denoted as* $c\ell E$ *and* $c\ell V$*, respectively.*

Proof. Lemma D.0.5 (D.6a):

Expand $c\ell^{(n)}(\hat{\phi}^{(n)}_\alpha)/n$ with respect to $\hat{\phi}^{(n)}_\alpha$ around $\hat{\phi}^{(n)}$, we get

$$\begin{aligned} c\ell^{(n)}(\hat{\phi}^{(n)}_\alpha)/n &= c\ell^{(n)}/n + \frac{\partial c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)}}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)})/n + \\ & \quad \frac{1}{2n}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)})' \frac{\partial^2 c\ell(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)} \partial \hat{\phi}^{(n)\prime}}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)}) + o_p(1). \end{aligned}$$

We know $n^{-1}\frac{\partial c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)}} = 0$. Move $c\ell^{(n)}(\hat{\phi}^{(n)})/n$ to the left side and times n to both side. We get

$$n(c\ell^{(n)}(\hat{\phi}^{(n)})/n - c\ell^{(n)}(\hat{\phi}^{(n)}_\alpha)/n) = \frac{1}{2}\sqrt{n}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)})' n^{-1} \frac{\partial^2 c\ell(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)} \partial \hat{\phi}^{(n)\prime}} \sqrt{n}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)}) + o_p(1)$$

Take the limitation for the above equation. Note $\operatorname{p\,lim}_{n\to\infty} n^{-1}\frac{\partial^2 c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)} \partial \hat{\phi}^{(n)\prime}} = H^*$. In addition, by equation (D.1) and (D.5a), we are able to obtain

$$\begin{aligned} & \operatorname{p}\lim_{n\to\infty} n(c\ell^{(n)}(\hat{\phi}^{(n)})/n - c\ell^{(n)}(\hat{\phi}^{(n)}_\alpha)/n) \\ =& \operatorname{p}\lim_{n\to\infty}[\frac{1}{2}\sqrt{n}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)})' n^{-1} \frac{\partial^2 c\ell^{(n)}(\hat{\phi}^{(n)})}{\partial \hat{\phi}^{(n)} \partial \hat{\phi}^{(n)\prime}} \sqrt{n}(\hat{\phi}^{(n)}_\alpha - \hat{\phi}^{(n)})^j \\ =& (\hat{\phi}^\diamond_\alpha - \hat{\phi}^\diamond)' H^* (\hat{\phi}^\diamond_\alpha - \hat{\phi}^\diamond)/2 \\ =& \|\hat{\phi}^\dagger_\alpha - \hat{\phi}^\dagger\|^2/2. \end{aligned}$$

Replacing $\hat{\phi}_\alpha^{(n)}$ by ϕ^*, the above equation becomes

$$
\begin{aligned}
\mathrm{p}\lim_{n\to\infty} n(c\ell^{(n)}(\hat{\phi}^{(n)})/n - c\ell^{(n)}(\phi^*)/n) &= (\hat{\phi}^\diamond H^* \hat{\phi}^\diamond)/2 \\
&= (H^{*1/2}\hat{\phi}^\diamond H^{*1/2}\hat{\phi}^\diamond)/2 \\
&= (\hat{\phi}^\dagger \hat{\phi}^\dagger)/2 \\
&= \|\hat{\phi}^\dagger\|^2/2
\end{aligned}
$$

Take the difference of these two equations, we obtain

$$
\begin{aligned}
&\mathrm{p}\lim_{n\to\infty} n(c\ell^{(n)}(\hat{\phi}^{(n)})/n - c\ell^{(n)}(\phi^*)/n) - \mathrm{p}\lim_{n\to\infty} n(c\ell^{(n)}(\hat{\phi}^{(n)})/n - c\ell^{(n)}(\hat{\phi}_\alpha^{(n)})/n) \\
&\mathrm{p}\lim_{n\to\infty}(c\ell^{(n)}(\hat{\phi}_\alpha^{(n)}) - c\ell^{(n)}(\phi^*)) \\
&= \|\hat{\phi}^\dagger\|^2/2 - \|\widehat{\phi_\alpha^\dagger} - \hat{\phi}^\dagger\|^2/2.
\end{aligned}
$$

Note $\hat{\phi}^{\dagger}_{\alpha} = P^{\dagger}_{\alpha}\hat{\phi}^{\dagger}$ and $P^{\dagger 2}_{\alpha} = P^{\dagger}_{\alpha}$. We can rewrite above equation as,

$$
\begin{aligned}
&\operatorname{p}\lim_{n\to\infty}(c\ell^{(n)}(\hat{\phi}^{(n)}_{\alpha}) - c\ell^{(n)}(\phi^{*}))\\
&= \left(||\hat{\phi}^{\dagger}||^{2} - ||\hat{\phi}^{\dagger} - \hat{\phi^{\dagger}_{\alpha}}||^{2}\right)/2\\
&= \left(||\hat{\phi}^{\dagger}||^{2} - ||\hat{\phi}^{\dagger} - P^{\dagger}_{\alpha}\hat{\phi}^{\dagger}||^{2}\right)/2\\
&= \left(||\hat{\phi}^{\dagger}||^{2} - ||(I_m - P^{\dagger}_{\alpha})\hat{\phi}^{\dagger}||^{2}\right)/2\\
&= \left(\hat{\phi}^{\dagger\prime}\hat{\phi}^{\dagger} - \hat{\phi}^{\dagger\prime}(I_m - P^{\dagger}_{\alpha})\hat{\phi}^{\dagger}\right)/2\\
&= \left(\hat{\phi}^{\dagger\prime}(I_m - I_m + P^{\dagger}_{\alpha})\hat{\phi}^{\dagger}\right)/2\\
&= ||P^{\dagger}_{\alpha}\hat{\phi}^{\dagger}||^{2}/2\\
&= ||\hat{\phi}^{\dagger}_{\alpha}||^{2}/2
\end{aligned}
$$

We proved equation (D.6a). □

Proof. Lemma D.0.5 (D.6b):

By applying exactly same proof approaches as Lemma D.0.5 (D.6b) but under model β, we can show $\operatorname{p}\lim_{n\to\infty}(c\ell^{(n)}(\hat{\phi}^{(n)}_{\alpha}) - c\ell^{(n)}(\phi^{*})) = ||\hat{\phi}^{\dagger}_{\beta}||^{2}/2$ □

Proof. Lemma D.0.5 (D.6c):

From above proof, we are able to get

$$\operatorname{p}\lim_{n\to\infty}(\hat{c\ell}^{(n)}_{\alpha} - c\ell^{(n)}(\phi^{*})) - \operatorname{p}\lim_{n\to\infty}(\hat{c\ell}^{(n)}_{\beta} - c\ell^{(n)}(\phi^{*})) = (||\hat{\phi^{\dagger}_{\alpha}}||^{2} - ||\hat{\phi}^{\dagger}_{\beta}||^{2})/2$$

$$\begin{aligned}
&\operatorname*{p\,lim}_{n\to\infty}(\hat{c\ell}_{\alpha}^{(n)} - \hat{c\ell}_{\beta}^{(n)}) \\
&\qquad = (\|\widehat{\phi_{\alpha}^{\dagger}}\|^2 - \|\hat{\phi}_{\beta}^{\dagger}\|^2)/2 \\
&\qquad = (\widehat{\phi_{\alpha}^{\dagger}}' \widehat{\phi_{\alpha}^{\dagger}} - \hat{\phi}_{\beta}^{\dagger\prime}\hat{\phi}_{\beta}^{\dagger})/2 \\
&\qquad = [(P_{\alpha}^{\dagger}\hat{\phi}^{\dagger})' P_{\alpha}^{\dagger}\hat{\phi}^{\dagger} - (P_{\beta}^{\dagger}\hat{\phi}^{\dagger})' P_{\beta}^{\dagger}\hat{\phi}^{\dagger}]/2 \\
&\qquad = [\hat{\phi}^{\dagger\prime}(P_{\alpha}^{\dagger\prime}P_{\alpha}^{\dagger} - P_{\beta}^{\dagger\prime}P_{\beta}^{\dagger})\hat{\phi}^{\dagger}]/2 \\
&\qquad = \hat{\phi}^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\hat{\phi}^{\dagger}/2
\end{aligned}$$

We complete the proof for equation (D.6c). □

Proof. Lemma D.0.5 (D.6d):

$$\begin{aligned}
&\operatorname*{p\,lim}_{n\to\infty} n(C_{\alpha}^{(n)} - C_{\beta}^{(n)}) \\
&= \operatorname*{p\,lim}_{n\to\infty} n[-\hat{c\ell}_{\alpha}^{(n)}/n + \mathrm{tr}(J_{\alpha}^{*}H_{\alpha}^{*-1})/n + \hat{c\ell}_{\beta}^{(n)}/n - \mathrm{tr}(J_{\beta}^{*}H_{\beta}^{*-1})/n] \\
&= -\operatorname*{p\,lim}_{n\to\infty}(\hat{c\ell}_{\alpha}^{(n)} - \hat{c\ell}_{\beta}^{(n)}) + \operatorname*{p\,lim}_{n\to\infty}[\mathrm{tr}(J_{\alpha}^{*}H_{\alpha}^{*-1}) - \mathrm{tr}(J_{\beta}^{*}H_{\beta}^{*-1})] \\
&= -\hat{\phi}^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\hat{\phi}^{\dagger}/2 + \mathrm{tr}[(J_{\alpha}^{*}H_{\alpha}^{*-1}) - (J_{\beta}^{*}H_{\beta}^{*-1})].
\end{aligned}$$

Equation (D.6d) is proved. □

Proof. Lemma D.0.5 (D.6e):

From equation (D.5d), we know $\widehat{\phi_{\alpha}^{\dagger}} \sim N(\phi^{\dagger}, W^{*}))$. Let $\hat{\eta}^{\dagger} = W^{*-\frac{1}{2}}\hat{\phi}^{\dagger}$. We have $\hat{\eta}^{\dagger} \sim N(W^{*-\frac{1}{2}}\phi^{\dagger}, I_m)$. Because $\hat{\eta}^{\dagger} = W^{*-\frac{1}{2}}\hat{\phi}^{\dagger}$, we get $\hat{\phi}^{\dagger} = W^{*\frac{1}{2}}\hat{\eta}^{\dagger}$. Note $E(X'AX) = b'Ab + \mathrm{tr}A$ and $\mathrm{Var}(X'AX) = 4b'A^2b + 2\mathrm{tr}A^2$, where X is a $m \times 1$ random vector and distributed as

$N(b, I_m)$, and A is a $m \times m$ symmetric matrix.

$$\begin{aligned}
& c\ell E(nC_\alpha^{(n)} - nC_\beta^{(n)}) \\
= \; & E\left[-\hat{\phi}^{\dagger\prime}(P_\alpha^\dagger - P_\beta^\dagger)\hat{\phi}^\dagger/2 + \text{tr}(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1})\right] \\
= \; & E\left[-(W^{*\frac{1}{2}}\hat{\eta}^\dagger)'(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}}\hat{\eta}^\dagger/2 + \text{tr}(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1})\right] \\
= \; & -E[\underbrace{\hat{\eta}^{\dagger\prime}}_{\text{X}}\underbrace{W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}}}_{\text{A}}\underbrace{\hat{\eta}^\dagger}_{\text{X}}]/2 + \text{tr}(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1}) \\
= \; & -\frac{1}{2}[(W^{*-\frac{1}{2}}\phi^\dagger)'(W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}})(W^{*-\frac{1}{2}}\phi^\dagger) \\
& +\text{tr}(W^{*\frac{1}{2}}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}})] + \text{tr}\left(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1}\right) \\
= \; & -\phi^{\dagger\prime}(P_\alpha^\dagger - P_\beta^\dagger)\phi^\dagger/2 - \text{tr}[(P_\alpha^\dagger - P_\beta^\dagger)W^*]/2 + \text{tr}(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1})
\end{aligned}$$

Now, we want to prove $\text{tr}\left[(P_\alpha^\dagger - P_\beta^\dagger)W^*\right] = \text{tr}\left(J_\alpha^* H_\alpha^{*-1} - J_\beta^* H_\beta^{*-1}\right)$. This is equivalent to prove $\text{tr}(P_\alpha^\dagger W^*) = \text{tr}(J_\alpha^* H_\alpha^{*-1})$, and $\text{tr}(P_\beta^\dagger W^*) = \text{tr}(J_\beta^* H_\beta^{*-1})$. Before we start prove, let's recall:

$$\begin{aligned}
P_\alpha^\dagger &= B_\alpha^\dagger(B_\alpha^{\dagger\prime}B_\alpha^\dagger)^{-1}B_\alpha^{\dagger\prime}, \\
B_\alpha^\dagger &= H^{*\frac{1}{2}}B_\alpha^*, \\
W^* &= H^{*-\frac{1}{2}}J^*H^{*-\frac{1}{2}}.
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{tr}(P_\alpha^\dagger W^*) &= \mathrm{tr}\left[B_\alpha^\dagger (B_\alpha^{\dagger\prime} B_\alpha^\dagger)^{-1} B_\alpha^{\dagger\prime} W^*\right] \\
&= \mathrm{tr}\left[H^{*\frac{1}{2}} B_\alpha^* \left((H^{*\frac{1}{2}} B_\alpha^*)' H^{*\frac{1}{2}} B_\alpha^*\right)^{-1} (H^{*\frac{1}{2}} B_\alpha^*)' H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}}\right] \\
&= \mathrm{tr}\left[(B_\alpha^{*\prime} J^* B_\alpha^*)(B_\alpha^{*\prime} H^* B_\alpha^*)^{-1}\right].
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
\mathrm{tr}(P_\beta^\dagger W^*) &= \mathrm{tr}\left[B_\beta^\dagger (B_\beta^{\dagger\prime} B_\beta^\dagger)^{-1} B_\beta^{\dagger\prime} W^*\right] \\
&= \mathrm{tr}\left[H^{*\frac{1}{2}} B_\beta^* \left((H^{*\frac{1}{2}} B_\beta^*)' H^{*\frac{1}{2}} B_\beta^*\right)^{-1} (H^{*\frac{1}{2}} B_\beta^*)' H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}}\right] \\
&= \mathrm{tr}\left[(B_\beta^{*\prime} J^* B_\beta^*)(B_\beta^{*\prime} H^* B_\beta^*)^{-1}\right].
\end{aligned}$$

If we can show $J_\alpha^* = B_\alpha^{*\prime} J^* B_\alpha^*$, $H_\alpha^* = B_\alpha^{*\prime} H^* B_\alpha^*$, $J_\beta^* = B_\beta^{*\prime} J^* B_\beta^*$ and $H_\beta^* = B_\beta^{*\prime} H^* B_\beta^*$, then the equation (D.6e) will be proved. Recall $B_\alpha^* = B_\alpha(\theta_\alpha^*) = \frac{\partial\phi_\alpha}{\partial\theta_\alpha}$ and$\phi^* = \phi_\alpha(\theta_\alpha^*)$. We can derive

$$\begin{aligned}
B_\alpha^{*\prime} J^* B_\alpha^* &= \left(\frac{\partial\phi_\alpha}{\partial\theta_\alpha}\right)' \mathrm{E}\left[\left(\frac{\partial c\ell(\phi^*, Y)}{\partial\phi^*}\right)\left(\frac{\partial c\ell(\phi^*, Y)}{\partial\phi^*}\right)'\right]\left(\frac{\partial\phi_\alpha}{\partial\theta_\alpha}\right) \\
J_\alpha^* &= \mathrm{E}\left[\left(\frac{\partial c\ell(\phi_\alpha(\theta_\alpha^*), Y)}{\partial\phi_\alpha(\theta_\alpha^*)}\right)\left(\frac{\partial c\ell(\phi_\alpha(\theta_\alpha^*), Y)}{\partial\phi_\alpha(\theta_\alpha^*)}\right)'\right] \\
&= \left(\frac{\partial\phi_\alpha}{\partial\theta_\alpha}\right)' \mathrm{E}\left[\left(\frac{\partial c\ell(\phi^*, Y)}{\partial\phi^*}\right)\left(\frac{\partial c\ell(\phi^*, Y)'}{\partial\phi^*}\right)\right]\left(\frac{\partial\phi_\alpha}{\partial\theta_\alpha}\right) \\
&= B_\alpha^{*\prime} J^* B_\alpha^*.
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
B_{\alpha}^{*\prime} H^{*} B_{\alpha}^{*} &= \left(\frac{\partial \phi_{\alpha}}{\partial \theta_{\alpha}}\right)' \mathrm{E}\left[\frac{\partial^{2} c\ell(\phi^{*}, Y)}{\partial \phi^{*} \phi^{*\prime}}\right]\left(\frac{\partial \phi_{\alpha}}{\partial \theta_{\alpha}}\right) \\
H_{\alpha}^{*} &= \mathrm{E}\left[\frac{\partial^{2} c\ell(\phi_{\alpha}(\theta_{\alpha}^{*}), Y)}{\partial \phi_{\alpha}(\theta_{\alpha}^{*}) \phi_{\alpha}(\theta_{\alpha}^{*})'}\right] \\
&= \left(\frac{\partial \phi_{\alpha}}{\partial \theta_{\alpha}}\right)' \mathrm{E}\left[\frac{\partial^{2} c\ell(\phi^{*}, Y)}{\partial \phi^{*} \phi^{*\prime}}\right]\left(\frac{\partial \phi_{\alpha}}{\partial \theta_{\alpha}}\right) \\
&= B_{\alpha}^{*\prime} H^{*} B_{\alpha}^{*}.
\end{aligned}$$

By applying same approach, we can show $B_{\beta}^{*\prime} J^{*} B_{\beta}^{*} = J_{\beta}^{*}$, and $B_{\beta}^{*\prime} H^{*} B_{\beta}^{*} = H_{\beta}^{*}$. Therefore, we proved $\mathrm{tr}\left((P_{\alpha}^{\dagger} - P_{\beta}^{\dagger}) W^{*}\right) = \mathrm{tr}(J_{\alpha}^{*} H_{\alpha}^{*-1} - J_{\beta}^{*} H_{\beta}^{*-1})$. Consequently,

$$\begin{aligned}
c\ell E(nC_{\alpha}^{(n)} - nC_{\beta}^{(n)}) &= -\phi^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\phi^{\dagger}/2 - \mathrm{tr}\left[(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger}) W^{*}\right]/2 + \mathrm{tr}\left(J_{\alpha}^{*} H_{\alpha}^{*-1} - J_{\beta}^{*} H_{\beta}^{*-1}\right) \\
&= -\phi^{\dagger\prime}(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger})\phi^{\dagger}/2 + \mathrm{tr}\left[(P_{\alpha}^{\dagger} - P_{\beta}^{\dagger}) W^{*}\right]/2.
\end{aligned}$$

Hence, we complete the proof for equation (D.6e). □

Proof. Lemma D.0.5 (D.6f):

$$\begin{aligned}
c\ell V(nC_\alpha^{(n)} - nC_\beta^{(n)}) &= \text{var}[-\hat{\phi}^{\dagger\prime}(P_\alpha^\dagger - P_\beta^\dagger)\hat{\phi}^\dagger/2 + \text{tr}[(J_\alpha H_\alpha^{-1}) - (J_\beta H_\beta^{-1})]] \\
&= \text{var}[-(W^{*\frac{1}{2}}\hat{\eta}^\dagger)'(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}}\hat{\eta}^\dagger/2] \\
&= \text{var}[(\hat{\eta}^{\dagger\prime}W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}}\hat{\eta}^\dagger]/4 \\
&= [4(W^{*-\frac{1}{2}}\phi^\dagger)'(W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}})^2(W^{*-\frac{1}{2}}\phi^\dagger) \\
&\quad +2\text{tr}(W^{*\frac{1}{2}}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}})^2]/4 \\
&= [\phi^{\dagger\prime}W^{*-\frac{1}{2}\prime}W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)'W^{*\frac{1}{2}}W^{*\frac{1}{2}\prime}(P_\alpha^\dagger - P_\beta^\dagger)W^{*\frac{1}{2}}W^{*-\frac{1}{2}}\phi^\dagger \\
&\quad +\text{tr}((P_\alpha^\dagger - P_\beta^\dagger)W^*)^2/2] \\
&= \phi^{\dagger\prime}(P_\alpha^\dagger - P_\beta^\dagger)'W^*(P_\alpha^\dagger - P_\beta^\dagger)\phi^\dagger + \text{tr}[(P_\alpha^\dagger - P_\beta^\dagger)W^*]^2/2
\end{aligned}$$

We complete proof for Equation (D.6f). □

Lemma D.0.6. *For $\alpha, \beta \in M$,*

$$c\ell E[nD_{KL}(\phi^{(n)}, \hat{\phi}_\alpha^{(n)})] = \|\phi^\dagger - \phi_\alpha^\dagger\|^2/2 + \text{tr}(J_\alpha^* H_\alpha^{*-1})/2, \tag{D.7a}$$

$$c\ell E[nD_{KL}(\phi^{(n)}, \hat{\phi}_\alpha^{(n)}) - nD_{KL}(\phi^{(n)}, \hat{\phi}_\beta^{(n)})] \tag{D.7b}$$

$$= -\phi^{\dagger\prime}(P_\alpha^\dagger - P_\beta^\dagger)\phi^\dagger/2 + \text{tr}\left[(P_\alpha^\dagger - P_\beta^\dagger)W^*\right]/2.$$

We verify that equation (D.7b) equals equation (D.6e)

Proof. Lemma D.0.6 (D.7a):

From equation (D.5a), we know $\text{p}\lim_{n\to\infty} nD_{KL}(\phi_1^{(n)}, \phi_1^{(n)}) = \|\phi_1^\dagger - \phi_2^\dagger\|^2/2$. We also know

$(\phi^{\dagger} - \phi^{\dagger}_{\alpha})'(\phi^{\dagger}_{\alpha} - \widehat{\phi^{\dagger}_{\alpha}}) = (\phi^{\dagger} - \phi^{\dagger}_{\alpha})' B^{\dagger}_{\alpha}(\theta^{\diamond}_{\alpha} - \hat{\theta}^{\diamond}_{\alpha}) = 0$. Therefore,

$$\begin{aligned} \mathrm{p}\lim_{n\to\infty} nD(\phi^{(n)}, \hat{\phi}^{(n)}_{\alpha}) &= ||\phi^{\dagger} - \widehat{\phi^{\dagger}_{\alpha}}||^2/2 \\ &= ||\phi^{\dagger} - \phi^{\dagger}_{\alpha}||^2/2 + ||\phi^{\dagger}_{\alpha} - \widehat{\phi^{\dagger}_{\alpha}}||^2/2. \end{aligned}$$

Noting $||\hat{\phi}^{\dagger}_{\alpha} - \phi^{\dagger}_{\alpha}||^2$ has weighted χ^2 distribution whose expectation value is $\mathrm{tr}(J^{*}_{\alpha} H^{*-1}_{\alpha})$. This was proved by Varin *et al.* (2011). Hence, equation (D.7a) is proved. □

Proof. Lemma D.0.6 (D.7b):

$$\begin{aligned} &c\ell E[nD_{KL}(\phi^{(n)}, \hat{\phi}^{(n)}_{\alpha}) - nD_{KL}(\phi^{(n)}, \hat{\phi}^{(n)}_{\beta})] \\ &= ||\phi^{\dagger} - \phi^{\dagger}_{\alpha}||^2/2 + \mathrm{tr}(J^{*}_{\alpha} H^{*1}_{\alpha})/2 - ||\phi^{\dagger} - \phi^{\dagger}_{\beta}||^2/2 - \mathrm{tr}(J^{*}_{\beta} H^{*-1}_{\beta})/2 \\ &= [||\phi^{\dagger} - \phi^{\dagger}_{\alpha}||^2 - ||\phi^{\dagger} - \phi^{\dagger}_{\beta}||^2]/2 - [\mathrm{tr}(J^{*}_{\alpha} H^{*-1}_{\alpha}) - \mathrm{tr}(J^{*}_{\beta} H^{*-1}_{\beta})]/2, \end{aligned}$$

where

$$\begin{aligned}
&||\phi^\dagger - \phi^\dagger_\alpha||^2 - ||\phi^\dagger - \phi^\dagger_\beta||^2 \\
&= (\phi^\dagger - \phi^\dagger_\alpha)'(\phi^\dagger - \phi^\dagger_\alpha) - (\phi^\dagger - \phi^\dagger_\beta)'(\phi^\dagger - \phi^\dagger_\beta) \\
&= \phi^{\dagger\prime}\phi^\dagger - \phi^{\dagger\prime}\phi^\dagger_\alpha - \phi^{\dagger\prime}_\alpha\phi^\dagger + \phi^{\dagger\prime}_\alpha\phi^\dagger_\alpha - \phi^{\dagger\prime}\phi^\dagger + \phi^{\dagger\prime}\phi^\dagger_\beta + \phi^{\dagger\prime}_\beta\phi^\dagger - \phi^{\dagger\prime}_\beta\phi^\dagger_\beta \\
&(\because \phi^\dagger_\alpha = P^\dagger_\alpha\phi^\dagger) \\
&= -\phi^{\dagger\prime}P^\dagger_\alpha\phi^\dagger - (P^\dagger_\alpha\phi^\dagger)'\phi^\dagger + (P^\dagger_\alpha\phi^\dagger)'(P^\dagger_\alpha\phi^\dagger) + \phi^{\dagger\prime}P^\dagger_\beta\phi^\dagger + (P^\dagger_\beta\phi^\dagger)'\phi^\dagger - (P^\dagger_\beta\phi^\dagger)'(P^\dagger_\beta\phi^\dagger) \\
&= -\phi^{\dagger\prime}P^\dagger_\alpha\phi^\dagger - \phi^{\dagger\prime}P^{\dagger\prime}_\alpha\phi^\dagger + \phi^{\dagger\prime}P^{\dagger\prime}_\alpha P^\dagger_\alpha\phi^\dagger + \phi^{\dagger\prime}P^\dagger_\beta\phi^\dagger + \phi^{\dagger\prime}P^{\dagger\prime}_\beta\phi^\dagger - \phi^{\dagger\prime}P^{\dagger\prime}_\beta P^\dagger_\beta\phi^\dagger \\
&(\because P^{\dagger 2}_\alpha = P^\dagger_\alpha) \\
&= -\phi^{\dagger\prime}(P^{\dagger\prime}_\alpha - P^{\dagger\prime}_\beta)\phi^\dagger = -\phi^{\dagger\prime}(P^\dagger_\alpha - P^\dagger_\beta)'\phi^\dagger
\end{aligned}$$

We get

$$\begin{aligned}
&c\ell E[nD_{K-L}(\phi^{(n)}, \hat{\phi}^{(n)}_\alpha) - nD_{KL}(\phi^{(n)}, \hat{\phi}^{(n)}_\beta)] \\
&= -\phi^{\dagger\prime}(P^\dagger_\alpha - P^\dagger_\beta)'\phi^\dagger/2 + \left[\mathrm{tr}(J^*_\alpha H^{*-1}_\alpha) - \mathrm{tr}(J^*_\beta H^{*-1}_\beta)\right]/2 \\
&= -\phi^{\dagger\prime}(P^\dagger_\alpha - P^\dagger_\beta)'\phi^\dagger/2 + \mathrm{tr}\left[(P^\dagger_\alpha - P^\dagger_\beta)W^*\right]/2
\end{aligned}$$

Equation (D.7b) is completely proved. □

Lemma D.0.7. *The two terms in equation* (4.3.2) *are asymptotically*

$$\mathrm{p}\lim_{n\to\infty} nV_{\alpha\beta} = \hat{\phi}^{\dagger\prime}(P^\dagger_\alpha - P^\dagger_\beta)W^*(P^\dagger_\alpha - P^\dagger_\beta)\hat{\phi}^\dagger, \tag{D.8a}$$

$$\mathrm{p}\lim_{n\to\infty} \nu_{\alpha\beta} = \mathrm{tr}[(P^\dagger_\alpha - P^\dagger_\beta)W^*]^2/2. \tag{D.8b}$$

Proof. Lemma D.0.7 (D.8a):

Expand $\sqrt{n}\left(c\ell(\hat{\phi}_\alpha; Y^{(i)}) - c\ell(\hat{\phi}_\beta; Y^{(i)})\right)$ with respect to $\hat{\phi}_\alpha$ and $\hat{\phi}_\beta$ around $\hat{\phi}$. We expand $c\ell(\hat{\phi}_\alpha; Y^{(i)})$ and $c\ell(\hat{\phi}_\beta; Y^{(i)})$ two terms respectively as the follow,

$$c\ell(\hat{\phi}_\alpha) = c\ell(\hat{\phi}) + \frac{\partial c\ell(\hat{\phi})}{\partial\hat{\phi}}(\hat{\phi}_\alpha - \hat{\phi}) + o_p(1),$$

and

$$c\ell(\hat{\phi}_\beta) = c\ell(\hat{\phi}) + \frac{\partial c\ell(\hat{\phi})}{\partial\hat{\phi}}(\hat{\phi}_\beta - \hat{\phi}) + o_p(1).$$

Subtract above two equations we obtain

$$\sqrt{n}\left(c\ell(\hat{\phi}_\alpha; Y^{(i)}) - c\ell(\hat{\phi}_\beta; Y^{(i)})\right) = \sqrt{n}\left(\frac{\partial c\ell(\hat{\phi}; Y^{(i)})}{\partial\hat{\phi}}(\hat{\phi} - \hat{\phi}_\beta)\right).$$

Furthermore, we are able to get

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\left[\sqrt{n}\left(c\ell(\hat{\phi}_\alpha; Y^{(i)}) - c\ell(\hat{\phi}_\beta; Y^{(i)})\right)\right]^2 \\
&= n^{-1}\sum_{i=1}^{n}\sqrt{n}(\hat{\phi}_\alpha - \hat{\phi}_\beta)'\frac{\partial c\ell(\hat{\phi}; Y^{(i)})}{\partial\hat{\phi}}\left(\frac{\partial c\ell(\hat{\phi}; Y^{(i)})}{\partial\hat{\phi}}\right)'\sqrt{n}(\hat{\phi}_\alpha - \hat{\phi}_\beta).
\end{aligned}$$

Note $\frac{1}{n}(\hat{c\ell}_\alpha^{(n)} - \hat{c\ell}_\beta^{(n)})^2 = o_p(1)$, and $n^{-1}\sum_{n=1}^{N}\frac{\partial c\ell(\hat{\phi};Y^{(i)})}{\partial\hat{\phi}}\left(\frac{\partial c\ell(\hat{\phi};Y^{(i)})}{\partial\hat{\phi}}\right)' = J^*$. So, we have

$$\begin{aligned}
&\operatorname*{p\,lim}_{n\to\infty}\left[\sum_{i=1}^{n}\left(c\ell(\hat{\phi}_\alpha; Y^{(i)}) - c\ell(\hat{\phi}_\beta; Y^{(i)})\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}c\ell(\hat{\phi}_\alpha; Y^{(i)}) - \sum_{i=1}^{n}c\ell(\hat{\phi}_\beta; Y^{(i)})\right)^2\right] \\
&= \operatorname*{p\,lim}_{n\to\infty}\sum_{i=1}^{n}\sqrt{n}(\hat{\phi}_\alpha - \hat{\phi}_\beta)' n^{-1}\frac{\partial c\ell(\hat{\phi})}{\partial\hat{\phi}}\left(\frac{\partial c\ell(\hat{\phi})}{\partial\hat{\phi}}\right)'\sqrt{n}(\hat{\phi}_\alpha - \hat{\phi}_\beta) \\
&= (\hat{\phi}_\alpha^\diamond - \hat{\phi}_\beta^\diamond)' J^* (\hat{\phi}_\alpha^\diamond - \hat{\phi}_\beta^\diamond).
\end{aligned}$$

We are able to prove that

$$\begin{aligned}
&\mathrm{p}\lim_{n\to\infty} nV_{\alpha\beta} \\
&= \mathrm{p}\lim_{n\to\infty} n\left[n^{-1}\sum_{i=1}^{n}\left(c\ell(\hat{\phi}_\alpha; Y^{(i)}) - c\ell(\hat{\phi}_\beta; Y^{(i)})\right)^2 - \left(\hat{c\ell}_\alpha^{(n)}/n - \hat{c\ell}_\beta^{(n)}/n\right)^2\right] \\
&= (\hat{\phi}_\alpha^\diamond - \hat{\phi}_\beta^\diamond)' J^* (\hat{\phi}_\alpha^\diamond - \hat{\phi}_\beta^\diamond) \\
&= (H^{*-\frac{1}{2}}\hat{\phi}_\alpha^\dagger - H^{*-\frac{1}{2}}\hat{\phi}_\beta^\dagger)' J^* (H^{*-\frac{1}{2}}\hat{\phi}_\alpha^\dagger - H^{*-\frac{1}{2}}\hat{\phi}_\beta^\dagger) \\
&= (\hat{\phi}_\alpha^\dagger - \hat{\phi}_\beta^\dagger)' H^{*-1/2} J^* H^{*-\frac{1}{2}} (\hat{\phi}_\alpha^\dagger - \hat{\phi}_\beta^\dagger) \\
&= (P_\alpha^\dagger \hat{\phi}^\dagger - P_\beta^\dagger \hat{\phi}^\dagger)' W^* (P_\alpha^\dagger \hat{\phi}^\dagger - P_\beta^\dagger \hat{\phi}^\dagger) \\
&= \hat{\phi}^{\dagger\prime} (P_\alpha^\dagger - P_\beta^\dagger)' W^* (P_\alpha^\dagger - P_\beta^\dagger) \hat{\phi}^\dagger
\end{aligned}$$

We proved equation (D.8a). □

Proof. Lemma D.0.7 (D.8b):

To prove (D.8b) is equivalent to prove

$$\begin{aligned}
&\mathrm{tr}[(P_\alpha^\dagger - P_\beta^\dagger)W^*]^2/2 \\
&= \mathrm{p}\lim_{n\to\infty} \mathrm{tr}\left(H_{\alpha\alpha}^{(n)-1} J_{\alpha\alpha}^{(n)} H_{\alpha\alpha}^{(n)-1} J_{\alpha\alpha}^{(n)} + H_{\beta\beta}^{(n)-1} J_{\beta\beta}^{(n)} H_{\beta\beta}^{(n)-1} J_{\beta\beta}^{(n)}\right)/2 \\
&- tr(H_{\alpha\alpha}^{(n)-1} J_{\alpha\beta}^{(n)} H_{\beta\beta}^{(n)-1} J_{\beta\alpha}^{(n)})
\end{aligned}$$

Let's look the left side of the equation first.

$$\mathrm{tr}[(P_\alpha^\dagger - P_\beta^\dagger)W^*]^2/2 \quad = \mathrm{tr}\left[P_\alpha^\dagger W^* P_\alpha^\dagger W^* - 2P_\alpha^\dagger W^* P_\beta^\dagger W^* + P_\beta^\dagger W^* P_\beta^\dagger W^*\right]/2 \qquad \text{(D.9)}$$

We now calculate each element in the above equation. Note $J^*_\alpha = B^{*\prime}_\alpha J^* B^*_\alpha$ and $H^*_\alpha = B^{*\prime}_\alpha H^* B^*_\alpha$, which are proved when we prove equation (D.6d). By similar approach, we can prove $J^*_\beta = B^{*\prime}_\beta J^* B^*_\beta$ and $H^*_\beta = B^{*\prime}_\beta H^* B^*_\beta$. In addition, it's easy to verify that

$$\begin{aligned} J^*_{\alpha\beta} &= E\left[\frac{\partial c\ell(\phi_\alpha(\theta^*_\alpha), Y)}{\partial\phi_\alpha(\theta^*_\alpha)}\left(\frac{\partial c\ell(\phi_\beta(\theta^*_\beta), Y)}{\partial\phi_\beta(\theta^*_\beta)}\right)'\right] \\ &= \left(\frac{\partial\phi_\alpha}{\partial\theta_\alpha}\right)' E\left[\frac{\partial c\ell(\phi^*, Y)}{\partial\phi^*}\left(\frac{\partial c\ell(\phi^*, Y)}{\partial\phi^*}\right)'\right]\frac{\partial\phi_\beta}{\partial\theta_\beta} \\ &= B^{*\prime}_\alpha J^* B^*_\beta, \end{aligned}$$

By applying same approach, we can show $J^*_{\beta\alpha} = B^{*\prime}_\beta J^* B^*_\alpha$. We are now able to calculate each element in equation (D.9)

$$\begin{aligned} &\mathrm{tr}(P^\dagger_\alpha W^* P^\dagger_\alpha W^*) \\ &= \mathrm{tr}[\Big(B^\dagger_\alpha(B^{\dagger\prime}_\alpha B^\dagger_\alpha)^{-1}B^{\dagger\prime}_\alpha H^{*-\frac{1}{2}}J^*H^{*-\frac{1}{2}}B^\dagger_\alpha(B^{\dagger\prime}_\alpha B^\dagger_\alpha)^{-1}B^{\dagger\prime}_\alpha H^{*-\frac{1}{2}}J^*H^{*-\frac{1}{2}}\Big) \\ &= \mathrm{tr}(H^{*\frac{1}{2}}B^*_\alpha)\left((H^{*\frac{1}{2}}B^*_\alpha)'(H^{*\frac{1}{2}}B^*_\alpha)\right)^{-1}(H^{*\frac{1}{2}}B^*_\alpha)'(H^{*-\frac{1}{2}}J^*H^{*-\frac{1}{2}}) \\ &(H^{*\frac{1}{2}}B^*_\alpha)\left((H^{*\frac{1}{2}}B^*_\alpha)'(H^{*\frac{1}{2}}B^*_\alpha)\right)^{-1}(H^{*-\frac{1}{2}}B^*_\alpha)'(H^{*-\frac{1}{2}}J^*H^{*-\frac{1}{2}})] \\ &= \mathrm{tr}\left[(B^{*\prime}_\alpha H^* B^*_\alpha)^{-1}(B^{*\prime}_\alpha J^* B^*_\alpha)(B^{*\prime}_\alpha H^* B^*_\alpha)^{-1}(B^{*\prime}_\alpha J^* B^*_\alpha)\right] \\ &= \mathrm{tr}(H^{*-1}_\alpha J^*_\alpha H^{*-1}_\alpha J^*_\alpha) \\ &= \mathrm{p}\lim_{n\to\infty}\mathrm{tr}\left(H^{(n)-1}_{\alpha\alpha}J^{(n)}_{\alpha\alpha}H^{(n)-1}_{\alpha\alpha}J^{(n)}_{\alpha\alpha}\right). \end{aligned}$$

Similarly, we can show

$$\mathrm{tr}(P^\dagger_\beta W^* P^\dagger_\beta W^*) = \mathrm{tr}(H^{*-1}_\beta J^*_\beta H^{*-1}_\beta J^*_\beta) = \mathrm{p}\lim_{n\to\infty}\mathrm{tr}\left(H^{(n)-1}_{\beta\beta}J^{(n)}_{\beta\beta}H^{(n)-1}_{\beta\beta}J^{(n)}_{\beta\beta}\right).$$

We now calculate for $\mathrm{tr}(P^\dagger_\alpha W^* P^\dagger_\beta W^*)$

$$
\begin{aligned}
&\mathrm{tr}(P^\dagger_\alpha W^* P^\dagger_\beta W^*)\\
&= \mathrm{tr}[\left(B^\dagger_\alpha (B^{\dagger\prime}_\alpha B^\dagger_\alpha)^{-1} B^{\dagger\prime}_\alpha H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}} B^\dagger_\beta (B^{\dagger\prime}_\beta B^\dagger_\beta)^{-1} B^{\dagger\prime}_\beta H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}}\right)\\
&= \mathrm{tr}(H^{*\frac{1}{2}} B^*_\alpha) \left((H^{*\frac{1}{2}} B^*_\alpha)'(H^{*\frac{1}{2}} B^*_\alpha)\right)^{-1} (H^{*\frac{1}{2}} B^*_\alpha)'(H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}})\\
&(H^{*\frac{1}{2}} B^*_\beta) \left((H^{*\frac{1}{2}} B^*_\beta)'(H^{*\frac{1}{2}} B^*_\beta)\right)^{-1} (H^{*-\frac{1}{2}} B^*_\beta)'(H^{*-\frac{1}{2}} J^* H^{*-\frac{1}{2}})]\\
&= \mathrm{tr}\left[(B^{*\prime}_\alpha H^* B^*_\alpha)^{-1}(B^{*\prime}_\alpha J^* B^*_\beta)(B^{*\prime}_\beta H^* B^*_\beta)^{-1}(B^{*\prime}_\beta J^* B^*_\alpha)\right]\\
&= \mathrm{tr}(H^{*-1}_\alpha J^*_{\alpha\beta} H^{*-1}_\beta J^*_{\beta\alpha})\\
&= \mathrm{p}\lim_{n\to\infty} \mathrm{tr}\left(H^{(n)-1}_{\alpha\alpha} J^{(n)}_{\alpha\beta} H^{(n)-1}_{\beta\beta} J^{(n)}_{\beta\alpha}\right).
\end{aligned}
$$

Plug above results into equation (D.9), we show

$$
\begin{aligned}
&\mathrm{tr}[(P^\dagger_\alpha - P^\dagger_\beta)W^*]^2/2\\
&= \left[\mathrm{p}\lim_{n\to\infty} \mathrm{tr}\left(H^{(n)-1}_{\alpha\alpha} J^{(n)}_{\alpha\alpha} H^{(n)-1}_{\alpha\alpha} J^{(n)}_{\alpha\alpha}\right) + \mathrm{p}\lim_{n\to\infty} \mathrm{tr}\left(H^{(n)-1}_{\beta\beta} J^{(n)}_{\beta\beta} H^{(n)-1}_{\beta\beta} J^{(n)}_{\beta\beta}\right)\right]/2\\
&- \mathrm{p}\lim_{n\to\infty} \mathrm{tr}\left(H^{(n)-1}_{\alpha\alpha} J^{(n)}_{\alpha\beta} H^{(n)-1}_{\beta\beta} J^{(n)}_{\beta\alpha}\right),
\end{aligned}
$$

i.e.

$$
\mathrm{p}\lim_{n\to\infty} \nu^{(n)}_{\alpha\beta} = \mathrm{tr}[(P^\dagger_\alpha - P^\dagger_\beta)W^*]^2/2.
$$

We complete proof for equation(D.8b). □

Note that the sum of equation (D.8a) and (D.8b), if $\hat{\phi}^\dagger$ is replaced with $\phi^\dagger$, gives equation (D.6f). From the Lemma D.0.1 to Lemma D.0.7, consequently, Theorem 4.3.2 can be derived.

Proof. : Equation (4.8)

Note $H^* = J^*$, $\text{tr}(I_{m_\alpha}) = m_\alpha$, and $\text{tr}(I_{m_\beta}) = m_\beta$. Let I denote identical matrix. Hence, equation (4.5) can be simplified as:

$$\begin{aligned}
\nu_{\alpha\beta}^{(n)} &= \text{tr}(H_{\alpha\alpha}^{(n)-1} H_{\alpha\alpha}^{(n)} H_{\alpha\alpha}^{(n)-1} H_{\alpha\alpha}^{(n)} + H_{\beta\beta}^{(n)-1} H_{\beta\beta}^{(n)} H_{\beta\beta}^{(n)-1} H_{\beta\beta}^{(n)})/2 \\
&\quad -\text{tr}(H_{\alpha\alpha}^{(n)-1} H_{\alpha\beta}^{(n)} H_{\beta\beta}^{(n)-1} H_{\beta\alpha}^{(n)}) \\
&= \text{tr}(I_{m_\alpha} + I_{m_\beta})/2 - \text{tr}(H_{\alpha\alpha}^{(n)-1} H_{\alpha\beta}^{(n)} H_{\beta\beta}^{(n)-1} H_{\beta\alpha}^{(n)}) \\
&= \text{tr}(m_\alpha + m_\beta)/2 - \text{tr}(H_{\alpha\alpha}^{(n)-1} H_{\alpha\beta}^{(n)} H_{\beta\beta}^{(n)-1} H_{\beta\alpha}^{(n)}).
\end{aligned}$$

We proved equation (4.8). □