Composite likelihood: Multiple comparisons and non-standard conditions in hypothesis testing

MAHDIS AZADBAKHSH

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

Graduate Program in Mathematics and Statistics

York University

Toronto, Ontario

September 2017

Abstract

Computational intensity in using full likelihood estimation of multivariate and correlated data is a valid motivation to employ composite likelihood as an alternative that eases the process by using marginal or conditional densities and reducing the dimension .

We study the problem of multiple hypothesis testing for multidimensional clustered data. The problem of multiple comparisons is common in many applications. We propose to construct multiple comparisons procedures based on composite likelihood statistics. The simultaneous multivariate normal quantile is chosen as the threshold that controls the multiplicity. We focus on data arising in four cases: multivariate Gaussian, probit, quadratic exponential models and gamma. To assess the quality of our proposed methods, we assess their empirical performance via Monte Carlo simulations. It is shown that composite likelihood based procedures maintain good control of the familywise type I error rate in the presence of intra-cluster correlation, whereas ignoring the correlation leads to invalid performance. Using data arising from a depression study and also kidney study, we show how our composite likelihood approach makes an otherwise intractable analysis possible.

Moreover, we study distribution of composite likelihood ratio test when the true parameter is not an interior point of the parameter space. We approached the problem looking at the geometry of the parameter space and approximating it at the true parameter by a cone under Chernoff's regularity. First, we established the asymptotic properties of the test statistic for testing continuous differentiable linear and non-linear combinations of parameters and then we provide algorithms to compute the distribution of both full and composite likelihood ratio tests for different cases and dimensions. The proposed approach is evaluated by running simulations.

This work is dedicated to my parents.

Having them is one of the greatest blessings in my life.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Dr. Xin Gao and Dr. Hanna Jankowski for all their invaluable support and guidance along this journey. I would also like to thank my committee members, Dr. Cindy Fu and Dr. Steven Wang for their collaboration, suggestions and insightful comments.

I am also thankful to many staff members, faculty and fellow students from the York University community for supporting me in various ways throughout the completion of this degree.

Contents

List of Tables

List of Figures

List of Symbols

Symbol	**Description**
Θ	total parameter space
Θ_0	null parameter space
θ_0	true value of the parameter
$CL(\theta; y)$	composite likelihood function
$cl(\theta; y)$	log-composite likelihood function
$G(\theta)$	Godambe information matrix
$H(\theta)$	Hessian matrix
U	$H(\theta) = U^T U$
$\widehat{\theta}_n^c$	composite likelihood estimator
$\mathcal{C}$	cone
$\mathbb{P}$	polyhedral cone
$\mathbb{P}_0$	null polyhedral cone under
$T_\theta(\theta)$	tangent cone
$\tilde{z}$	Uz transferred z also $\tilde{T}_\theta(\theta), \tilde{\theta}, \tilde{\mathbb{P}}$
$\mathbb{I}$	indicator function
$\mathcal{I}$	subset of $\{1, \ldots, q\}$ such that $\mathcal{A}_\mathcal{I}\theta = 0$
I_k	k-d identity matrix
$ri(F_\mathcal{I})$	relative interior set
$\mathbb{J}(\theta)$	Jacobian matrix
$\mathcal{A}(\theta_0)$	Jacobian matrix evaluated at θ_0
$\Pi(y\|\mathbb{P})$	projection of y onto $\mathbb{P}$
w_{ij}	weight of the distribution of $\bar{\chi}^2$, $P(\bar{\chi}^2 \leq c) = \sum_{i,j} P(S_{ij} \leq c) w_{ij}$

Chapter 1

Composite Likelihood Estimation

1.1 Introduction

Composite likelihood methods are extensions of the likelihood method that project high-dimensional likelihood functions to low-dimensional ones that results in less complex computations [7, 30]. This dimension reduction is achieved by compounding valid marginal or conditional densities instead of using the joint density. It has been shown that, under regularity conditions, the composite likelihood estimator has desirable properties, such as consistency and asymptotic normality [7, 30, 49, 50]. This makes it an appealing alternative in inferential procedures. Xu and Reid [54] also discussed efficiency and robustness of the composite likelihood method. They concluded that composite likelihood method is often more reliable than full likelihood as the high dimensional joint density is more likely to be mis-specified than lower dimensional densities. Furthermore, composite likelihood is often more computationally convenient than full likelihood at a cost of some mild loss of efficiency. The magnitude of this loss depends on the dimension of the multivariate vector and its dependency structure.

One situation that motivates us to use composite likelihood estimation is when applying maximum likelihood on multivariate distributions encounter computational challenges, especially when sub-groups of data are correlated. For example, evaluating the full likelihood of a multivariate probit model involves multi-dimensional integration, which quickly becomes computationally prohibitive. Composite likelihood reduce this computational burden by using marginal densities instead of the joined density. Ignoring correlations among the subjects in order to lower the complexity in maximum likelihood approach can lead to invalid inferences.

Moreover, composite likelihood can help us to make the models simpler and the computation less complex. For the quadratic exponential model, the normalizing constant has to be computed through summation of all possible configurations of the clustered data and computational intensity increases with the cluster size. By using conditional density in composite likelihood the normalizing constant will be removed.

Composite likelihood is applicable in any situation that maximum likelihood fits and makes the computation easier or possible.

1.1.1 Previous work

Composite likelihood methods was suggested by Lindsay [30]. It has been shown that under regularity conditions, the composite likelihood estimator has desirable properties, such as consistency and asymptotic normality (Cox and Reid [7], Lindsay [30], Varin [49], Varin et al. [50]). Although usually there is a little loss of efficiency compared to maximum likelihood method, composite likelihood is robust in the sense that the inference is still valid in the case of misspecification of the statistical models or parameters of the densities (Xu and Reid [54]) and also more reliable than full likelihood since in modelling the high dimensional joint density, misspecification is more likely to happen with lower dimensional densities. The magnitude of loss of efficiency depends on the dimension of the multivariate vector and its dependency structure. In some cases even the composite likelihood is fully efficient. Let $\widehat{\theta}_n^c$ denote the maximum composite likelihood estimator (MCLE). Xu and Reid [54] give precise conditions under which $\widehat{\theta}_n^c$ is consistent for θ.

Composite likelihood methodology has been applied to numerous statistical problems: Zhao and Joe [55] proposed composite likelihood methods for multivariate data analysis. Renard et al. [36] used it in the generalized linear mixed model; Fearnhead and Donnelly [11] proposed to maximize the compounded marginal probabilities in genetics; Geys et al. [16] presented a composite likelihood method for clustered binary data in the quadratic exponential model. Composite likelihood method has also been successfully applied in other areas including spatial statistics (Heagerty and Lele [19], Hjort and Omre [20], Varin and Vidoni [51]), Markov random fields (Besag [3]), and multivariate survival analysis (Li and Lin [29], Parner [34]). However, the potential of composite likelihood in multiple testing has yet to be explored.

1.1.2 Outline

In this thesis, we focus on studying composite likelihood in two areas.

- Chapter 1 covers the concept of composite likelihood estimation and its asymptotic behaviour and shows the consistency and asymptotic normality of the estimator in context of clustered data.

- In chapters $2-3$, we employ composite likelihood estimation as an alternative to full likelihood and a multivariate normal quantile as the threshold to perform multiple hypothesis testing in multidimensional and correlated data. We examine the proposed approach via simulation and then apply it on some real data sets.

- In chapters $4-5$, we are concerned with the cases that some regularity assumptions in composite likelihood ratio testing does not hold. we develop a theory to find the limiting distribution of composite likelihood ratio test in the situations that the true parameter has non-standard conditions. Then we derive the limiting distribution and propose some algorithms to compute the test statistic in high dimensions for testing a continuously differentiable function of the parameters. Then the approach is validated through simulations.

- In chapter 6, two areas that the suggested approach in likelihood ratio testing can be applied , are described as future works.

Each study is described in detail in the next chapters.

1.2 Composite likelihood approach

Let $y_{n\times m} = (y_1, \ldots, y_n)$ denote a sample of size n from a joint $m-$variate density function $f(y;\theta)$ where $\theta \in \Theta \subset \mathbb{R}^p$.

For $A_k \subset \{(i,j) : j = 1, \ldots, m, i = 1, \ldots, n\}$, let $y_{A_k} = \{y_{ij}, (i,j) \in A_k\}$ denote a subset of the data, where $k = 1, \ldots, K$. The composite likelihood function is then defined as

$$CL(\theta; y) \;\; = \;\; \prod_{k=1}^{K} f(y_{A_k};\theta)^{w_{A_k}},$$

where $f(y_{A_k};\theta)$ is the density for the subset vector y_{A_k}, and w_{A_k} are some suitably chosen weights. The composite likelihood function can be constructed in two main ways.

Composite marginal likelihood function is built based on lower dimensional marginal densities. For example, the univariate marginal composite likelihood function is

$$CL(\theta) = \prod_{i=1}^{n}\prod_{j=1}^{m} f(y_{ij};\theta)$$

where any dependence structure is ignored. The second class of composite likelihood functions is constructed by univariate conditional likelihood functions

$$CL(\theta) = \prod_{i,j}(y_{ij}|y_{i(-j)};\theta)$$

where $y_{i(-j)}$ denotes the sub-vector of y_i with its jth element removed.

The composite log likelihood function is denoted by

$$cl(\theta) = \log CL(\theta)$$

and the maximum composite likelihood estimate (MCLE) is defined as $\widehat{\theta}_n^c = \text{argmax}_{\theta \in \Theta} cl(\theta)$.

In general, composite likelihood is a compounded form of marginal or conditional likelihoods, which is often easier to maximize than full likelihood. In practice, the type of composite likelihood should be chosen so that the resulting composite score equation is consistent for the parameters, and the computation complexity is sufficiently manageable.

Xu and Reid [54] give precise conditions under which $\widehat{\theta}_n^c$ is consistent for θ. Under appropriate assumptions, $\sqrt{n}(\widehat{\theta}_n^c - \theta)$ is also asymptotically normally distributed with mean zero and limiting variance given by the inverse of the the Godambe information matrix [30, 51], where

$$G^{-1}(\theta) \quad = \quad H^{-1}(\theta) J(\theta) H^{-1}(\theta), \tag{1.1}$$

with $H(\theta) = \lim_n E(-cl^{(2)}(\theta; y))/n$ and $J(\theta) = \lim_n \text{var}(cl^{(1)}(\theta; y))/n$. Here, $cl^{(1)}$ is the vector of first derivatives and $cl^{(2)}$ is the matrix of second order derivatives of $cl(\theta; y) = \log CL(\theta; y)$ with respect to θ. The matrix $H(\theta)$ can be estimated as the negative Hessian matrix evaluated at the maximum composite likelihood estimator, whereas the matrix $J(\theta)$ can be estimated as the sample covariance matrix of the composite score vectors. Both estimators, which we denote as $\widehat{H}_n$ and $\widehat{J}_n$, are consistent [51].

1.3 Asymptotic properties of composite likelihood estimator in clustered data

Xu and Reid (2011) provided a detailed proof of consistency under misspecification, along with a precise list of required conditions. One can obtain from their work sufficient conditions for consistency even in the well-specified setting. Here, for reference, we give a proof of some asymptotic properties of the composite likelihood estimator provided that the model is correctly specified and data is formed by n independent clusters, each with fixed sample size m. For the composite marginal likelihoods, regularity conditions can be stated with slight modification of the conditions in full likelihood context.

Regularity conditions:

(A1). The marginal density function of y_{ij}, $f(y;\theta)$ is distinct for different values of y, i.e. if $\theta_1 \neq \theta_2$ then $P(f(y_{ij};\theta) \neq f(y_{ij};\theta)) > 0$, for all $j = 1, \ldots, m$.

(A2). The marginal densities of y_{ij} have common support for all θ.

(A3). The true value θ_0 is an interior point of Ω, the space of possible values of the parameter θ.

(A4). Let α and $\partial\alpha$ denote the index and partial derivative operator, respectively, as in the standard multi-index notation from multivariable calculus. The marginal density $\log f$ is three times continuously differentiable in a closed ball around θ_0. Moreover, there exists a constant c and an integrable function $M(y)$ such that

$$|(\partial\alpha\partial\theta_i \log f)(y;\theta)| \leq M(y),$$

for all $||\theta - \theta_0||_2 < c$, all $|\alpha| = 2$, and any $i = 1, \ldots, p$. Here, $\theta \in \mathbb{R}^P$ and $||\cdot||_2$ denotes the Euclidean norm.

(A5). $J(\theta_0)$ is well-defined (i.e. exists and is finite) and invertible.

(A6). $H(\theta_0)$ is well-defined (i.e. exists and is finite) and (strictly) positive-definite.

Define the marginal composite log-likelihood function as

$$cl(\theta) \quad = \quad \log CL(\theta; y) \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{m} \log f(y_{ij}; \theta),$$

and let $cl_m(\theta; y_i) = \sum_{j=1}^{m} \log f(y_{ij}; \theta))$.

Theorem 1.3.1. *Under the regularity conditions (A1)-(A6), there exists a solution to the composite likelihood equation, $\widehat{\theta}_n^c$, which satisfies*

$$\sqrt{n}(\widehat{\theta}_n^c - \theta_0) \quad \Rightarrow \quad G^{-1/2}(\theta_0)\, Z$$

where $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$, and Z is a standard normal random vector.

The proof is provided in the appendix A.

1.3.1 Estimating $H(\theta)$ and $J(\theta)$

$\widehat{H}_n$ and $\widehat{J}_n$ are estimators of $H(\theta)$ and $J(\theta)$, respectively.

To estimate $H(\theta)$ and $J(\theta)$, it is proposed in Cox and Reid [7] that

$$\widehat{H}_n \quad = \quad -\frac{1}{n}\sum_{i=1}^{n} cl^{(2)}(\theta; y_i)\bigg|_{\theta=\widehat{\theta}_n^c},$$

that $\widehat{H}$ is the negative Hessian matrix evaluated at the maximum composite likelihood estimator. To estimate the matrix J, we can use the sample covariance matrix of the composite score vectors:

$$\widehat{J}_n \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(cl^{(1)}(\theta; y_i))^T cl^{(1)}(\theta; y_i).$$

Both estimators $\widehat{H}_n$ and $\widehat{J}_n$ are consistent [51, page 523]. For more details on the estimation of H and J, we refer to Cox and Reid [7] and Varin [49].

Chapter 2

Multiple Comparisons Using Composite Likelihood in Clustered Data

2.1 Introduction

”Clustered” is referred to correlated data with a grouped structure that individual in each group/sub-population are relating in some manners. For example, repeated measurements in clinical studies, that each individual can be considered as a cluster, parent-sibling data, such as data from different stages of disease spreading, or data from same pedigree, spatial data and longitudinal studies.

This correlation structure within the clusters should be taken into account in the analysis, otherwise ignoring it leads to invalid inferences. As computing full likelihood could be challenging for correlated data, composite likelihood is introduced as a feasible alternative. Composite likelihood method has been successfully applied in many areas. As clustered data is correlated, full likelihood again might encounter computational difficulty. Here, we explore it in multiple testing problems on clustered data.

Multiplicity of hypothesis tests is an intrinsic issue arising when the number of simultaneous comparisons is greater than one, leading to a family-wise type I error rate larger than α. The greater the number of comparisons, the more serious this effect becomes. Different multiple testing procedures try to adjust this issue [4, 21]. The classical Bonferroni method is the simplest procedure to adjust the overall type I error rate, but it is very conservative. The Dunn-Sidák procedure [43] generalizes the Bonferroni procedure by using a slightly less conservative p-value threshold for each comparison. Scheffe [40] established a method for testing all possible linear comparisons among a set of normally distributed variables, which tends to be over-conservative for a finite family of multiple comparisons. There are some stage-wise procedures as well to improve the power. Simes [44] modified the Bonferroni procedure based on ordered p-values. Holm [22] proposed a multi-stage procedure that adjusts the family-wise error rate in each step using the number of remaining null hypotheses. Hommel [23] suggested a stagewise rejective multiple test based on the principle of closed test procedures. All of these methods are less conservative and therefore more powerful than the Bonferroni method.

However, it is difficult to construct simultaneous confidence intervals based on stagewise procedures. As another alternative, Hothorn, Bretz and Westfall [24] proposed to use quantiles of the multivariate normal and multivariate t-distribution to perform multiple comparisons in parametric methods. Therefore, the correlation structure is taken into account in this procedure and it offers more accurate control of the family-wise type I error rate. The approach has been employed in many parametric and nonparametric settings to provide both multiple inferences and simultaneous intervals [24, 26, 27].

In this chapter, we propose a new procedure to handle multiple testing with a more efficient threshold and with employing composite likelihood methodology. This enables us to overcome problems with computational intensity and multiplicity issue in multiple testing of clustered data. We explore in detail different multivariate models for correlated clustered data including the multivariate normal, multivariate probit, gamma and quadratic exponential models to illustrate our multiple comparisons approach. Moreover, we explore the Bonferroni, Scheffé, Dunn-Sidák, Holm, and the multivariate normal quantile (MNQ) of Hothorn et al. [24] methods with both univariate and conditional composite likelihood formulations. Among these methods, the multivariate normal quantile threshold appears to have the best control of the familywise type I error rate in most simulation settings.

The structure of this chapter is as follows: In Section 2.2, we develop our composite likelihood based test statistics for multiple inferences and establish their asymptotic properties. In Section 2.3, we provide details on how to apply the general approach on different multivariate models, including normal, probit, quadratic exponential and gamma. We continue with examining the proposed approach in the next chapter.

2.2 Multiple comparisons procedures based on composite likelihood

In parametric statistical model, suppose $y \sim \{f(y;\theta), \theta \in \Theta\}$, where $\theta = (\theta_1, \ldots, \theta_p)^T$, and $\theta \in \Theta \subset \mathbb{R}^p$. Let $y = (y_1^T, \cdots, y_n^T)$ denote the response variables, where $y_i = (y_{i1}, \cdots, y_{im_i})^T$ is the vector of observations from cluster i, $i = 1, \cdots, n$ from a study population. It is assumed that observations across different clusters are independent, whereas observations within the same cluster may be dependent. Note that overall sample size is $\sum_{i=1}^n m_i$. We assume that the cluster size, m_i, is uniformly bounded.

Let $C = C_{p\times c} = (C_{(1)}, C_{(2)}, \cdots, C_{(c)})$ denote the contrast matrix and the family of c linear combinations of the parameters can then be specified by $C\theta$.

Consider the testing of the family of hypotheses $\{H_{0l} : C_{(l)}^T\theta = 0, l = 1, \cdots, c\}$. In multiple testing, the family-wise type I error (FWER) rate is the probability of false rejection of at least one individual null hypothesis when all null hypotheses are true:

$$\text{P(rejecting at least one} H_{0i} \mid \bigcap_k H_{0k}) = \alpha.$$

Let $\widehat{\theta}_n^c$ be the maximum composite likelihood estimator. It is shown that $\sqrt{n}(\widehat{\theta}_n^c - \theta) \longrightarrow N_p(0, G^{-1}(\theta))$ where G denote the Godambe information matrix (1.1). Consider the hypothesis test on a family of linear combinations of the parameters: $\{H_0 : C^T\theta = 0\}$. Denote by $\Gamma = G^{-1}(\theta)$, and let $\widehat{\Gamma}_n$ denote the consistent estimator of Γ, where $\widehat{\Gamma}_n = \widehat{H}_n^{-1}\widehat{J}_n\widehat{H}_n^{-1}$. We propose the following test statistics for our hypothesis test

$$T_{l,n} \;=\; \frac{{C_{(l)}}^T \widehat{\theta}_n^c}{\sqrt{\left({C_{(l)}}^T \widehat{\Gamma}_n C_{(l)}\right)/n}}, \quad l = 1, \ldots, c. \tag{2.1}$$

The limiting distribution of $T_n = (T_{1,n}, \cdots, T_{l,n})^T$ is multivariate normal $MVN(0, V)$, where

$$V \;=\; \mathrm{diag}(D)^{-1/2} D \, \mathrm{diag}(D)^{-1/2}, \quad D \;=\; CG^{-1}(\theta)C^T. \tag{2.2}$$

Furthermore, since $V_{i,i} = 1$, the marginal asymptotic distribution of each individual $T_{l,n}$ is standard normal. In practice, we estimate V by plugging $\widehat{\Gamma}_n$ as a consistent estimator of $G^{-1}(\theta)$ into (2.2). This results in a consistent estimator of V.

The proposed test statistics are Wald-type statistics which are not invariant under re-parametrization. Under re-parametrization, the new statistics follow the same type of limiting distributions, but the values of the statistics are not the same. This is a standard limitation that Wald-type statistics encounter.

The multivariate distribution of T_n can be approximated by a multivariate t distribution. The denominator ${C_{(l)}}^T \widehat{\Gamma}_n C_{(l)}/n$ has an asymptotic equivalent distribution as ${C_{(l)}}^T H^{-1} \widehat{J}_n H^{-1} C_{(l)}/n$ based on Slutsky' Theorem. Furthermore, $\widehat{J}_n = (cl^{(1)})^T cl^{(1)}$ asymptotically follows a Wishart (J, n). This entails that asymptotically ${C_{(l)}}^T H^{-1} \widehat{J}_n H^{-1} C_{(l)}$ follows $\sigma_l^2 \chi_n^2$, where $\sigma_l^2 =$

${C_{(l)}}^T H^{-1} J H^{-1} C_{(l)}$. Reformulate $T_{l,n}$ as

$$\frac{{C_{(l)}}^T \widehat{\theta}_n^c / \sigma_l}{\sqrt{\left((C_{(l)})^T \widehat{\Gamma}_n C_{(l)} \right) / (n\sigma_l^2)}},$$

where the numerator is asymptotically a multivariate normal $MVN\ (0, V)$, and the denominator is asymptotically $\sqrt{\chi_n^2 / n}$. Therefore, the multivariate distribution of T_n can be approximated as a multivariate $t(V, n)$, where V is the covariance matrix and n is the degrees of freedom.

Different procedures have been suggested to control the FWER and adjust individual test levels. In this work we illustrate a few of these approaches.

- The Bonferroni procedure: The global intersection hypothesis $\cap_{l=1}^m H_{0l}$ will be rejected if $\max_l |T_{l,n}| > Z_{\alpha/m}$. Each individual hypothesis H_{0l} will be rejected if $|T_{l,n}| > Z_{\alpha/m}$.

- The Holm's procedure: For each H_{0l}, evaluate the p-value $p_l = 2P(Z > |T_{l,n}|)$. Order the p-values from the least to the greatest as $p_{(1)}, \ldots, p_{(m)}$ and the corresponding hypotheses are reordered as $H_{(01)}, \ldots, H_{(0m)}$. The global intersection hypothesis $\cap_{l=1}^m H_{0l}$ will be rejected if $p_{01} \leq \alpha/m$. Let k denote the smallest l so that $p_{(l)} > \alpha/m - l + 1$. If $k > 1$, then the individual hypotheses $H_{01}, \ldots, H_{0,k-1}$ will be rejected.

- The MNQ procedure: The global intersection hypothesis $\cap_{l=1}^m H_{0l}$ will be rejected if $\max_l |T_{l,n}| > Q_{\alpha,V}$, where $Q_{\alpha,V}$ denote the equi-coordinate α quantile for a multivariate normal vector with covariance matrix V. Each individual hypothesis H_{0l} will be rejected if $|T_{l,n}| > Q_{\alpha,V}$.

The MNQ approach is handled based on the p-dimensional approximation

$$P\Big(\max|T_{l,n} \leq t) \cong \int_{-t}^{t} \dots \int_{-t}^{t} \varphi_p(x_1, \dots, x_p, V, \nu) dx_1, \dots, dx_p$$

where φ is the limiting p-variate normal density ($\nu = \infty$)or the exact multivariate t-distribution (with $\nu < \infty$). In MNQ approach, we find $Q_{\alpha,V}$ such that $P\Big(\max|T_{l,n} \leq Q_{\alpha,V}) = 1 - \alpha$.

The MNQ method also helps to constructe the simultaneous confidence intervals. The simultaneous $(1 - \alpha)100\%$ confidence interval for $C\theta$ is

$$\big(C_{(l)}{}^T\widehat{\theta}_n^c - Q_{\alpha,V}\sqrt{\Big(C_{(l)}{}^T\widehat{\Gamma}_nC_{(l)}\Big)/n},\ C_{(l)}{}^T\widehat{\theta}_n^c + Q_{\alpha,V}\sqrt{\Big(C_{(l)}{}^T\widehat{\Gamma}_nC_{(l)}\Big)/n}\big). \tag{2.3}$$

Consider $G(\theta) = (G_1(\theta), \dots, G_c(\theta))'$ being a general nonlinear mapping from p-dimensional θ to c-dimensional G. Let B denote the Jacobian matrix with $B_{lm} = \partial G_l / \partial\theta_m$ evaluated at θ_0, and $B_l = (B_{l1} \dots, B_{lp})^T$. Let $D^* = B\Gamma B^T$ and $V^* = \text{diag}(D^*)^{-1/2} D^* \,\text{diag}(D^*)^{-1/2}$. Using the Delta method, the approximate simultaneous $100(1 - \alpha)\%$ confidence interval will be

$$\big(G_l(\widehat{\theta}_n^c) - Q_{\alpha,V^*}\sqrt{\Big(B_l^T\widehat{\Gamma}_nB_l\Big)/n}, G_l(\widehat{\theta}_n^c) + Q_{\alpha,V^*}\sqrt{\Big(B_l^T\widehat{\Gamma}_nB_l\Big)/n}\big). \tag{2.4}$$

In some applications, the collection of effect sizes are nonlinear monotone transformations of the parameters. For example, we obtain odds ratio from log odds ratio by applying the exponential function. Then the simultaneous $(1 - \alpha)100\%$ confidence interval for

$G[(C_{(l)})^T\theta]$, $l = 1, \ldots, c$, is

$$\left(G[C_{(l)}{}^T\widehat{\theta}_n^c - Q_{\alpha,V}\sqrt{\left(C_{(l)}{}^T\widehat{\Gamma}_n C_{(l)}\right)/n}], G[C_{(l)}{}^T\widehat{\theta}_n^c + Q_{\alpha,V}\sqrt{\left(C_{(l)}{}^T\widehat{\Gamma}_n C_{(l)}\right)/n}]\right). \tag{2.5}$$

2.3 Four multivariate models

To examine our methodology, we consider four different multivariate distributions: The multivariate normal, multivariate probit, quadratic exponential and gamma distributions. Gamma distribution is considered as an example of a skewed multivariate model. For the all mentioned distributions except the third one, the composite likelihood is constructed as sum of univariate likelihoods, whereas for the third distribution, the composite likelihood is constructed as conditional likelihood. Our methodology is not limited to these distributions and can be applied to other distributions as well.

Let X_i denote an $m_i \times p$ matrix containing the values of p covariates for the m_i individuals in the i^{th} cluster and $\beta = (\beta_1, \ldots, \beta_p)^T$ denote the vector of regression coefficients. Let $\vec{x}_{ij}$ denote the j^{th} row of the matrix X_i (this is the vector of covariates for individual j in cluster i).

2.3.1 Multivariate Gaussian distribution

Let $\{(y_i, X_i),\ i = 1, \cdots n\}$, denote the response and covariates arising from a multivariate normal model, with $y_i = X_i\beta + \epsilon_i, i = 1, \ldots, n$, and $m_i = m$. We assume that $\epsilon_i \sim N_m(0, \Sigma)$ where $\Sigma = (\sigma_{ij}), i, j = 1, \ldots, m$, is an arbitrary covariance matrix. The univariate composite likelihood is thus equal to

$$cl\,(\beta) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m}\left(-\frac{1}{2}\log(2\pi\sigma_{jj}) - \frac{1}{2\sigma_{jj}}(y_{ij} - \vec{x}_{ij}\beta)^2\right),$$

where the σ_{jj}'s are nuisance parameters and with $\vec{x}_{ij}$ denoting the j^{th} row of the matrix X_i. To estimate the regression coefficients, an iterative algorithm is used: Given the current estimate for the nuisance parameters σ_{jj}'s, we maximize the composite likelihood to obtain an estimate of

$$\widehat{\beta}_n^c = (\sum_{i=1}^{n} X_i^T W X_i)^{-1}\sum_{i=1}^{n} X_i^T W y_i,$$

where $W = \mathrm{diag}(\Sigma)^{-1}$, and given a current estimate for β, we use the sample covariance matrix of residuals to estimate Σ. To estimate the sample covariance we use the unbiased empirical estimator

$$\widehat{\sigma}_{jk}\big|_{\beta} \;=\; \frac{1}{n-p}\sum_{i=1}^{n}(y_{ij} - \vec{x}_{ij}\beta)(y_{ik} - \vec{x}_{ik}\beta), \quad 1 \le j,k \le m.$$

Based on the estimates $\widehat{\beta}_n^c$ and $\widehat{\Sigma}$, we obtain estimates for $H(\beta) = n^{-1}\left(\sum_{i=1}^{n} X_i^T W X_i\right)$ and $J(\beta) = n^{-1}\left(\sum_{i=1}^{n} X_i^T W\,\Sigma\,W X_i\right)$, with W being replaced by its estimate $\widehat{W} = \mathrm{diag}(\widehat{\Sigma})$. This is repeated until convergence is observed for $\widehat{\beta}_n^c$. The correlation is taken into account in estimating the covariance matrix, since $\widehat{W}\widehat{\sigma}_{jk}\widehat{W} = \widehat{\Sigma}$

$$Cov(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i^T \widehat{W} X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i^T \widehat{W}\widehat{\sigma}_{jk}\widehat{W} X_i\right)\left[\left(\sum_{i=1}^{n} X_i^T \widehat{W} X_i\right)^{-1}\right]^T$$

The result is similar to the ones in generalized estimating equations (GEE). However, GEE uses the model of the mean of the data and covariance matrix needs to be specified too. But in composite likelihood estimation, the model could be any marginal or conditional densities.

2.3.2 Multivariate probit model

Let $y_i^* = X_i\beta + \epsilon_i$ with $\epsilon_i \sim N_m(0, \Sigma)$ and $\Sigma = \sigma R$, where R is an $m \times m$ correlation matrix. The variables y_i^* are the latent response variables, and their dichotomized version of the latent variable with $y_{ij} = I(y_{ij}^* > 0), \ \ j = 1, \cdots, m$ yield the multivariate probit model. We therefore have that $P(y_{ij} = 1|X_i) \quad = \quad \Phi(\vec{x}_{ij}\beta/\sigma)$ where Φ denotes the univariate standard normal cumulative distribution function. It follows that the parameters β and σ are not fully identifiable in the model, and we can only estimate the ratio β/σ. To simplify notation, σ is set equal to 1 in what follows. The univariate composite log-likelihood function of the probit model is then formulated as

$$cl(\beta; y) \quad = \quad \sum_{i=1}^n \sum_{j=1}^m [y_{ij} \ \log \ \Phi\left(\vec{x}_{ij}\beta\right) + (1 - y_{ij}) \ \log\left(1 - \Phi\left(\vec{x}_{ij}\beta\right)\right)].$$

Denoting $\mu_{ij} = P(y_{ij} = 1|X_i)$, and $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$, we have

$$cl^{(1)}(\beta; y) \quad = \quad \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial_\beta}\right)^T \Pi_i^{-1}(y_i - \mu_i),$$

where $\Pi_i = \text{diag}(\text{var}(y_{i1}), \cdots, \text{var}(y_{im}))$, and $\text{var}(y_{ij}) = \mu_{ij}(1-\mu_{ij})$. This yields

$$H(\beta) = n^{-1} \textstyle\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial_\beta}\right)^T \Pi_i^{-1} \left(\frac{\partial \mu_i}{\partial_\beta}\right) \quad \text{and} \quad J(\beta) = n^{-1} \textstyle\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial_\beta}\right)^T \Pi_i^{-1} \,\text{cov}(y_i)\, \Pi_i^{-1} \left(\frac{\partial \mu_i}{\partial_\beta}\right).$$

To find the estimates $\widehat{\beta}_n^c$, we use the Newton-Raphson algorithm. Denote $\widehat{\mu}_{in} = \{\widehat{\mu}_{i1n}, \widehat{\mu}_{i2n}, \ldots, \widehat{\mu}_{imn}\}^T$, where $\widehat{\mu}_i = \Phi(X_i \widehat{\beta}_n^c)$. Let $\widehat{\Pi}_{in}$ denote the estimator of Π_i obtained by substituting $\widehat{\mu}_{ijn}$ for μ_{ij}. We estimate $H(\beta)$ and $J(\beta)$ as

$$\begin{aligned} \widehat{H}_n &= n^{-1} \textstyle\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial_\beta}\Big|_{\widehat{\beta}_n^c}\right)^T \widehat{\Pi}_{in}^{-1} \left(\frac{\partial \mu_i}{\partial_\beta}\Big|_{\widehat{\beta}_n^c}\right) \\ \widehat{J}_n &= n^{-1} \textstyle\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial_\beta}\Big|_{\widehat{\beta}_n^c}\right)^T \widehat{\Pi}_{in}^{-1} \;\; \widehat{\text{cov}}_n(y_i) \;\; \widehat{\Pi}_{in}^{-1} \left(\frac{\partial \mu_i}{\partial_\beta}\Big|_{\widehat{\beta}_n^c}\right), \end{aligned}$$

calculating the empirical variance as $\widehat{\text{cov}}_n(y_i) = (y_i - \widehat{\mu}_{in})(y_i - \widehat{\mu}_{in})^T$.

Computational Aspects

The score vector of the probit model is

$$S = \sum_{y_{ij}=1} \frac{\phi_1(\vec{x}_{ij}\beta)}{\Phi_1(\vec{x}_{ij}\beta)} (\vec{x}_{ij})^T - \sum_{y_{ij}=0} \frac{\phi_1(\vec{x}_{ij}\beta)}{1 - \Phi_1(\vec{x}_{ij}\beta)} (\vec{x}_{ij})^T$$

and the expectation of Hessian matrix

$$E(-\frac{\partial}{\partial \beta} S) = \sum_{i=1}^n \sum_{j=1}^m \frac{\phi_1^2(\vec{x}_{ij}\beta)}{\Phi_1(\vec{x}_{ij}\beta)(1 - \Phi_1(\vec{x}_{ij}\beta))} (\vec{x}_{ij})^T \vec{x}_{ij} \tag{2.6}$$

and

$$
\begin{aligned}
Var(S) = & \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\phi_1^2(\vec{x}_{ij}\beta)}{\Phi_1(\vec{x}_{ij}\beta)(1-\Phi_1(\vec{x}_{ij}\beta))}(\vec{x}_{ij})^T\vec{x}_{ij} \\
& + \sum_{i=1}^{n}\sum_{j\neq k} \frac{E(y_{ij}y_{ik})-\Phi_1(\vec{x}_{ij}\beta)\Phi_1(\vec{x}_{ik}\beta)}{\Phi_1(\vec{x}_{ij}\beta)(1-\Phi_1(\vec{x}_{ij}\beta))\Phi_1(\vec{x}_{ik}\beta)(1-\Phi_1(\vec{x}_{ik}\beta))}\phi_1(\vec{x}_{ij}\beta)\phi_1(\vec{x}_{ik}\beta)(\vec{x}_{ij})^T\vec{x}_{ik}
\end{aligned}
\tag{2.7}
$$

Where $E(y_{ij}y_{ik}) = \Phi_2(\vec{x}_{ij}\beta, \vec{x}_{ik}\beta, \rho)$. As the correlation of the latent variables $\rho = Corr(y^*_{ij}, y^*_{ik})$ can't be computed explicitly, we estimate it using the response variables,

$$
E(y_{ij}y_{ik}) = \sum_{j<k} \frac{y_{ij}y_{ik}}{\binom{m}{2}} \quad i = 1\cdots,n, \;\; j,k = 1,\cdots,m.
$$

In estimation of $\widehat{\beta}^c_n$, occasionally the value of $\vec{x}_{ij}\beta$ is either very small or large, which leads $\Phi(\vec{x}_{ij}\beta)$ to be very close to zero or one, which in turn makes the computation unstable. To avoid this problem, we adopt Demidenko [8] suggestion of approximation based on the following limits for the standard normal density function

$$
\begin{aligned}
\lim_{s\longrightarrow-\infty}\frac{\phi(s)}{s\Phi(s)} = -1, &\qquad \lim_{s\longrightarrow\infty}\frac{\phi(s)}{s(1-\Phi(s))} = 1, \\
s\Phi(s)+\phi(s) > 0, &\qquad \phi(s)-s(1-\Phi(s)) > 0, \forall s \in R
\end{aligned}
$$

Therefore first and second terms of the $Var(S)$ are approximated as

$$\frac{\phi_1^2(s)}{\Phi_1(s)(1-\Phi_1(s))} = \begin{cases} \frac{\phi_1^2(s)}{\Phi_1(s)(1-\Phi_1(s))}, & \text{if } |s| \le 5. \\ \phi(s) \times s, & \text{if } s > 5. \\ -\phi(s) \times s, & \text{if } s < -5. \end{cases}$$

Let $E = E(y_{ij}y_{ik})$, then

$$\frac{E-\Phi_1(s)\Phi_1(t)}{\Phi(s)(1-\Phi(s))\Phi(t)(1-\Phi(t))}\phi(s)\phi(t) = \begin{cases} \frac{E-\Phi(s)\Phi(t)}{\Phi(s)(1-\Phi(s))\Phi(t)(1-\Phi(t))}\phi(s)\phi(t), & \text{if } |t| \le 5, |s| \le 5. \\ (E-\Phi(t))\frac{\phi(t)s}{\Phi(t)(1-\Phi(t))}, & \text{if } |t| \le 5, s > 5. \\ \frac{-Es\phi(t)}{\Phi(t)(1-\Phi(t))}, & \text{if } |t| \le 5, s < -5. \\ (E-\Phi(s))\frac{\phi(s)t}{\Phi(s)(1-\Phi(s))}, & \text{if } t > 5, |s| \le 5. \\ st(E-1), & \text{if } t > 5, s > 5. \\ -Est, & \text{if } t > 5, s < -5. \\ \frac{-Et\phi(s)}{\Phi(s)(1-\Phi(s))}, & \text{if } t < -5, |s| \le 5. \\ -Est, & \text{if } t < -5, s > 5. \\ Est, & \text{if } t < -5, s < -5. \end{cases}$$

2.3.3 Quadratic exponential model

The quadratic exponential model is a popular tool that captures the mean function and within family correlation structure simultaneously. Therefore, it is used to model clustered binary data with intra-cluster interactions (Geys et al. [16]). In this model, the binary observations take values $y_{ij} \in \{-1, 1\}$. This coding for the response variable is used instead of 0 and 1 to provides a parametrization that is more suitable when success and failure demonstrate completely reversed situations. The joint distribution is given by

$$f_y(y_i) \quad \propto \quad \exp\left\{\sum_{j=1}^{m_i} \mu_{ij}^* y_{ij} + \sum_{j<j'} w_{ijj'}^* y_{ij} y_{ij'}\right\}, \tag{2.8}$$

where μ_{ij}^* is a parameter which describes the main effect of the measurements and $w_{ijj'}^*$ describes the association between pairs of measurements within the cluster y_i. Independence corresponds to the case that $w_{ijj'}^* = 0$ and positive or negative correlation corresponds to $w_{ijj'}^* > 0$ or $w_{ijj'}^* < 0$, respectively. For simplicity, we consider the case that $\mu_{ij}^* = \mu_i^*$ and $w_{ijj'}^* = w_i^*$, noting that our methodology can be readily applied to the general scenario as well. Under this simplification, Molenberghs and Ryan [33], showed that the joint distribution can be equivalently written in terms of $z_i = \sum_{j=1}^{m_i} \mathbb{I}(y_{ij} = 1)$ (the number of successes in the ith cluster) as

$$f_y(y_i) \propto \exp\{\mu_i z_i - w_i z_i (m_i - z_i)\},$$

where $w_i = 2w_i^*$ and $\mu_i = 2\mu_i^*$.

Specifying the normalizing constant in (2.8) is famously difficult, yet necessary to compute the full likelihood function. Employing composite likelihood approach helps to get rid of such an intensive calculation. Replacing the joint distribution function with the conditional distributions leads to a conditional composite likelihood function $cl(\mu, w; y) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \log f(y_{ij} | \{y_{ij'}\}, j' \neq j)$, which does not require computation of the normalizing constant.

The two conditional probabilities are defined as

$$p_{is} = \frac{\exp\{\mu_i - w_i(m_i - 2z_i + 1)\}}{1 + \exp\{\mu_i - w_i(m_i - 2z_i + 1)\}}, \qquad p_{if} = \frac{\exp\{-\mu_i + w_i(m_i - 2z_i - 1)\}}{1 + \exp\{-\mu_i + w_i(m_i - 2z_i - 1)\}}.$$

where p_{is} is the conditional probability of one more success, given $z_i - 1$ successes and $m_i - z_i$ failures, while p_{if} is the conditional probability of one more failure, given z_i successes and $m_i - z_i - 1$ failures. Note that $p_{if} \neq 1 - p_{is}$, because of the term $m_i - 2z_i \pm 1$. The composite likelihood can now be expressed as

$$cl(\mu, w; y) = \sum_{i=1}^{n} \left(z_i \log p_{is} + (m_i - z_i) \log p_{if} \right). \tag{2.9}$$

The obtained form of the composite likelihood shows that a logistic regression approach can be used to estimate the parameters. We model a covariate effect by using the linear model $\mu_i = X_i\beta$, with $w_i = w$ interpreted as an additional parameter. That is, for the parameter w, the value of the covariate is set to $-(m_i - 2z_i + 1)$ when $y_{ij} = 1$ and $-(m_i - 2z_i - 1)$ when $y_{ij} = -1$. This allows us to obtain MCLE estimates of both β and w using iterative re-weighted least squares, commonly used to solve logistic regression maximization problems. To estimate the covariance of $\widehat{\beta}_n^c$, we computed $\widehat{J}_n$ as the empirical variance of the score vector,

$$\begin{pmatrix} \frac{\partial f_{y_i}}{\partial \mu} \\ \\ \frac{\partial f_{y_i}}{\partial w} \end{pmatrix} = \begin{pmatrix} z_i(1-p_{is}) - (m_i - z_i)(1-p_{if}) \\ \\ -z_i(m_i - 2z_i + 1)(1-p_{is}) - (m_i - z_i)(m_i - 2z_i - 1)(1-p_{if}) \end{pmatrix}$$

plugging in estimates of μ_i^*, w^* throughout. The Hessian matrix $\widehat{H}_n$ is estimated using the result from fitting the logistic model in R, see Geys et al. [16].

2.3.4 Multivariate gamma distribution

Given n independent multivariate gamma vectors $y = (y_1, y_2, \ldots, y_n)^T$, with $y_i = (y_{i1}, \ldots, y_{im})^T$. The univariate composite log-likelihood function for the multivariate gamma model can be formulated as

$$cl(\beta; y) \;\; = \;\; \sum_{i=1}^{n} \sum_{j=1}^{m} \left(-\frac{\nu y_{ij}}{\mu_{ij}} - \nu \log \mu_{ij} + \nu log \nu + (\nu - 1) \log y_{ij} - \log \Gamma(\nu) \right),$$

where $\mu_{ij} = E(y_{ij})$, ν is the shape parameter, and μ_{ij}/ν is the scale parameter. We used the log link to define the mean parameter: $\mu_{ij} = \exp\{\vec{x}_{ij}\beta\}$. Denote $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$. Under this set up, we have

$$cl^{(1)}(\beta; y) \;\; = \;\; \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V(\mu)_i^{-1} (y_i - \mu_i),$$

where $V_i = \text{diag}(\mu_{i1}^2, \cdots, \mu_{im}^2)/\nu$, and

$$\begin{aligned} H(\beta) &= n^{-1}\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T V_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right), \\ J(\beta) &= n^{-1}\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T V_i^{-1}\,\text{cov}(y_i)\,V_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right). \end{aligned}$$

The dispersion parameter is $\frac{1}{\nu} = \frac{D(6(n-p)+nD)}{6(n-p)+2nD}$, where $D = \frac{2}{nm-p}\sum_{i,j}\left(\frac{y_{ij}-\mu_{ij}}{\mu_{ij}} + \log\frac{\mu_{ij}}{y_{ij}}\right)$. Let $\widehat{V}_{in}$ denote the estimator of V_i obtained by substituting $\widehat{\mu}_{ijn}$ for μ_{ij}. We estimate $H(\beta)$ and $J(\beta)$ as

$$\begin{aligned} \widehat{H}_n &= n^{-1}\sum_{i=1}^{n} X_i^T \widehat{V}_{in}^{-1} X_i, \\ \widehat{J}_n &= n^{-1}\sum_{i=1}^{n} X_i^T \widehat{V}_{in}^{-1}\ \widehat{\text{cov}}_n(y_i)\ \widehat{V}_{in}^{-1} X_i, \end{aligned}$$

with empirical variance $\widehat{\text{cov}}_n(y_i) = (y_i - \widehat{\mu}_{in})(y_i - \widehat{\mu}_{in})^T$, where where $\widehat{\mu}_i$ is the vector $\widehat{\mu}_i = \exp\{X_i\widehat{\beta}_n^c\}$.

Chapter 3

Simulation Results and Application of Multiple Comparisons Using Composite Likelihood

In this chapter, we evaluate MNQ approach numerically. In Section 3.1, we conduct simulation studies to evaluate empirical performance of the proposed method. Finally, in Section 3.2 we analyze the depression data and kidney function data sets to demonstrate the practical utility of the method. We conclude the chapter with a brief discussion of the results.

3.1 Simulation results

To examine how our proposed approach works for multiple hypothesis testing in clustered data, we evaluated it as well as some well-known testing procedures through simulations. Two different global null hypothesis on the regression coefficients $\beta_1, \cdots, \beta_p$ are tested: (a) many-to-one comparisons, $H_{01} : \cap_{i=2}^{p}\{\beta_1 = \beta_i\}$, (b) all pairwise comparisons $H_{02} : \cap_{1\leq i,j\leq n}\{\beta_i = \beta_j\}$.

In addition to MNQ approach, multiple comparisons is performed based on several approaches including Bonferroni, Dunn-Sidak, as well as multi-stage Holm and Scheffé's method. In order to show the result in the situation that correlation structure is ignored, a naive method is also considered in which the threshold is computed by ignoring the existed intra-cluster correlation.

We use univariate composite likelihood estimation, then such a misspecification is equivalent to $H(\theta) = J(\beta)$ in (1.1). This results in an estimate of $\widehat{\Gamma}_n = \widehat{H}_n^{-1}$. This misspecified scenario is included for comparison, and we consider it only with the MNQ multiple comparison method (that is, the MNQ cutoff is calculated based on V estimated by plugging in $\widehat{\Gamma}_n = \widehat{H}_n^{-1}$). The equi-coordinate critical values for multivariate normal and multivariate t distributions are obtained using the R package `mvtnorm` [25].

In our simulations, we study the four models described in the previous section. For each model, a different sample size is needed for our asymptotic approximations to be valid. We determine this sample size with an initial simulation. For each simulation setting, 10 000 simulated data sets were generated and the family-wise type I error rate was set to

0.05. The standard deviation for the observed FWER is hence approximately 0.002. These preliminary simulation results are given in Table 3.1. We observe that $n = 200, 500, 700$ and 3000 are required for the multivariate normal, multivariate probit, quadratic exponential and gamma models to maintain FWER within two standard deviations away from 0.05, respectively. These are the sample sizes used for the simulation results which follow.

Table 3.1: FWER for different sample sizes

model	Sample size				
	200	500	700	1000	4000
multivariate normal	0.0509	0.0492	0.0483	0.0495	0.050
multivariate probit	0.0576	0.0501	0.0511	0.0506	0.0511
quadratic exponential	0.0580	0.0543	0.0519	0.0520	0.0504

To compute the power of each of the different methods, we consider two different alternative scenarios: (1) a_1 with only one non-zero parameter with a large effect size, (2) a_2 with five true non-zero parameters but with small effect sizes for all. We are interested in the ability of the test to reject both the global and individual null hypotheses. Under the alternative scenario a_1, we calculate the power to reject the global hypothesis (denoted as "a_1" in the tables) and for the alternative configuration a_2, we calculate both the power to reject the global null hypothesis (denoted as "a_2" in the tables) and the sum of the five powers to rejected the five individual true alternatives (denoted as "ind a_2" in the tables).

3.1.1 Multivariate Gaussian model

We consider the multivariate normal model with $n = 200$ clusters, cluster size $m = 4$ or 10, and the number of covariates set to $p = 10$ or 20. Four different Σ scenarios are considered: 1) three exchangeable structures with $\sigma^2 = 0.8$ and $\rho = \text{cov}(y_{ij}, y_{ik}) = 0$, 0.2 or 0.5; 2) one arbitrary structure, where $\Sigma = ((1.3, 0.9, 0.5, 0.3)^T, (0.9, 1.9, 1.3, 0.3)^T, (0.5, 1.3, 1.3, 0.1)^T, (0.3, 0.9, 0.1, 0.7)^T)$. In each simulation, the $m \times p$ covariate matrix X_i is obtained by randomly sampling from normal distributions.

We consider here the many-to-one comparisons where the first parameter is taken as the baseline. Under the global null hypothesis H_0, the true value of the regression parameters is set to $\beta^T = 0$, and the power is calculated under two different alternative configurations $\beta_{a_1}^T = (0, 0, 0, 0.032, 0, \ldots, 0)$ and $\beta_{a_2}^T = (0, 0.008, 0.01, -0.03, 0.005, -0.01, 0, \ldots, 0)$. Under β_{a_1}, there is only one true alternative, and we evaluate the power to reject the global null hypothesis. Under β_{a_2}, there are five true alternatives and we evaluate both the power to reject the global null and the sum of five powers to reject the five true alternatives.

Table 3.2 (three exchangeable Σ scenarios) and Table 3.3 (general Σ) summarize the results of our simulations. Overall, it is shown that the MNQ method has the best performance among all of the multiple comparison procedures. A comparison of MNQ and naive MNQ clearly shows the cost of ignoring these correlations: the FWER of MNQ is superior to that of naive MNQ for $\rho \neq 0$ (when $\rho = 0$ the two methods are almost identical). Notably, the power of the naive MNQ is occasionally higher than that of MNQ, however, this is only due to the over-inflation of the naive MNQ's FWER. The small effect sizes chosen under a_2 allow us to detect more subtle differences in the performance of the methods. Notice that for the rejection of the global null hypothesis, Holm's method has exactly the same power as that of the Bonferroni method. However, for the individual powers, Holm's method has higher power to reject individual hypothesis than the Bonferroni method.

We also evaluate the efficiency of the maximum composite likelihood estimator versus maximum likelihood estimator. That is, we compute the ratio of the standard error of the MLE versus that of the MCLE. For small ρ, the ratio is close to one and as ρ increases, the ratio decreases. This demonstrates that the efficiency of composite likelihood estimator decreases with the increase of the intra-cluster correlation, as expected.

It is observed that with increasing ρ and p for the multivariate distribution of the clustered data, the power of Bonferroni was not substantially smaller. The increase of ρ will increase the variability of each estimate $C_{(l)T}\widehat{\beta}$ and hence decrease the power. When ρ increases from 0 to 0.5, we observe about 10% increase in the variability of the estimates and this is in compatible with the 5-10% power loss that we observe. We also conduct simulations with smaller sample sizes $n = 50$, and $n = 100$. It is shown that with n greater than 50, the statistics based on the plug-in estimate of the Godambe information matrix has satisfactory performance. Table 3.4 shows for $n = 50$, MNQ and Bonferroni maintains the FWER only for normal distribution, whereas for other two distributions, MNQ and Bonferroni tend to be liberal. The control of FWER is greatly improved with $n = 100$ for all three multivariate distributions. As the multivariate distribution of T_n can be approximated as a multivariate t distribution with n degrees of freedom, we conduct simulations to investigate the multivariate t approximation. Table 3.5 shows that the multivariate t approximation provides improved control of FWER for normal, probit and quadratic exponential distribution compared to multivariate normal approximation with the same sample of $n = 50$.

Table 3.2: Simulations results for many to one comparisons in the multivariate normal model with exchangeable Σ

	ρ	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé	efficiency
FWER	0	4	10	0.0545	0.0553	0.0419	0.0427	0.0419	0.0007	0.9983
a_1				0.8164	0.8166	0.7894	0.7918	0.7894	0.2801	
a_2				0.8057	0.8053	0.7617	0.7738	0.7617	0.2226	
ind a_2				0.9080	0.9079	0.8417	0.8503	0.8848	0.2242	
FWER			20	0.0511	0.0502	0.0352	0.0363	0.0352	0.0000	0.9980
a_1				0.7487	0.7476	0.7062	0.7086	0.7062	0.0259	
a_2				0.7150	0.7134	0.6687	0.6628	0.6687	0.0162	
ind a_2				0.7698	0.7674	0.7081	0.7007	0.7518	0.0162	
FWER		10	10	0.0479	0.0471	0.0375	0.0378	0.0375	0.0001	0.9989
a_1				0.9983	0.9983	0.9979	0.9980	0.9979	0.9284	
a_2				0.9993	0.9993	0.9986	0.9990	0.9986	0.8792	
ind a_2				1.4822	1.4816	1.4219	1.4284	1.4896	0.8933	
FWER			20	0.0487	0.0485	0.0363	0.0373	0.0363	0.0000	0.9986
a_1				0.9981	0.9980	0.9967	0.9969	0.9967	0.5428	
a_2				0.9978	0.9977	0.9963	0.9957	0.9963	0.4137	
ind a_2				1.3439	1.3406	1.2759	1.2776	1.3267	0.4139	
FWER	0.2	4	10	0.0494	0.0670	0.0389	0.0397	0.0389	0.0001	0.9453
a_1				0.7760	0.8113	0.7453	0.7476	0.7453	0.2481	
a_2				0.7630	0.8044	0.7224	0.7280	0.7224	0.1831	
ind a_2				0.8556	0.9268	0.7939	0.8032	0.8317	0.1845	
FWER			20	0.0533	0.0734	0.0390	0.0397	0.0390	0.0000	0.9430
a_1				0.7044	0.7490	0.6591	0.6617	0.6591	0.0191	
a_2				0.6713	0.7200	0.6187	0.6148	0.6187	0.0102	
ind a_2				0.7106	0.7777	0.6476	0.6438	0.6937	0.0102	
FWER		10	10	0.0467	0.1019	0.0357	0.0365	0.0357	0.0003	0.8685
a_1				0.9912	0.9974	0.9875	0.9880	0.9875	0.8098	
a_2				0.9925	0.9983	0.9877	0.9897	0.9877	0.7295	
ind a_2				1.3506	1.5795	1.2871	1.2968	1.3407	0.7374	
FWER			20	0.0468	0.1057	0.0320	0.0331	0.0320	0.0000	0.8636
a_1				0.9868	0.9970	0.9813	0.9819	0.9813	0.3114	
a_2				0.9820	0.9964	0.9758	0.9752	0.9758	0.2146	
ind a_2				1.2100	1.4205	1.1495	1.1545	1.1891	0.2146	
FWER	0.5	4	10	0.0513	0.0977	0.0390	0.0398	0.0390	0.0007	0.7491
a_1				0.7235	0.8129	0.6867	0.6904	0.6867	0.1947	
a_2				0.6922	0.8042	0.6615	0.6571	0.6615	0.1497	
ind a_2				0.7570	0.9391	0.7208	0.7074	0.7533	0.1502	
FWER			20	0.0510	0.1029	0.0377	0.0385	0.0377	0.0000	0.7343
a_1				0.6420	0.7526	0.5950	0.5985	0.5950	0.0140	
a_2				0.6031	0.7322	0.5437	0.5508	0.5437	0.0076	
ind a_2				0.6369	0.8035	0.5677	0.5750	0.6109	0.0076	
FWER		10	10	0.0520	0.2079	0.0410	0.0417	0.0410	0.0000	0.6070
a_1				0.9570	0.9936	0.9466	0.9469	0.9466	0.6125	
a_2				0.9555	0.9982	0.9438	0.9431	0.9438	0.5062	
ind a_2				1.1914	1.6903	1.1367	1.1372	1.1877	0.5096	
FWER			20	0.0459	0.2271	0.0328	0.0337	0.0328	0.0000	0.5898
a_1				0.9403	0.9938	0.9224	0.9243	0.9224	0.1408	
a_2				0.9222	0.9948	0.8932	0.8968	0.8932	0.0871	
ind a_2				1.0589	1.5362	0.9907	0.9983	1.0306	0.0871	

a_1, a_2: two global powers, ind a_2: individual power, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, ρ : correlation, m: cluster size, p : length of β

Table 3.3: Simulations results for many to one comparisons in multivariate normal with unstructured Σ

	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé
FWER	4	10	0.0464	0.0729	0.0345	0.0358	0.0345	0.0004
a_1			0.6348	0.7089	0.5962	0.5992	0.5962	0.1358
a_2			0.5123	0.6045	0.4763	0.4714	0.4763	0.0614
ind a_2			0.1100	0.1341	0.1019	0.1022	0.1071	0.0123
FWER		20	0.0390	0.0664	0.0285	0.0290	0.0285	0.0000
a_1			0.5205	0.6081	0.4694	0.4736	0.4694	0.0046
a_2			0.3913	0.4864	0.3378	0.3428	0.3378	0.0011
ind a_2			0.0811	0.1018	0.0687	0.0702	0.0749	0.0002
FWER	10	10	0.0472	0.0407	0.0360	0.0367	0.0360	0.0004
a_1			0.6310	0.6102	0.5906	0.5940	0.5906	0.1198
a_2			0.5025	0.4779	0.4560	0.4599	0.4560	0.0537
ind a_2			0.1088	0.1028	0.0974	0.0982	0.1032	0.0107
FWER		20	0.0361	0.0302	0.0262	0.0267	0.0262	0.0000
a_1			0.5078	0.4865	0.4585	0.4615	0.4585	0.0025
a_2			0.3668	0.3448	0.3148	0.3167	0.3148	0.0010
ind a_2			0.0742	0.0698	0.0637	0.0641	0.0692	0.0002

a_1, a_2: two global powers, ind a_2: individual power, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, ρ : correlation, m: cluster size, p : length of β

Table 3.4: Simulations results for different models with small sample sizes

		Normal			Probit			Quad Exp		
	n	MNQ	naive	Bonf	MNQ	naive	Bonf	MNQ	naive	Bonf
FWER	50	0.0531	0.0962	0.0419	0.0839	0.0837	0.0674	0.1139	0.0003	0.0912
a_1		0.1750	0.2645	0.1502	0.1353	0.1411	0.1139	0.1551	0.0015	0.1273
a_2		0.1760	0.2628	0.1496	0.2849	0.3034	0.2398	0.1810	0.0034	0.1480
FWER	100	0.0474	0.0931	0.0379	0.0631	0.0808	0.0496	0.0704	0.0000	0.0551
a_1		0.3751	0.4836	0.3382	0.1832	0.2203	0.1594	0.1458	0.0004	0.1190
a_2		0.3563	0.4754	0.3181	0.5159	0.5835	0.4604	0.2016	0.0007	0.1703

a_1, a_2: two global powers, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, n : number of clusters

Table 3.5: Simulations results using multivariate t approximation with $n = 50$

model		MNQ	naive	Bonf
Normal	FWER	0.0509	0.0896	0.0509
	a_1	0.1513	0.2245	0.1547
	a_2	0.1470	0.2241	0.1503
Probit	FWER	0.0667	0.0660	0.0673
	a_1	0.1112	0.1130	0.1152
	a_2	0.1206	0.1218	0.1143
	ind a_2	0.0227	0.0237	0.0214
Quad Exp	FWER	0.0934	0.0001	0.0912
	a_1	0.1305	0.0008	0.1269
	a_2	0.1629	0.0013	0.1600

This result is consistent with the result from GEE since both have misspecified covariance structure.

3.1.2 Multivariate Probit model

Here, we consider $n = 500$ clusters with a cluster size $m = 4$, or 10. The binary variables are generated by dichotomizing latent multivariate normal variables with a threshold of zero. For each cluster, an $m \times p$ covariate matrix X_i, with $p = 10$ or 20, is obtained by randomly sampling from normal distributions. The regression coefficients under the global null hypothesis is $\beta^T = 0$ and the two alternative configurations are $\beta_{a1}^T = (0, 0, 0, 0.03, 0, \ldots, 0)$ and $\beta_{a2}^T = (0, 0.008, 0.01, -0.03, 0.005, -0.01, 0, \ldots, 0)$. The latent multivariate random vector has a mean $X_i\beta$ and a correlation matrix with ρ on the off-diagonals and $\sigma = 1$. Here, we consider $\rho = 0$, or 0.5.

The empirical results are given in Table 3.6. The results show that the MNQ method has overall the best performance. We note though that for the two settings when $\rho = 0.5$

and $p = 20$, the MNQ method has FWER more than 2 standard deviations away from 0.05. Similarly to the multivariate normal setting, the naive MNQ for the multivariate probit model has large FWER when $\rho = 0.5$. For the global hypothesis, the Sidák method has higher power than that of the Bonferroni and Holm method, whereas the Holm method has higher power to reject individual null hypotheses than the Bonferroni and Sidák method.

3.1.3 Quadratic exponential model

Here, we take a total of $n = 700$ clusters, and $p = 10$ or 20 predictors. The number of observations within each clusters, m_i, varies between clusters and is uniformly sampled from $\{4, 5, 6, 7, 8\}$. The $m_i \times p$ covariate matrix X_i is sampled from a standard normal distribution. We also consider two different values for the interaction parameter: $w = 0$ or 0.5. The null value of the regression coefficients is $\beta^T \equiv 0$ and the two alternative configurations are to $\beta_{a1}^T = (0, 0, 0, 0.12, 0, \ldots, 0)$ and $\beta_{a2}^T = (0, 0.08, 0.12, -0.03, 0.05, -0.08, 0, \ldots, 0)$. The empirical FWER and power are computed and summarized in Table 3.7. Overall, MNQ has clearly the best performance.

Table 3.6: Simulation results for many to one comparisons in the probit model

	ρ	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé
FWER	0	4	10	0.0530	0.0506	0.0413	0.0424	0.0413	0.0001
a_1				0.8700	0.8705	0.8477	0.8496	0.8477	0.3420
a_2				0.9114	0.9109	0.8885	0.8907	0.8885	0.3572
ind a_2				1.0828	1.0779	1.0193	1.0305	1.0682	0.3590
FWER			20	0.0528	0.0503	0.0389	0.0395	0.0389	0.0000
a_1				0.8258	0.8232	0.7902	0.7924	0.7902	0.0460
a_2				0.8547	0.8511	0.8149	0.8159	0.8149	0.0410
ind a_2				0.9436	0.9389	0.8847	0.8825	0.9308	0.0410
FWER		10	10	0.0526	0.0515	0.0423	0.0428	0.0423	0.0005
a_1				0.9996	0.9996	0.9995	0.9995	0.9995	0.9641
a_2				1.0000	1.0000	1.0000	1.0000	1.0000	0.9695
ind a_2				1.6649	1.6594	1.5839	1.5939	1.6658	1.0024
FWER			20	0.0527	0.0508	0.0364	0.0375	0.0364	0.0000
a_1				0.9993	0.9995	0.9985	0.9985	0.9985	0.6596
a_2				1.0000	0.9999	0.9997	0.9997	0.9997	0.6603
ind a_2				1.4867	1.4780	1.4057	1.4062	1.4624	0.6607
FWER	0.5	4	10	0.0508	0.0793	0.0393	0.0404	0.0393	0.0003
a_1				0.8102	0.8601	0.7808	0.7841	0.7808	0.2726
a_2				0.8530	0.9038	0.8305	0.8258	0.8305	0.2689
ind a_2				0.9852	1.1028	0.9321	0.9334	0.9768	0.2708
FWER			20	0.0585	0.0915	0.0406	0.0415	0.0406	0.0000
a_1				0.7578	0.8196	0.7082	0.7106	0.7082	0.0264
a_2				0.7891	0.8534	0.7365	0.7428	0.7365	0.0247
ind a_2				0.8637	0.9712	0.7855	0.7963	0.8330	0.0247
FWER		10	10	0.0513	0.1437	0.0402	0.0412	0.0402	0.0005
a_1				0.9900	0.9979	0.9871	0.9876	0.9871	0.8017
a_2				0.9952	0.9997	0.9966	0.9939	0.9966	0.8038
ind a_2				1.4075	1.7926	1.3520	1.3552	1.4154	0.8147
FWER			20	0.0543	0.1622	0.0382	0.0389	0.0382	0.0000
a_1				0.9862	0.9974	0.9784	0.9787	0.9784	0.3081
a_2				0.9935	0.9998	0.9894	0.9883	0.9894	0.3006
ind a_2				1.2873	1.6251	1.2248	1.2218	1.2777	0.3006

a_1, a_2: two global powers, ind a_2: individual power, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, ρ : correlation, m: cluster size, p : length of β

Table 3.7: Simulation results for many to one comparisons in the quadratic exponential model

	w	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé
FWER	0	4	10	0.0514	0.0562	0.0400	0.0403	0.0400	0.0001
a_1				0.5390	0.5534	0.5010	0.5046	0.5010	0.0777
a_2				0.7067	0.7240	0.6573	0.6636	0.6573	0.0935
ind a_2				0.9779	1.0283	0.8826	0.8888	0.9373	0.0986
FWER			20	0.0561	0.0767	0.0404	0.0412	0.0404	0.0000
a_1				0.4551	0.4853	0.3990	0.4021	0.3990	0.0025
a_2				0.6040	0.6365	0.5237	0.5403	0.5237	0.0027
ind a_2				0.7731	0.8347	0.6514	0.6679	0.7029	0.0027
FWER		10	10	0.0491	0.0549	0.0381	0.0384	0.0381	0.0001
a_1				0.5391	0.5535	0.5010	0.5046	0.5010	0.0779
a_2				0.7066	0.7239	0.6573	0.6636	0.6573	0.0934
ind a_2				0.9780	1.0284	0.8826	0.8890	0.9373	0.0985
FWER			20	0.0561	0.0767	0.0404	0.0412	0.0404	0.0000
a_1				0.4548	0.4849	0.3989	0.4020	0.3989	0.0026
a_2				0.5971	0.6309	0.5255	0.5361	0.5255	0.0013
ind a_2				0.7681	0.8316	0.6527	0.6688	0.7043	0.0013
FWER	0.5	4	10	0.0521	0.0000	0.0417	0.0424	0.0417	0.0002
a_1				0.7864	0.0307	0.7546	0.7582	0.7546	0.2329
a_2				0.9050	0.0444	0.8800	0.8772	0.8800	0.2531
ind a_2				1.5136	0.0452	1.4102	1.4089	1.4915	0.2753
FWER			20	0.0509	0.0000	0.0377	0.0383	0.0377	0.0000
a_1				0.7214	0.0158	0.6739	0.6769	0.6739	0.0178
a_2				0.8460	0.0148	0.7998	0.7976	0.7998	0.0132
ind a_2				1.2902	0.0150	1.1532	1.1612	1.2141	0.0134
FWER		10	10	0.0521	0.0000	0.0417	0.0424	0.0417	0.0002
a_1				0.7864	0.0307	0.7546	0.7582	0.7546	0.2329
a_2				0.9141	0.0407	0.8800	0.8855	0.8800	0.2518
ind a_2				1.5326	0.0416	1.4102	1.4261	1.4915	0.2746
FWER			20	0.0509	0.0000	0.0378	0.0384	0.0378	0.0000
a				0.7202	0.0161	0.6731	0.6760	0.6731	0.0178
a_2				0.8460	0.0148	0.7998	0.7976	0.7998	0.0132
ind a_2				1.2902	0.0150	1.1532	1.1612	1.2141	0.0134

a_1, a_2: two global powers, ind a_2: individual power, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, w : association parameter, m: cluster size, p : length of β

Table 3.8: Simulations results for all pairwise comparisons in the multivariate normal, probit, and quadratic exponential models

model		ρ	MNQ	naive	Bonf	S-D	Scheffé	Tukey
normal	FWER	0	0.0537	0.0562	0.0411	0.0420	0.0038	0.0536
	a1		0.9274	0.9266	0.9096	0.9113	0.6115	0.9256
	a2		0.9800	0.9807	0.9735	0.9740	0.8173	0.9792
	FWER	0.5	0.0484	0.1101	0.0358	0.0365	0.0032	0.0489
	a1		0.8611	0.9245	0.8325	0.8346	0.4769	0.8587
	a2		0.9492	0.9775	0.9346	0.9361	0.6854	0.9482
probit	FWER	0	0.0534	0.0494	0.0409	0.0412	0.0026	0.0524
	a1		0.9792	0.9790	0.9745	0.9747	0.7972	0.9791
	a2		0.9961	0.9961	0.9946	0.9946	0.9321	0.9959
	FWER	0.5	0.0523	0.0864	0.0394	0.0394	0.0023	0.0514
	a1		0.9586	0.9754	0.9467	0.9484	0.6991	0.9577
	a2		0.9885	0.9938	0.9842	0.9848	0.8707	0.9884
quad. exp.	FWER	0	0.0534	0.0631	0.0399	0.0407	0.0018	0.0530
	a1		0.7710	0.7869	0.7270	0.7301	0.3224	0.7678
	a2		0.9706	0.9741	0.9613	0.9621	0.7348	0.9701
	FWER	0.5	0.0548	0.0000	0.0388	0.0393	0.0014	0.0535
	a1		0.9360	0.0197	0.9199	0.9213	0.6417	0.9356
	a2		0.9976	0.2855	0.9957	0.9958	0.9408	0.9974

a_1, a_2: two global powers, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak, p : length of β

3.1.4 Multivariate gamma distribution

To generate a multivariate gamma model, let g_1 be $m \times 1$ independent vectors from a gamma distribution with shape parameters γ_1, a positive vector of dimension m. Define $G = Kg_1$, where K is a full rank matrix with all entries equal to either zero or one that follows some properties [38]. (K is called the incidence matrix). Then G has a multivariate gamma distribution with shape parameter $\alpha = K\gamma_1$ and covariance matrix $\Sigma = K\Gamma_1 K^T$, where the (diagonal) matrix Γ_1 is the variance matrix of g_1.

In the simulation $\nu = 1$, and under the global null hypothesis H_0, the true value of the regression parameters is set to $\beta = 0.75$, and the power is calculated under two different alternative configurations $\beta_{a_1}^T = (0.75, 0.75, 0.68, 0.75, \ldots, 0.75)$ and also $\beta_{a_2}^T =$

$(0.75, 0.80, 0.68, 0.70, 0.79, 0.69, 0.75, \ldots, 0.75)$. We simulate 10 000 data sets with $m = 3$, and $p = 10$. We perform many-to-one comparisons with the MNQ, naive MNQ, Bonferroni, Dunn-Sidák, Holm and Scheffé method. We consider both independent and correlated cases. We simulate with the sample size $n = 3\,000$ as we found that it takes at least $n = 3\,000$ for the MNQ method to have the FWER fall within 2 standard deviations away from 0.05. This larger sample size is expected for a skewed distribution such as the multivariate gamma. Among all the methods, the MNQ method continues to achieve the highest power and exhibits the best performance. The results are presented in Table 3.9.

Table 3.9: FWER and power for many to one comparisons in multivariate gamma distribution

		MNQ	naive	Bonf	S-D	Schéffe
FWER	independent	0.0554	0.0507	0.0437	0.0444	0.0003
a_1		0.8763	0.8777	0.8508	0.8531	0.3055
a_2		0.9906	0.9899	0.9856	0.9862	0.4526
FWER	correlated	0.0588	0.3427	0.0468	0.0479	0.0003
a_1		0.8223	0.9883	0.7853	0.7877	0.2378
a_2		0.9778	0.9999	0.9638	0.9653	0.3683

a_1, a_2: two global powers, MNQ: multivariate normal quantile method, S-D: Dunn-Sidak

3.2 Application to real data

In this section, the method is applied to two different data set.

3.2.1 Analysis of kidney function data

To examine the performance of the proposed methodology, we analyze data from a diabetic nephropathy (DN) study at the University of Michigan. DN is damage to the kidneys, caused by the destruction of the kidney's blood vessels by high blood sugar levels. This study was performed to determine if any biomarkers, among 500 candidate genes, have important influence on the risk of DN, as part of a therapeutic program.

In the study, Glomerular filtration rate (GFR), relating to renal function of 35 patients with abnormal DN was assessed at multiple time points, and the number of these measurements varies between 10 to 15 across the patients and in total 402 measurements were collected. A binary factor showing the kind of treatment was also recorded.

In the original data set, there were 500 candidate covariates, each one representing a gene. In the analysis, a binary response variable is created by dichotomizing GFR using 100 as the cut-off point, that is $y_{ij} = 1$ if GFR of the patient j at time i was atleast 100. First we need to find a suitable model that defines the effects of genes and treatments on GFR. As there are more than 500 covariates, including such large number of covariates in a model fitting process may cause some problems such as over-fitting or singularity. So, it is beneficial to shrink the number of covariate by finding the most significant ones and using only those factors in the analysis. To achieve this goal, generalized regression with L1-constraint on the parameters (LASSO, Tibshirani [47]) is used. As the response variable is binary, a logistic regression model is fitted to 500 covariates and the parameters are estimated by minimizing the least-squares, adding a penalty term that keeps the absolute size of the regression parameter, $\alpha||\beta||_1$, less than a known value. The constraint causes some coefficients to be shrunken to zero exactly and then 9 covariates with the most significant effect on the response variable are chosen. So just 9 genes, which have the most effect on GFR will be used in the next steps of the analysis(using the R package `lasso2` [31]).

Next, we would like to compare the effect of the 9 selected genes on the binary response variable based on GFR. The presence of correlation among repeated measurements within each patient can not be ignored. This problem can be handled by performing a multiple comparison test using composite likelihood estimation. Considering the binary response variable, we fit a quadratic exponential model to the data. Each patient is considered as a cluster and equal intra-correlation, $w_i = w$ for all clusters is assumed.

We define a quadratic exponential model such as $f_Y(y_i;\Theta) = \exp\{\Theta_i W_i - A(\Theta)\}$ and the vector of parameters Θ is assumed to be the linear model $X\beta$, with the design matrix X contains the level of genes and β represents the vector of gene's effects. To achieve a composite likelihood function as (2.9), m_i and z_i are considered as the number of measurements for the patient i and the number of GFR values measured from patient i which are greater than or equal to 100, respectively. Association parameter w^* in can be treated as an intercept that is multiplied by the value $(m_i - z_i + 1)$ for the response value 1(the GFR values greater than or equal to 100) and value $(m_i - z_i - 1)$ otherwise.

The parameters are estimated fitting a logistic model to the 9 chosen genes adjusting for the type of treatment that each patient received, on the defined binary response variable. Covariance of regression coefficients (gene effects), is estimated by MNQ method regarding the correlation structure and also by the naive method which ignores it.

Here, we compare all pair-wise effects of the predictors (genes) on the response (GFR), namely, $H_{0,i,j} = \{\beta_i = \beta_j, 1 \leq i, j \leq 9\}$, for a total of 36 null hypotheses. We consider the MNQ, Bonferroni and naive multiple testing approaches. All three methods reject the null hypothesis $H_0 = \cap_{i<j} H_{0,i,j}$. However, on the individual level, the results are quite different. Using the MNQ, only three individual hypotheses are rejected, whereas using the naive method, 29 are rejected and as it is expected Bonferroni as the most conservative method, rejects only 2 individual hypotheses. The test statistics for the rejected individual comparisons based on MNQ and naive methods are provided in Table 3.10. Bonferroni uses the same value of test statistics as MNQ with threshold 3.196.

Table 3.10: value of test statistics for the rejected individual null hypothesis from MNQ and naive method in analysis of kidney function

rejected H_0 in naive	MNQ	naive
$\beta_3 = \beta_4$	-5.858	-27.257
$\beta_3 = \beta_6$	-2.856	-14.137
$\beta_4 = \beta_7$	6.831	18.314
threshold	2.728	3.163

In order to explain the drastic difference between the MNQ and naive methods, recall that there is correlation between the repeated measurements across the time points. The interaction is shown to be very significant with $w^* = -0.49$. By ignoring this correlation, which is done by the naive method, the dependence between the points is underestimated and the standard error tends to be very liberal. On the other hand, the MNQ approach takes the inter-correlation into account properly.

3.2.2 Analysis of depression data

The Health and Retirement Study (HRS) dataset is used to show the application of the proposed approach. Information about health, financial situation, family structure and so on were collected by RAND center. Here the effect of some factors on depression during the elderly is studied. Depression is considered as the binary response variable (0 for no depression and 1 for depression). 7 factors are chosen as independent variables. Age (in month), smoke (0 for no and 1 for yes), restless sleep (0 for no and 1 for yes), diabetes, high blood pressure, frequent vigorous physical activity and difficulty in walking (0 for no and 1 for yes) are considered as covariates. For each individual we just considered the years that all the factors are recorded. So there was no missing in the data, but the number of repeated measurements varied across people. In this data set 33636 people have been measured in 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010 and 2012. As the response variable is binary, the quadratic exponential model is one natural choice to analyze this data that also easily allows us to perform multiple comparison when the clusters have different sizes. The effect of these 7 factors on depression was compared. Also we entered the augmented w parameter to account for the within-person correlations.

Here, we compare all pair-wise comparisons for the factors, $H_{0,i,j} = \{\beta_i = \beta_j\}$, for a total of 21 null hypotheses. We use the MNQ approach, MNQ "naive", and Bonferroni method. The hypothesis test based on the MNQ method rejected three hypotheses. Each of these test were also rejected by the MNQ naive method (and the three were the top 3 most significant), but this approach rejected in total 18 hypotheses out of the possible 21 and the

Table 3.11: estimate of the coefficients in analysis of depression data

	Estimate	Std. Error	P-value
sleepless	1.333	0.0156	$< 2e - 16$
diabetes	0.071	0.0146	$8.96e - 07$
smoke	0.2826	0.0200	$< 2e - 16$
age	0.0007	$5.964e - 05$	$< 2e - 16$
activity	-0.0156	0.0036	$2.35e - 05$
high blood pres.	0.0764	0.0114	$2.07e - 11$
difficulty in walking	0.0695	0.0054	$< 2e - 16$
w	0.2877	0.0023	$< 2e - 16$

conservative Bonferroni method fails to reject the null hypotheses.

In order to explain the drastic difference between the two methods, recall that there is correlation between the repeated measurements across the time points of size $\widehat{w}_n = 0.285$. By ignoring this correlation, as shown in the naive method, the dependence between the points is underestimated and the standard error tends to be very liberal, leading to false results.

Table 3.12: Results of MNQ, Bonferroni, and naive method in testing individual null hypotheses in analysis of depression data. A: fail to reject, R: rejectH_0

H_0	MNQ	naive	Bonf.	H_0	MNQ	naive	Bonf.
$\beta_{sleep} = \beta_{diabet}$	A	R	A	$\beta_{smoke} = \beta_{age}$	A	R	A
$\beta_{sleep} = \beta_{smoke}$	A	R	A	$\beta_{smoke} = \beta_{activity}$	A	R	A
$\beta_{sleep} = \beta_{age}$	R	R	A	$\beta_{smoke} = \beta_{hibp}$	A	R	A
$\beta_{sleep} = \beta_{activity}$	R	R	A	$\beta_{smoke} = \beta_{dif\ walk}$	A	R	A
$\beta_{sleep} = \beta_{hibp}$	A	R	A	$\beta_{age} = \beta_{activity}$	A	R	A
$\beta_{sleep} = \beta_{dif\ walk}$	R	R	A	$\beta_{age} = \beta_{hibp}$	A	R	A
$\beta_{diabet} = \beta_{smoke}$	A	R	A	$\beta_{age} = \beta_{dif\ walk}$	A	R	A
$\beta_{diabet} = \beta_{age}$	A	R	A	$\beta_{activity} = \beta_{hibp}$	A	R	A
$\beta_{diabet} = \beta_{activity}$	A	R	A	$\beta_{activity} = \beta_{dif\ walk}$	A	R	A
$\beta_{diabet} = \beta_{hibp}$	A	A	A	$\beta_{hibp} = \beta_{dif\ walk}$	A	A	A
$\beta_{diabet} = \beta_{dif\ walk}$	A	A	A				

3.3 Discussion

In many correlated multivariate models, it is often difficult to perform multiple comparisons based on the full likelihood. In this work, we construct the multiple comparison procedures based on the composite likelihood method to overcome this computational difficulty. Theory is developed based on the asymptotic properties of the composite likelihood test statistic. Then the simultaneous quantile of multivariate normal is used as a threshold for test statistics to handle larger errors in multiple comparisons. Therefore, we address issues of computational intensity and multiplicity. We illustrated the theory for four different models: multivariate normal, multivariate probit, quadratic exponential and gamma. The comparison between the proposed method and some well-known traditional approaches including Bonferroni, Dunn-Sidak, Holm, and Schéffe shows that the MNQ method, which is based on composite likelihood test statistics and uses multivariate normal quantiles to derive cut-off values for the test statistics, possesses a more acceptable family-wise type I error rate in most simulation settings, compared to the other test procedures.

Chapter 4

Asymptotic Distribution of Composite Likelihood Ratio Test in Non-standard Conditions

4.1 Introduction

Let $y = (y_1, \ldots, y_n)$ be a sample taken from a population with density function $f(y;\theta)$ of parameter space $\Theta \subseteq \mathbb{R}^k$. We wish to test

$$H_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad H_a : \theta \in \Theta_1,$$

where $\Theta = \Theta_0 \cup \Theta_1$. In every classical hypothesis testing procedure, there is a set of regularity conditions, consisting of two kinds of assumptions. Firstly, there are some probabilistic assumptions about the model $f(y;\theta)$. Secondly, there are some assumptions about the parameter space Θ, the space under the null hypothesis Θ_0, and the local geometry of the true parameter point θ_0. In likelihood ratio testing, the second set of assumptions is expressed as

- In a simple null hypothesis, it is assumed that Θ contains an open region ω and that θ_0 is an interior point of ω.

- In a composite hypothesis test, the parameter space Θ is linear and the null space Θ_0 is a linear sub-spaces of Θ [53]. In other words, the parameter θ_0 lies on an r-dimensional hyperplane Θ_0 of a k-dimensional space $\Theta \subseteq \mathbb{R}^k$, where $r = \dim(\Theta_0)$ [6].

Holding each of the regularity conditions is essential for the model approximation and if any of these assumptions fails, the known asymptotic results may no longer be valid. For instance, for testing a simple hypothesis, $H_0 : \theta = \theta_0$, if the true value of the parameter lies on the boundary of the parameter space Θ, the standard condition does not hold, since θ_0 is not an interior point of Θ. Also, it might be of interest to test if the parameter θ is on a subspace of $\mathbb{R}^k$, e.g. in the positive side of a k-dimensional Euclidean space, or inside the unit ball. In one-sided hypothesis, the null parameter space is not linear sub-space. If this happens, the standard theory may not be applicable. These situations that give rise to the limiting distributions other than the known classical one, are referred to as non-standard conditions.

To understand what is happening theoretically, let $\widehat{\theta}^c_{n,0}$ and $\widehat{\theta}^c_n$ denote the maximum composite likelihood estimators in Θ_0 and Θ under the null and alternative hypotheses, respectively, similar to section 1.2. Using Taylor's series the composite likelihood ratio test is written as

$$\begin{aligned}
\tilde{\lambda}_n(\theta) &= 2(\sup_{\theta\in\Theta}\sum_{i=1}^{n}\log CL(\theta,y_i) - \sup_{\theta\in\Theta_0}\sum_{i=1}^{n}\log CL(\theta,y_i)) \\
&= \sum_{i=1}^{n}\Big(cl(\widehat{\theta}^c_n,y_i) - cl(\widehat{\theta}^c_{n,0},y_i)\Big) \\
&= n(\widehat{\theta}^c_n - \widehat{\theta}^c_{n,0})^T g(\widehat{\theta}^c_n,y) + n(\widehat{\theta}^c_n - \widehat{\theta}^c_{n,0})^T H(\widehat{\theta}_n,y)(\widehat{\theta}^c_n - \widehat{\theta}^c_{n,0}) + o_p(1) \qquad (4.1)
\end{aligned}$$

where $g(\widehat{\theta}^c_n,y) = \frac{\partial cl(\theta,y)}{\partial\theta}$ and $H(\widehat{\theta}_n,y) = \frac{\partial^2 cl(\theta,y)}{\partial\theta^2}$ are first and second derivatives of $cl(\theta,y_i)$ at $\widehat{\theta}^c_n$. If the $\widehat{\theta}^c_n$ is an interior point of Θ, the first term vanishes and the second term converges to

$$\sqrt{n}(\widehat{\theta}^c_n - \widehat{\theta}_{n,0})^T H_{\theta_0}\sqrt{n}(\widehat{\theta}^c_n - \widehat{\theta}_{n,0}) + o_p(1)$$

which is, in fact, the transformed quadratic distance between the estimated values under each hypothesis. However, this can be quite different if the true parameter point lies on the boundary since the first term in (4.1) may not disappear and the limiting distribution would be affected.

In order to illustrate the issues, suppose the observation $y = (y_1,y_2)$ has a bivariate normal distribution with mean $\mu = (\mu_1,\mu_2)$ and variance $\Sigma = I$. In this case, the full likelihood ratio test statistic is

$$-2\log\Lambda(\mu) = \inf_{\mu\in\Theta_0} ||y-\mu||^2 - \inf_{\mu\in\Theta} ||y-\mu||^2 = Q_{\Theta_0} - Q_{\Theta}$$

which is the difference between the squared distance of a normal observation and null/ alternative parameter space. The following example from Chernoff [6] demonstrates a simple case of the non-standard situation. In this example, the true parameter is a boundary point of Θ and the distribution of the LRT becomes a mixture of chi-square distributions.

Example 4.1.1. Let $\Theta = \{(\mu_1, \mu_2) : \mu_1 \geq 0,\ \mu_2 \geq 0\}$ and $\Theta_0 = \{(\mu_1, \mu_2) : \mu_1 = \mu_2 = 0\}$, the origin alone. Consider the likelihood ratio test of $H_0 : \mu_1 = \mu_2 = 0$ versus $H_a : \mu_1 \geq 0 \mu_2 \geq 0$, then $Q_{\Theta_0} = y_1^2 + y_2^2$ and

	$y_1 \leq 0$	$y_1 \geq 0$
$y_2 \geq 0$	y_2^2	0
$y_2 \leq 0$	$y_1^2 + y_2^2$	y_1^2

Table 4.1: Q_Θ

	$y_1 \leq 0$	$y_1 \geq 0$
$y_2 \geq 0$	y_1^2	$y_1^2 + y_2^2$
$y_2 \leq 0$	0	y_2^2

Table 4.2: $-2\log\Lambda(\theta)$

Then

$$P(2\log\Lambda(\theta) \leq c) = \begin{cases} \frac{1}{4}P(\chi_2^2 \leq c) + \frac{1}{2}P(\chi_1^2 \leq c), & \text{if } c \geq 0. \\ 0, & \text{if } c < 0. \end{cases}$$

□

This work focuses on the composite likelihood ratio test on situations where the second set of the regularity assumptions are violated. In particular, the focus is on some non-standard conditions where the true parameter θ_0 lies on the boundary of the parameter space. We stablish the asymptotic properties of the composite likelihood estimator and limiting distribution of the composite likelihood ratio test when some parameters lie on the boundary.

In addition to the theoretical work, the limiting distribution of composite likelihood ratio tests is derived by partitioning the parameter space into smaller subsets called relative interior sets and projecting the observation points onto each set. The result can be applied to the full likelihood case too. However, the full likelihood ratio test in non-standard conditions has already been studied by several authors; existing theoretical results do not provide a direct method to derive the limiting distribution of the likelihood ratio test when the dimension of the parameters on the boundary is greater than four or the Hessian matrix is non-diagonal.

We assume that Θ_0 and Θ_1 follow Chernoff [6]'s assumption and can be approximated by polyhedral tangent cones. Then we expand Shapiro [42]'s approach to obtain a general form of the test statistic which is a weighted sum of the mixture of chi-square variables. We propose some algorithms to compute the elements of the test statistic. For a p dimensional parameter, the cone has 2^p faces. So the number of dimensions increases exponentially and this is the reason of difficulty of finding the distribution in higher dimensions.

In non-standard parametric problem, bootstrap is another method that seems useful. However, some authors such as Andrews [1] and Drton [10] show that bootstrap is inconsistent in estimating the limiting distribution. Andrews [1] suggests some variations of bootstrap that can be consistent, but a tuning parameters is required to be found, which is not easy. Drton [10] shows that bootstrap likelihood ratio test with the boundary problem is always anti-conservative.

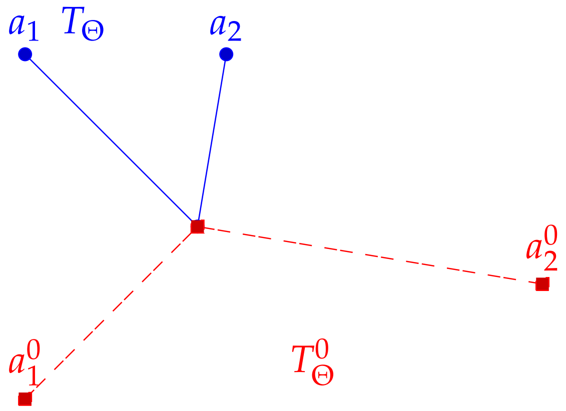


Figure 4.1: A 2-d polyhedral cone T_Θ and its polar cone T_Θ^0. That is $a_1 \perp a_1^0$ and $a_2 \perp a_2^0$

Section (4.2) covers a review on likelihood ratio test (LRT) and composite likelihood ratio test (CLRT) in standard conditions, as well as some geometrical and mathematical concepts, commonly applied in statistics and used in this chapter. The limiting distribution of likelihood ratio test is stated in section (4.3), done by Chernoff [6]. Then consistency and limiting distribution of composite likelihood ratio test is studied in section (4.4) which is developed in this thesis. In section (4.5) the parameter space assumed in this work is introduced and the method of approximating it by the tangent cone is described. In section (4.6), first a general form of limiting distribution of the likelihood ratio test, which is a mixture of weighted sum of chi-squares, is defined. This is a more complicated form of the chi-bar distribution and contains coefficients and weights that need to be estimated. In the next subsections, some algorithms for computing the weight and methods for finding other elements of the test statistic and quantile is discussed.

4.1.1 Full likelihood ratio test (LRT)

The likelihood ratio test (LRT) is a classical hypothesis test that is the most powerful hypothesis approach for testing simple hypotheses by the Neyman-Pearson lemma. For testing composite hypothesis, the generalized LRT is often uniformly most powerful test.

Let the parameter space Θ be an open set and $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^k$. The hypotheses of interest are $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_1$ $(\Theta = \Theta_0 \cup \Theta_1)$. Suppose the sample $y = (y_1, \ldots, y_n)$ is iid with density $f(y;\theta)$ and the likelihood function $L(\theta; y) = \prod_{i=1}^{n} f(y_i;\theta)$. Then the likelihood ratio test rejects the H_0 for the small values of

$$\Lambda_n = \frac{\sup_{\theta\in\Theta_0} L(\theta; y)}{\sup_{\theta\in\Theta} L(\theta; y)}.$$

Then the $-2\log$ likelihood ratio test statistic is

$$\lambda_n = -2\log\Lambda_n = -2\left(\sup_{\theta\in\Theta_0} L(\theta; y) - \sup_{\theta\in\Theta} L(\theta; y)\right).$$

Let $\widehat{\theta}_n$ and $\widehat{\theta}_{n,0}$ denote the maximum likelihood estimators in Θ and Θ_0, respectively. Under classical regularity conditions, the limiting distribution of Λ is determined. It is known that

1. For testing a simple hypothesis $H_0 : \theta = \theta_0$, the likelihood ratio statistic is $\Lambda_n = \frac{L(\theta_0; y)}{L(\widehat{\theta}_n; y)}$ and $\lambda_n(\theta) \to \chi^2_k$ as n grows to infinity.

2. To test the composite null hypothesis $H_0 : \theta \in \Theta_0$, where $\Theta_0 = \{\theta : A(\theta - b) = 0\}$ and A is a $r \times k$ matrix with rank r and $k \times 1$ vector b, the likelihood ratio statistic is

$\Lambda_n = \frac{L(\widehat{\theta}_{n,0};y)}{L(\widehat{\theta}_n;y)}$, and $\lambda_n \rightarrow \chi^2_r$.

3. More generally, suppose that $\Theta_0 = \{\theta : g = (g_1(\theta), ..., g_r(\theta))^T = 0\}$, where $g_i(\theta)$ is a continuously differentiable function from $\mathbb{R}^k \rightarrow \mathbb{R}$. It is shown that $\lambda_n \rightarrow \chi^2_r$.

In fact, the number of degrees of freedom is equal to the difference between the dimension of Θ and Θ_0.

Holding each of the regularity conditions is essential for chi-square approximation and if any of these assumptions fail, the limiting distribution may not be chi-square any more.

4.1.2 Composite likelihood ratio test (CLRT)

Let $f(y;\theta)$ be the density function of the random variables $y = (y_1, \ldots, y_n)$ with parameter $\theta \in \Theta \subseteq \mathbb{R}^k$ and the composite log-likelihood function is defined as $cl(\theta;y) = \log CL(\theta;y) = \sum_{s \in \mathbb{S}} \log f_s(y_s;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$, and $f_s(y_s;\theta)$ is a marginal or conditional density function corresponding to the subset s, and $\mathbb{S}$ is a set of indices. Let A_n be a vector of score function with elements $A_j = \frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(y_i;\theta)}{\partial \theta_j}$. From central limit theorem $\sqrt{n}A_n \longrightarrow N(0, J(\theta))$, where $J(\theta) = Var(\frac{\partial \log f(y_i;\theta)}{\partial \theta_j})$. Let $\widehat{\theta}^c_n$ be the maximum composite likelihood estimator (MCLE) in Θ. Then $\sqrt{n}(\widehat{\theta}^c_n - \theta) \longrightarrow N_k(0, G^{-1}(\theta))$, where $G(\theta) = H(\theta)J^{-1}(\theta)H(\theta)$ and where $H(\theta) = \lim_n E(-cl^{(2)}(\theta;y))/n$ and $J(\theta) = \lim_n \text{var}(cl^{(1)}(\theta;y))/n$. Here, $cl^{(1)}$ is the vector of first derivatives and $cl^{(2)}$ is the matrix of second order derivatives of $cl(\theta;y)$. Also

$$(\widehat{\theta}^c_n - \theta)G(\widehat{\theta}^c_n - \theta) \rightarrow \chi^2_k, \qquad nA_nJ^{-1}A_n \rightarrow \chi^2_k.$$

Let the true parameter θ_0 lie in a Euclidean space $\Theta \subseteq \mathbb{R}^k$ and $\Theta_0 \subseteq \Theta$. The composite likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_1$, is $\tilde{\Lambda}_n = \frac{\sup_{\theta \in \Theta_0} CL(\theta;y)}{\sup_{\theta \in \Theta} CL(\theta;y)}$. Then define

$$\tilde{\lambda}_n = -2 \log \tilde{\Lambda}_n = -2 \left(\sup_{\theta \in \Theta_0} cl(\theta; y) - \sup_{\theta \in \Theta} cl(\theta; y) \right)$$

The distribution of composite likelihood ratio statistic $\tilde{\lambda}_n$ under classical regularity condition is shown to converge to a mixture of chi-squared random variables with one degree of freedom, where the weight depends on the elements of the Godambe information matrix.

Theorem 4.1.1. *[56] Suppose that the true parameter θ_0 is a smooth interior point of the parameter space $\Theta \subseteq \mathbb{R}^k$. Under the regularity condition [reference], the composite likelihood ratio statistic $\tilde{\lambda}_n(\theta)$ is asymptotically distributed as $\sum_{i=1}^k \lambda_i V_i$, where V_i, $i = 1, \ldots, k$, are independent χ_1^2 and λ_i's, $i = 1, \ldots, k$ are the eigenvalues of $J(\theta)H^{-1}(\theta)$.*

4.1.3 Previous work

Likelihood ratio test under non-standard condition has been discussed by several authors. Under the assumption that the parameter spaces can be approximated by a cone, Chernoff [6] provided a representation of the limiting distribution of the likelihood ratio statistic when the true value of the parameter is on the boundary of the parameter space. Shapiro [42] showed that the distribution of a class of tests including likelihood ratio when the true parameter is a boundary point of Θ_0 and the space Θ is open, is asymptotically a mixture of chi-square random variables, which is referred to a chi-bar-squared statistic that is a mixture of chi-squares. He also proposed a method to find the weights corresponding to chi-bar for the maximum of four dimensions.

Self and Liang [41] studied distribution of likelihood ratio test with boundary problem in Θ_0 and Θ for higher dimension parameters with diagonal fisher information matrix. In addition to the existence and consistency of maximum likelihood estimator for the large sample distribution, they followed Chernoff (1954) to show that the limiting distribution of the maximum likelihood estimator is the same as the distribution of the projection of the Gaussian random variable onto the region of possible values for the parameter. Chen and Liang [5] studied the asymptotic distribution of a pseudo-likelihood ratio test statistics. They used the pseudo-likelihood approach, studied by Gong and Samaniego (1981), for testing the parameter of interest θ in presence of a nuisance parameter ϕ when ϕ is estimated by a method other than maximum likelihood. They studied the cases that the parameter of interest or the nuisance parameter lies on the boundary of the two-dimensional parameter space for the total of dimension two. These situations usually lead to mixtures of chi-square distributions with a sandwich type covariance structure.

Following Chernoff's work, Drton [9] introduced a situation that Chernoff's regularity assumption always hold. That is when the parameter space under the null hypothesis can be defined by a finite union of polynomial equalities and inequalities, so called a semi-algebraic set. The boundary problem is not the only non-standard condition that might raise. Drton [9] used tools from algebraic geometry to study the asymptotic distribution of likelihood ratio test when the parameter is a singularity point. Assume that $\Theta_0 = \{\theta \in \mathbb{R}^2 | \theta_2^2 = \theta_1^3 + \theta_1^2\}$. This space is a curve with a self-intersection at $\theta = 0$

(see Drton [9] example 1.1). It can be seen that the distribution of likelihood ratio test when the true parameter is zero converges to the minimum of two chi-squared random variables with one degree of freedom. Such points (such as self-intersections and cusps), while possibly interior, are not smooth and give rise to a non-standard condition and are called singularities. Drton [9] used tools from algebraic geometry to study the asymptotic distribution of likelihood ratio test when the parameter is a singularity.

Recently, Susko [46] developed an approach as an alternative to the chi-bar statistic for independent data with interior nuisance parameter. He did this by conditioning on the number of parameters that are interior of parameter space.

4.2 Some useful background

4.2.1 Cone

A *cone* $C(\theta)$ with vertex at θ is the set of the vectors such that if $x \in C(\theta)$ then $a(x-\theta)+\theta \in C(\theta)$, where a is a non-negative real number. Let $\mathbf{a}_1, \ldots, \mathbf{a}_r$ be points in $\mathbb{R}^r$. The cone is *finitely generated* by $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ if it can be written as $C = \{x_1\mathbf{a}_1 + \ldots + x_r\mathbf{a}_r : x_i \geq 0, i = 1, \ldots, r\}$. Let $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ be a $k \times r$ matrix that each column is one of the points, then the cone C can be defined as $C = \{A\mathbf{x} : \mathbf{x} \geq 0\}$. The cone C is called *tight*, if it can not be generated by a sub-matrix of columns of A.

A *polyhedral cone* P is a set of vectors $P = \{\mathbf{x} \in \mathbb{R}^k : A^T\mathbf{x} \geq 0\}$, where $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ is a $k \times r$ $(r \leq k)$ matrix with rank r. P is a closed convex cone, that is the intersection of half-spaces $\{\mathbf{x} : \mathbf{a}_1^T\mathbf{x} \geq 0\}$,. . ., and $\{\mathbf{x} : \mathbf{a}_r^T\mathbf{x} \geq 0\}$.

Theorem 4.2.1. *(Mikowski-Weyl's theorem): A cone is polyhedral if and only if it is finitely generated.*

The orthogonal space to the cone C is called the polar (or negative dual) cone, C^0. A *polar cone* of the cone C, is the set of vectors $C^0 = \{\mathbf{y} \in \mathbb{R}^k : \mathbf{a}_i^T \mathbf{y} \leq 0, i = 1, \ldots, r\} = \{\mathbf{y} : A^T \mathbf{y} \leq 0\}$. The polar polyhedral cone for the polyhedral cone is $\mathbb{P}^0 = \{A\mathbf{y} : \mathbf{y} \leq 0\}$, a space spanned by the columns of A.

4.2.2 Projection onto the cone

Let $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ be $q \times k$, $r - q \times k$ and $k - r \times k$ real valued matrices. Consider the set $\mathbb{P} = \mathbb{P}_1 \cap \mathbb{P}_2 \cap \mathbb{P}_3$ where

$$\begin{aligned}
\mathbb{P}_1 &= \{\theta \in \mathbb{R}^k : \mathcal{A}_1\theta \geq 0\} = \cap_{i=1}^{q}\{\theta \in \mathbb{R}^k : a_i^T\theta \geq 0\} \\
\mathbb{P}_2 &= \{\theta \in \mathbb{R}^k : \mathcal{A}_2\theta = 0\} = \cap_{i=q+1}^{r}\{\theta \in \mathbb{R}^k : a_i^T\theta = 0\} \qquad (4.2) \\
\mathbb{P}_3 &= \{\theta \in \mathbb{R}^k : \mathcal{A}_3\theta \in \mathbb{R}^{k-r}\} = \{\theta_{r+1} \in \mathbb{R}\} \times \{\theta_{r+2} \in \mathbb{R}\} \times \cdots \times \{\theta_k \in \mathbb{R}\},
\end{aligned}$$

where $\mathcal{A}_3$ is the matrix equal to zero, except the right hand $(k - r) \times (k - r)$ submatrix is the identity matrix I_{k-r}. The matrix $\mathcal{A} = [\mathcal{A}_1^T \;\; \mathcal{A}_2^T]$ has rank r.

For any $\mathcal{I} \subseteq \{1, \ldots, q\}$ we define the face of the polyhedral cone $\mathbb{P}_1$ as

$$F_{\mathcal{I}} = \left\{\cap_{i\in\mathcal{I}}\{\theta : a_i^T\theta = 0\}\right\} \cap \left\{\cap_{i\in\mathcal{I}^c}\{\theta : a_i^T\theta \geq 0\}\right\}.$$

The dimension of the face is given by the size of the set $\mathcal{I}^c$ denoting by $|\mathcal{I}^c|$. If $\mathcal{I}$ is the empty set we recover the $\mathbb{P}_1$ itself, while if $\mathcal{I} = \{1, \ldots, q\}$ we obtain a single vertex. Let $\mathcal{F} = \{F_{\mathcal{I}}, \mathcal{I} \in 2^{\{1,\ldots,q\}}\}$ denote the set of faces of $\mathbb{P}_1$. Here, $2^{\{1,\ldots,q\}}$ denotes the power set of $\{1, \ldots, q\}$. Therefore, the number of possible faces is 2^q. For each face, we also define the relative interior of the face as

$$\mathrm{ri}\,(F_{\mathcal{I}}) \;=\; \left\{\cap_{i\in\mathcal{I}}\{\theta : a_i^T\theta = 0\}\right\} \cap \left\{\cap_{i\in\mathcal{I}^c}\{\theta : a_i^T\theta > 0\}\right\}.$$

The collection of relative interiors partitions the polyhedral cone $\mathbb{P}_1$ into 2^q disjoint spaces. Except for the vertex, that is $\mathrm{ri}(F_{\{1,\ldots,q\}}) = F_{\{1,\ldots,q\}}$, the rest of these sets are open in a linear subspace and hence there is no boundary issue and the standard results can be applied to each part separately.

Example 4.2.1. Let $A = \begin{pmatrix} 2 & 4 \\ -1 & 1 \end{pmatrix}$ and P be a polyhedral cone generated by A. That is $P = \{\mathbf{x} : 2x_1 - x_2 \geq 0, 4x_1 + x_2 \geq 0\}$. The faces of P are

$$F_1 = \{\mathbf{x} : 2x_1 - x_2 \geq 0, 4x_1 + x_2 \geq 0\} \quad , \qquad F_2 = \{\mathbf{x} : 2x_1 - x_2 = 0, 4x_1 + x_2 \geq 0\},$$
$$F_3 = \{\mathbf{x} : 2x_1 - x_2 \geq 0, 4x_1 + x_2 = 0\} \quad , \qquad F_4 = \{\mathbf{x} : 2x_1 - x_2 = 0, 4x_1 + x_2 = 0\}.$$

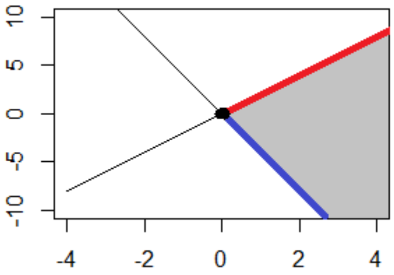


Figure 4.2: Four different relative interior sets of the cone P. gray: $ri(F_1)$, red: $ri(F_2)$, blue: $ri(F_3)$, black: $ri(F_4)$.

The relative interiors correspond to each face are described below and also can be seen in Figure 4.2.

$$
\begin{aligned}
ri(F_1) &= \text{The cone } P \text{ excluding the boundary,} \\
ri(F_2) &= \text{the upper boundary of } P \text{ excluding the origin,} \\
ri(F_3) &= \text{the lower boundary of } P \text{ excluding the origin,} \\
ri(F_4) &= \text{the origin}
\end{aligned}
$$

□

For any point $y \in \mathbb{R}^k$, we define the projection of y onto $\mathbb{P}$, $\Pi(y|\mathbb{P})$, as the point satisfying

$$\inf_{\theta \in \mathbb{P}} \|y - \theta\| = \|y - \Pi(y|\mathbb{P})\|. \tag{4.3}$$

That is, $\Pi(y|\mathbb{P})$ is the closest point in $\mathbb{P}$ to y. Due to the convex structure of the cone, the point $\Pi(y|\mathbb{P})$ is unique. There exist $\mathcal{I} \in 2^{\{1,\ldots,q\}}$ such that $\Pi(y|\mathbb{P}) = \Pi(y|\mathbb{P}_1 \cap \mathbb{P}_2 \cap \mathbb{P}_3) = \Pi(y|\operatorname{ri}(F_{\mathcal{I}}) \cap \mathbb{P}_2 \cap \mathbb{P}_3)$. This is because the relative interior sets partition the cone $\mathbb{P}_1$. Note that each element y_l in $\mathbb{P}_3$ belong to $\mathbb{R}$, therefore $\Pi(y_l|\mathbb{P}_3) = y_l$.

Proposition 4.2.2. *Let $\mathbb{P} = \mathbb{P}_1 \cap \mathbb{P}_2 \cap \mathbb{P}_3$ be as defined earlier. Let $\mathcal{A}_{1,\mathcal{I}} = [\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}]$, where $\mathcal{I} = \{i_1, \ldots, i_m\}$ and $\mathcal{A}_{\mathcal{I}}$ is the $r - q + m \times r$ upper matrix of $[\mathcal{A}_{1,\mathcal{I}}^T \;\; \mathcal{A}_2^T]$. Then, $\Pi(y|\mathbb{P})$ is unique and*

$$\|y - \Pi(y|\mathbb{P})\|^2 \;=\; \sum_{\mathcal{I} \in 2^{\{1,\ldots,q\}}} y^T \mathcal{Q}_{\mathcal{I}}\, y \;\; \mathbb{I}\Big(\Pi(y|\mathbb{P}_1 \cap \mathbb{P}_2) \in \operatorname{ri}(F_{\mathcal{I}}) \cap \mathbb{P}_2\Big).$$

where $\mathcal{Q}_{\mathcal{I}} = A_{\mathcal{I}}^T(A_{\mathcal{I}}A_{\mathcal{I}}^T)^{-1}A_{\mathcal{I}}$ and $y = (y_1, \ldots, y_r) \in \mathbb{R}^r$ denotes the subvector of the first r elements of y.

A proof of

$$\Pi(y|\operatorname{ri}(F_{\mathcal{I}})) \;=\; A_{\mathcal{I}}^T(A_{\mathcal{I}}A_{\mathcal{I}}^T)^{-1}A_{\mathcal{I}}y. \tag{4.4}$$

is provided in the Appendix (D).

4.2.3 Tangent approximation

When θ is an interior point of Θ, a tangent space can be defined in an open neighborhood of θ. *Tangent space* is defined as the first order linear approximation of Θ around θ. When θ lies on the boundary of Θ, the tangent space can not be defined and the concept of tangent space is replaced by the tangent cone. Chernoff [6] defined a tangent approximation to Θ

in a neighborhood of the parameter point θ.

Definition 4.2.1. The set $\Theta \subseteq \mathbb{R}^k$ is approximated at point θ by a cone $C_\Theta(\theta)$, if

$$d(C_\Theta, y) = o(||y - \theta||), \text{ for all } y \in \Theta,$$

and

$$d(\Theta, x) = o(||x - \theta||), \text{ for all } x \in C_\Theta,$$

where $d(\Phi, x) = inf_{y \in \Phi}||x - y||$, the distance between point x and its projection on Θ.

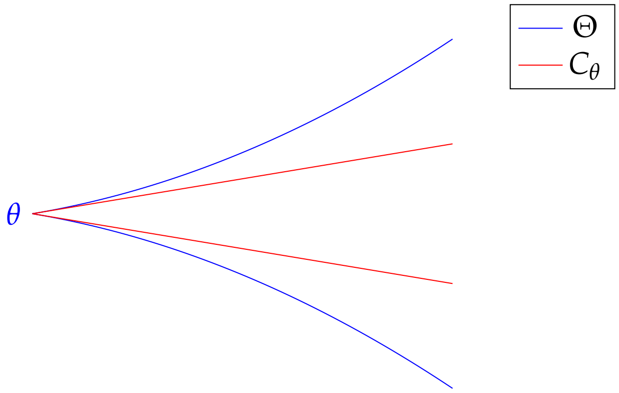


Figure 4.3: Approximating cone based on the Chernoff's assumption

Figure 4.3 demonstrates the approximating cone C_Θ of the parameter space Θ. Chernoff's approximation can be applied to many different forms of parameter spaces.

Tangent cone is the other concept that can describe the geometry of the θ in the parameter space. A *tangent cone* $T_\Theta(\theta)$ is the set of all the directions from which sequences in Θ converge to θ. Assume there exist a sequence $\{\theta_n\}$ in Θ converging to θ, and a sequence of positive real numbers $\{a_n\}$. Then $lim_{n\to\infty}a_n(\theta_n - \theta)$ is a tangent vector. The set of all tangent vectors form a *tangent cone* $T_\Theta(\theta)$ of Θ at the point θ.

So the tangent cone is a cone with vertex at the origin and can be defined by the limit of $a_n^{-1}(\Theta_n - \theta_0)$, when $n \to \infty$.

Next definition is related to directional derivative, is called Chernoff-regularity .

Definition 4.2.2. The parameter space $\Theta \subseteq \mathbb{R}^k$ is *Chernoff-regular* at the point θ if for every vector τ in the tangent cone $T_\Theta(\theta)$ and a sequence $\{a_n\}$ converging to zero, there exists a sequence $\{\theta_n\}$ converging to θ in Θ such that $\tau = \lim_{n\to\infty} a_n^{-1}(\theta_n - \theta)$.

Intuitively, it says each vector τ in the tangent cone $T_\Theta(\theta)$ corresponds to a smooth curve in Θ starting from θ, with slope parallel to τ at θ. For example, the set $\{\theta \in \mathbb{R}^2 : \theta_2 = \theta_1 \sin(\theta_1^{-1})\}$ is not Chernoff regular at $\theta_0 = 0$, although θ_0 is an interior point of Θ,.

Geyer [15], Theorem 2.1, shows that Θ is approximated by a cone at θ_0 if and only if Θ is Chernoff-regular.

Therefore the tangent cone, $T_\Theta(\theta)$, can be defined by the limit of $a_n^{-1}(\Theta_n - \theta_0)$, when $n \to \infty$. As we work with the sequence of full and composite likelihood estimators and it is shown that both are $n^{1/2}-$consistent estimator of θ_0, then $a_n = n^{-1/2}$ seems a proper choice and the set$\sqrt{n}(\Theta - \theta_0)$ converges to $T_\Theta(\theta_0)$.

As an specific case, assume the parameter space Θ can be shown as the Cartesian product $\Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_k$, For a simple hypothesis test, if the j^{th} coordinate of θ_0 lies on the boundary, then $\lim_{n\to\infty} \sqrt{n}(\Theta_j - \theta_{0j}) = \{0\}$. In general, let $\Theta_j = [a_j, b_j]$, $j = 1, \ldots, k$ which is any interval in $\mathbb{R}$, (an open, a close or a half interval), then

true value of θ_{0j} is on	$\Theta_j - \theta_{0j}$	$\lim_{n\to\infty} \sqrt{n}(\Theta_j - \theta_{0j})$
$c \in (a_j, b_j)$(interior)	$[a_j - c, b_j - c]$	$\mathbb{R}$
a_j (boundary)	$[0, b_j - a_j]$	$[0, \infty)$
b_j (boundary)	$[a_j - b_j, 0]$	$(-\infty, 0]$

It can be seen that if $\Theta_j,\ j = 1, \ldots, k$ corresponds to any nonnull subset of $\mathbb{R}$ other than the whole real line in the tangent cone, then its corresponding parameter may lay on the boundary.

4.3 LRT under non-standard conditions

The limiting distribution of likelihood ratio statistic, $\lambda_n(\theta)$ is examined by Chernoff [6] when the value of parameter is a boundary point of both parameter spaces corresponding to the null and alternative hypotheses.

Theorem 4.3.1. *[6] Under the regularity condition in Chernoff [6] , assume $\theta_0 \in \Theta_0 \subseteq \Theta \subseteq \mathbb{R}^k$ is the true parameter point at which the sets Θ and Θ_0 are Chernoff-regular. Let $z \sim N_k(0, I^{-1}(\theta_0))$ that $I(\theta)$ is the Fisher information matrix and the maximum likelihood estimator $\widehat{\theta}_n$ is consistent. Then the asymptotic distribution of the likelihood ratio statistic is the same as the distribution of*

$$\bar{\chi}^2 = \min_{\theta \in T_{\Theta_0(\theta_0)}} (z-\theta)^T I(\theta_0)(z-\theta) - \min_{\theta \in T_{\Theta(\theta_0)}} (z-\theta)^T I(\theta_0)(z-\theta) \tag{4.5}$$

which $T_{\Theta(\theta)}$ is the tangent cone of the set Θ at the point θ.

4.4 CLRT under non-standard conditions

Let $f(y;\theta)$ be the density function with parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. We wish to test $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$. When θ lies on the boundary, the score function at the composite likelihood estimation evaluated at $\widehat{\theta}_n^c$, may not be zero and consequently, the limiting distribution of the $\widehat{\theta}_n^c$ may not be normal.

Here, under Chernoff's regularity, the asymptotic properties of the composite likelihood ratio test statistic is studied. First, the classical regularity conditions is modified for the composite likelihood estimation in non-standard cases.

4.4.1 Regularity conditions

Let $y \subseteq \{y_1, \dots, y_n\}$ be the vector of observations with the density $f(y;\theta)$. Let $N_\delta(\theta_0)$ denote a neighbourhood around the point θ_0 and $cl(N_\delta(\theta_0))$ is a closure of $N_\delta(\theta_0)$. Then $f(y;\theta)$ should satisfy

(B1). The marginal density function of y, $f(y;\theta)$ is distinct for different values of y, i.e. if $\theta_1 \neq \theta_2$ then $P(f(y;\theta) \neq f(y;\theta)) > 0$, and θ is in the parameter space Θ.

(B2). The marginal densities of y have common support for all θ.

(B3). The marginal density $\log f$ is three times continuously differentiable in $\theta \in cl(N_\delta(\theta_0))$. Moreover, there exists an integrable function $M(y)$ such that

$$|(\partial\alpha\partial\theta_i \log f)(y;\theta)| \leq M(y),$$

for $i = 1, \cdots, k$. And for $\alpha = 3$ there is a constant B (independent of θ) such that $E\{M(y)\} < B$.

(B4). If $\theta \in N_\delta(\theta_0)$, $J(\theta_0)$ is well-defined (i.e. exists and is finite) and invertible.

(B5). If $\theta \in N_\delta(\theta_0)$, $H(\theta_0)$ is well-defined (i.e. exists and is finite) and (strictly) positive-definite.

4.4.2 Asymptotic behaviour of CLRT in non-standard condition

In this section, the main results about the asymptotic properties of the composite likelihood ratio test are discussed. The first lemma shows an expansion for the composite likelihood function. Then the root-n consistency of the composite maximum likelihood estimator when θ_0 lies on the boundary of Θ is discussed. For this result instead of assuming that θ_0 is an interior point, we require a closed set around θ_0 in Θ. Finally, the limiting distribution of composite likelihood ratio test statistic is concluded in a theorem. Proof of the lemmas and theorems are given in the appendix (C).

Lemma 4.4.1. *Under the regularity conditions $(B1) - (B5)$, when intersection of Θ and a closure of a neighbourhood around θ_0 is a closed set ϕ, there exists a sequence θ_n in ϕ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\log f_{\theta_n} - \log f_{\theta_0}\right) = (\theta_n - \theta_0)^T A_{n,\theta_0} - \frac{1}{2}(\theta_n - \theta_0)^T H(\theta_n - \theta_0) + o_p(|\theta_n - \theta_0|^2) \quad (4.6)$$

The next lemma shows that the root-n consistency of the maximum composite likelihood estimator is not affected by the location of the parameter.

Lemma 4.4.2. *Under the regularity conditions $(B1) - (B5)$, when θ_0 is a limiting point of Θ and the intersection of Θ and a closure of a neighbourhood around θ_0 is a closed set $\phi \subseteq \mathbb{R}^k$, then there exists a sequence $\widehat{\theta}_n^c$ in ϕ which is a consistent maximum composite likelihood estimator of θ_0 and $\sqrt{n}(\widehat{\theta}_n^c - \theta_0) = O_p(1)$.*

Proposition 4.4.3. *Under the Chernoff regularity, let $\widehat{\theta}_n^c$ be a consistent estimator of θ_0 in ϕ. Then,*

$$\widehat{\theta}_n^c \quad = \quad \theta_0 + H^{-1}(\theta_0) A_{n,\theta_0} + o_p(1/\sqrt{n})$$

Remark 1. The statement (4.7) is true if the sequence $\widehat{\theta}_n^c$ lies in Θ. Without Chernoff's assumption, if θ_0 lies on the boundary of Θ, $\sum_{i=1}^n \frac{\partial \log f(\widehat{\theta}_n^c, y_i)}{\partial \theta}$ may not be zero which is required in the proof. Because $\widehat{\theta}_n^c$ converges to θ_0, but it might converge from outside of Θ. Consequently, the limiting distribution of the composite likelihood ratio test might be different. This is why modification in the regularity condition seems necessary. By imposing Chernoff's regularity, the sequence $\widehat{\theta}_n^c$ converges to θ_0 in a closed set around θ_0 in Θ.

Now a representation for the limiting distribution of the composite likelihood ratio test can be derived. Suppose $\widehat{\theta}_n^c$ is the sequence of local MCLE in Θ. It is shown that the sequence of standardized local maximum composite likelihood estimators $\sqrt{n}(\widehat{\theta}_n^c - \theta_0)$ is bounded in probability. Define $\widehat{h}_n = \sqrt{n}(\widehat{\theta}_n^c - \theta_0)$, that converges in distribution to $\theta < \infty$, a vector in the tangent cone at θ_0. From definition of the tangent vector, it is seen that the limit of $\sqrt{n}(\widehat{\theta}_n^c - \theta_0)$ when $n \to \infty$ is tangent to θ at θ_0. The parameter space Θ contains all the sequences that converge to θ_0. Consider the local parameter space $\sqrt{n}(\Theta - \theta_0)$. Then

Chernoff's regularity holds, if the set $\sqrt{n}(\Theta - \theta_0)$ converges to a tangent cone at θ_0.

Theorem 4.4.4. *Assume the regularity assumptions $(B1) - (B5)$ are satisfied and $\theta_0 \in \mathbb{R}^k$ be the true parameter point at which the parameter spaces Θ_0 and $\Theta \in \mathbb{R}^k$ are Chernoff regular. Assume that $\widehat{\theta}_{n,0}$ and $\widehat{\theta}_n$ are the consistent composite maximum likelihood estimators of θ_0 in Θ_0 and Θ, respectively. Letz be a multivariate normal random variable with mean zero and covariance matrix $G^{-1}(\theta_0) = H^{-1}(\theta_0)J(\theta_0)H^{-1}(\theta_0)$. Then the limiting distribution of likelihood ratio test statistic $\tilde{\lambda}_n$ is the same as distribution of*

$$\bar{\chi}^2 = \inf_{\theta \in T_{\Theta_0}(\theta_0)} (z-\theta)^T H(z-\theta) - \inf_{\theta \in T_{\Theta}(\theta_0)} (z-\theta)^T H(z-\theta) \tag{4.7}$$

which $T_\Theta(\theta_0)$ is the tangent cone of the set Θ at the point θ_0.

which gives the subtraction of two squared Mahalanobis distance when the covariance matrix of y is misspecified as H^{-1}.

Lemma 4.4.5. *Under the condition of the theorem 4.4.4, $\sqrt{n}|\widehat{\theta}^c_n - \tilde{\theta}_n| = o_p(1)$, where*

$$\tilde{\theta}_n = argmax_{\theta \in \phi} \left(\sqrt{n}(\theta-\theta_0)^T - H^{-1}J^{1/2}z\right)^T H \left(\sqrt{n}(\theta-\theta_0)^T - H^{-1}J^{1/2}z\right).$$

That is, $\tilde{\theta}_n$ is asymptotically equivalent to $\widehat{\theta}^c_n$.

Hence, in order to estimate the limiting distribution of composite likelihood ratio we need to compute $\inf_{\theta \in T_\Theta(\theta)} (z-\theta)^T H(z-\theta)$ under the null and alternative hypothesis.

4.5 Parameter space

Here, we introduce the space that we assume the parameters lie onto in this work. Assume $\theta_0^T = (\theta_{01}, \theta_{02}, \ldots, \theta_{0k})^T$ be the true parameter point. The k-dimensional parameter vector θ is decomposed into four parts. The first two parts denote the elements that are to be tested, that one of them represents the parameters which may lie on the boundary and the other part contains the interiors. The last two parts are the elements that are not of interest in hypothesis testing, yet one part is for the boundary parameters and the last one for the parameters which are located in the interior of the parameter space.

Here, we consider the rather general form of the parameter space.

Let Ω denote an open subset of $\mathbb{R}^k$ with $0 \leq \kappa_1 \leq \kappa_2 \leq k$. Assume that

$$\Theta \;=\; \{\theta \in \Omega : g_i(\theta) \geq 0 \; i = 1, \ldots, \kappa_1, \;\; g_i(\theta) = 0 \; i = \kappa_1 + 1, \ldots, \kappa_2\}, \tag{4.8}$$

where the functions $g_i : \mathbb{R}^k \to \mathbb{R}$ are assumed to be continuously differentiable ($i = 1, \ldots, \kappa_2$). For $i = 1, \ldots, \kappa_2$, let $a_i = \{\partial_{\theta_1} g_i(\theta_0), \ldots, \partial_{\theta_k} g_i(\theta_0)\}^T$ denote the gradient vector at θ_0. Next, we relabel the indices $\{1, \ldots, \kappa_2\}$: Set $\{1, \ldots, q\} = \{i : g_i(\theta_0) \geq 0, 1 \leq i \leq \kappa_1\}$, and $\{q+1, \ldots, r\} = \{\kappa_1 + 1, \ldots, \kappa_2\}$.

The next proposition from Silvapulle, M.J., Sen, P.K. [45] shows that the tangent cone of the defined Θ is a polyhedral cone under some assumptions.

Proposition 4.5.1. *[45] Assume Θ is defined similar to (4.8). Let $\mathcal{A}^T = [a_1^T \; \ldots \; a_r^T]$ be the $r \times k$ Jacobian matrix at θ_0. Assume that*

- *$a_i, i = q+1, \ldots, r$ are linearly independent,*

  also, there exist a nonzero vector $\boldsymbol{b} \in \mathbb{R}^k$ such that

- *$a_i^T \boldsymbol{b} \geq 0, i = 1, \ldots, q$,*
- *$a_i^T \boldsymbol{b} = 0, i \in i = q+1, \ldots, r$.*

Then the tangent cone is equal to

$$T_\Theta(\theta) \quad = \quad \Big\{\theta \in \mathbb{R}^k : a_i^T\theta \geq 0 \; i = 1 \ldots, q, \;\; a_i^T\theta = 0 \; i = q+1, \ldots, r\Big\}. \tag{4.9}$$

The result suggests that the tangent cone is obtained by the first-order linear approximation at θ_0. Let $S = \{s : \; g_s(\theta_0) > 0, 1 \leq s \leq q\}$. If $S = \{1, \ldots, r\}$, then θ_0 is an interior point of Θ and its corresponding tangent space is $\mathbb{R}^k$. Also, (4.9) suggests that $g_s(\theta), \;\; s \in S$ does not have a role in approximation and constructing the tangent cone $T_\Theta(\theta_0)$.

Let $\mathcal{A} = [\mathcal{A}_1^T, \mathcal{A}_2^T]$ be a $k \times r$ matrix such that $T_\Theta(\theta_0) = \mathbb{P} = \mathbb{P}_1 \cap \mathbb{P}_2$, where

$$\begin{aligned}
\mathbb{P}_1 \quad &= \quad \{\theta \in \mathbb{R}^k : \mathcal{A}_1\theta \geq 0\} = \cap_{i=1}^q \{\theta \in \mathbb{R}^k : a_i^T\theta \geq 0\} \\
\mathbb{P}_2 \quad &= \quad \{\theta \in \mathbb{R}^k : \; \mathcal{A}_2\theta = 0\} = \cap_{i=q+1}^r \{\theta \in \mathbb{R}^k : a_i^T\theta = 0\}
\end{aligned}$$

are polyhedral cones and $\mathcal{A}_1$ is a $q \times k$ matrix with rank q, and $\mathcal{A}_2$ is a $(r-q) \times k$ matrix with rank $r - q$. The cone $\mathbb{P}_1$ represent the boundary parameters which are not of interest in hypothesis testing under H_0 or the parameters of interest that lie on the boundary under

H_a. The cone $\mathbb{P}_2$ represents the parameters of interest under H_0. It is shown that the cone

$$\mathbb{P}_3 = \{\theta_{r+1} \in \mathbb{R}\} \times \{\theta_{r+2} \in \mathbb{R}\} \times \cdots \times \{\theta_k \in \mathbb{R}\}$$

that represent the interior parameters does not have an impact on the overall tangent cone $T_\Theta(\theta)$. The cone $T_\Theta(\theta)$ is the linear space spanned by the columns of $\mathcal{A}$ and it can be written as

$$T_\Theta(\theta) = \{\theta \in \mathbb{R}^k : \mathcal{A}^T\theta \geq 0\}$$

As $T_\Theta(\theta)$ is a polyhedral cone and by Minkowski-Weyl's theorem (4.2.1), is finitely generated. Also we assume that the set of constraints defining $T_\Theta(\theta)$ is tight, so there is not a sub-matrix of $\mathcal{A}$ that generates the same cone.

Therefore, under the assumptions of proposition (4.5.1), the linear and non-linear parameter spaces can be well-approximated by a polyhedral cone.

Remark 2. If $g_i(\theta) = \theta_i$, $i = 1, \ldots, k$, and

$$\Theta_0 \quad = \quad \left\{\theta \in \mathbb{R}^k : \theta_i \geq 0 \; i = 1, \ldots, q, \;\; \theta_i = 0 \; i = q+1, \ldots, r\right\}$$

the parameter space Θ can be shown as $\Theta_1 \times \Theta_2 \times \ldots \times \Theta_k$, product of intervals of real line $\mathbb{R}$. Under the null hypothesis $\Theta_0 = \{0\}^{r-q} \times [0, \infty)^q \times \mathbb{R}^{k-r}$. The term $\{0\}^{r-q}$ contains both kinds of interior and boundary parameters of interest under the null hypothesis. Then $\Theta = [0, \infty)^p \times \mathbb{R}^{r-p-q} \times [0, \infty)^q \times \mathbb{R}^{k-r}$, where p is the number of parameters of interest which may lie on the boundary.

Then, computing the Jacobian matrix the general form of the tangent cones are $T_{\Theta_0}(\theta) = \{0\}^{r-q} \times [0,\infty)^q$ and $T_\Theta(\theta) = [0,\infty)^p \times [0,\infty)^q$.

From now on, we assume the parameter spaces are defined by continuously differentiable functions that satisfy the assumption of the proposition (4.5.1).

4.6 Methodology

In this section, we compute the distribution of $\bar{\chi}^2$. It is seen that $\tilde{\lambda} = 0$ when $\theta \in \mathbb{R}$ and the cone $\mathbb{P}_3$ is related to the interior areas of the parameter space and doesd not have an effect on constructing the tangent cone. Therefore, to reduce the dimension from now on we only consider the parameters which may lie on the boundaries and locates in the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ and their corresponding tangent cone can be built by $\mathbb{P}_1$ and $\mathbb{P}_2$. We consider $\theta \in \mathbb{R}^r$ the first r elements of θ, and so only consider the the $r \times r$ upper left sub-matrix of Σ, H and J. Similarly, let $\mathcal{A}$ be is the $r \times r$ upper matrix of $[\mathcal{A}_1^T \;\; \mathcal{A}_2^T]$. However, in general $\mathcal{A}$ is not a square matrix and is a $r \times m$ matrix with rank m, for m constraints imposed on the r parameters

Let $z \sim N_r(0, \Sigma)$, from theorem (4.3.1) and theorem (4.4.4) it is seen that full and composite likelihood ratio test statistic is represented in terms of Mahalanobis distances, $\inf_{\theta \in T_\Theta(\theta)} (z-\theta)^T \Sigma (z-\theta)$ where Σ in full likelihood is estimated by the Fisher information matrix while in composite likelihood setting the covariance is misspecified as the Hessian matrix H.

The matrix Σ is symmetric and positive definite. Using the matrix factorization, such as

Cholesky, define

$$\Sigma^{-1} = U^T U, \tag{4.10}$$

Hence the equations (4.5) and (4.7) can be represented as

$$\begin{aligned} \bar{\chi}^2 &= \inf_{\theta \in T_{\Theta_0}} (Uz - U\theta)^T (Uz - U\theta) - \inf_{\theta \in T_\Theta} (Uz - U\theta)^T (Uz - U\theta) & (4.11) \\ &= \inf_{\theta \in \tilde{T}_{\Theta_0}} ||\tilde{z} - \tilde{\theta}||^2 - \inf_{\theta \in \tilde{T}_\Theta} ||\tilde{z} - \tilde{\theta}||^2 & (4.12) \end{aligned}$$

where $\tilde{z} = Uz$ and $\tilde{T}_\Theta = \{\tilde{\theta} : \tilde{\theta} = Uh, \text{ for any } \theta \in T_\Theta(\theta)\}$. Then, $\inf_{\theta \in \tilde{T}_\Theta} ||\tilde{z} - \tilde{\theta}||^2$ is the squared distance between a normal random variable $\tilde{z}$ and its projection on the linearly transformed cone $\tilde{T}_\Theta$. In full likelihood $\tilde{z}$ has standard normal distribution and in composite likelihood $\tilde{z} \sim N(0, U^{-T} J U^{-1})$.

Consider the tangent cone $T_\Theta(\theta) = \left\{\theta \in \mathbb{R}^k : \mathcal{A}^T \theta \geq 0\right\}$. The transformed tangent cone for $\tilde{\theta} = U\theta$

$$\begin{aligned} \tilde{T}_\Theta(\theta) &= U T_\Theta(\theta) \\ &= \left\{U\theta \in \mathbb{R}^r : \mathcal{A}^T \theta \geq 0\right\} \\ &= \{\tilde{\theta} : \mathcal{A}^T U^{-1} \tilde{\theta} \geq 0\} \\ &= \{\tilde{\theta} : \mathcal{A}_1 U^{-1} \tilde{\theta} \geq 0\} \cap \{\tilde{\theta} : \mathcal{A}_2 U^{-1} \tilde{\theta} = 0\} \\ &= \tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2 \end{aligned}$$

Let $b_i = (\mathcal{A}^T U^{-1})_i, i = 1, \ldots, r$ denote the row of the matrix, then $\text{ri}\,(F_\mathcal{I}) = \{\cap_{i \in \mathcal{I}} \{\theta : b_i \theta = 0\}\} \cap$

$\{\cap_{i\in\mathcal{I}^c}\{\theta : b_i\theta > 0\}\}$. Similarly, define $\widetilde{T}_{\Theta_0}(\theta)$, the transformed tangent cone of the null parameter space and let $ri(F_{\mathcal{I}_0})$ denote relative interior set of the null cone, $\mathcal{I}_0 \subseteq \{1, \ldots, q_0\}$. From (4.3), it is deduced that each term of $\bar{\chi}^2$ can be computed by finding the projection of each point on the tangent cone, which is partitioned into its relative interior sets, then using (D.1) and proposition (4.2.2)

$$\begin{aligned}
\inf_{\theta\in\tilde{T}_\Theta} ||\tilde{z}-\tilde{\theta}||^2 &= \|\tilde{z}-\Pi(\tilde{z}|\tilde{T}_\Theta(\theta))\|^2 \\
&= \|\tilde{z}-\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\|^2 \\
&= \sum_{\mathcal{I}\in 2^{\{1,\ldots,q\}}} \|\tilde{z}-\Pi(\tilde{z}|ri(F_{\mathcal{I}_j})\cap\tilde{\mathbb{P}}_2)\|^2 \;\; \mathbb{I}\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in \mathrm{ri}(F_{\mathcal{I}})\cap\tilde{\mathbb{P}}_2\Big) \\
&= \sum_{\mathcal{I}\in 2^{\{1,\ldots,q\}}} \tilde{z}^T\mathcal{Q}_{\mathcal{I}}\;\tilde{z}\;\; \mathbb{I}\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in \mathrm{ri}(F_{\mathcal{I}})\cap\tilde{\mathbb{P}}_2\Big) \qquad (4.13)
\end{aligned}$$

Let $m = 2^q$ and $m_0 = 2^{q_0}$. From (4.13) the random variable $\bar{\chi}^2$ is written as

$$\begin{aligned}
\bar{\chi}^2 &= \sum_{\mathcal{I}_0\in 2^{\{1,\ldots,q_0\}}} \tilde{z}^T\mathcal{Q}_{0\mathcal{I}_0}\;\tilde{z}\;\; \mathbb{I}\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01}\cap\tilde{\mathbb{P}}_{02})\in \mathrm{ri}(F_{\mathcal{I}_0})\cap\tilde{\mathbb{P}}_{02}\Big) \\
&- \sum_{\mathcal{I}\in 2^{\{1,\ldots,q\}}} \tilde{z}^T\mathcal{Q}_{\mathcal{I}}\;\tilde{z}\;\; \mathbb{I}\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in \mathrm{ri}(F_{\mathcal{I}})\cap\tilde{\mathbb{P}}_2\Big) \\
&= \sum_{j=1}^{m_0}\sum_{i=1}^{m} \tilde{z}^T\Big(\mathcal{Q}_{0\mathcal{I}_{j,0}}-\mathcal{Q}_{\mathcal{I}_i}\Big)\;\tilde{z} \\
&\times\;\; \mathbb{I}\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01}\cap\tilde{\mathbb{P}}_{02})\in \mathrm{ri}(F_{\mathcal{I}_{0j}})\cap\tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in \mathrm{ri}(F_{\mathcal{I}_i})\cap\tilde{\mathbb{P}}_2\Big) \qquad (4.14)
\end{aligned}$$

where $\mathcal{Q}_{0\mathcal{I}_{j,0}}$ is the $r \times r$ matrix equal to $\mathcal{Q}_{0\mathcal{I}_0}$ in the upper left hand $r_0 \times r_0$ submatrix, and zero everywhere else. We compute $\mathcal{Q}_{0\mathcal{I}_{j,0}}$ and $\mathcal{Q}_{\mathcal{I}_i}$ in the next subsection. Then from (4.11), using (4.3), distribution of the random variable $\bar{\chi}^2$ is written as

$$\begin{aligned} P(\bar{\chi}^2 \leq c) &= \sum_{j=1}^{m_0}\sum_{i=1}^{m} P\left(\bar{\chi}^2 \leq c \mid \Pi(\tilde{z}|\tilde{T}_{\Theta_0}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{T}_{\Theta}) \in \mathrm{ri}(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\right) \\ &\times\ P(\Pi(\tilde{z}|\tilde{T}_{\Theta_0}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{T}_{\Theta}) \in \mathrm{ri}(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2) \\ &= \sum_{j=1}^{m_0}\sum_{i=1}^{m} P\left(\|\tilde{z} - \Pi(\tilde{z}|\tilde{\mathbb{P}}_{01} \cap \tilde{\mathbb{P}}_{02})\|^2 - \|\tilde{z} - \Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2)\|^2 \leq c\right) \\ &\times\ P\left(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01} \cap \tilde{\mathbb{P}}_{02}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2) \in \mathrm{ri}(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\right) \qquad (4.15) \end{aligned}$$

4.6.1 Computing the distribution of $\bar{\chi}^2$

By equation (4.15), we find a representation for the distribution of $\bar{\chi}^2$. Here, we discuss finding the random variable $S_{ij} = \|\tilde{z} - \Pi(\tilde{z}|\,\mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}\|^2 - \|\tilde{z} - \Pi(\tilde{z}|\,\mathrm{ri}(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\|^2$, as well as computing the quantile c and the probability $P(S_{ij} \leq c)$. The weight $w_{ij} = P\left(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01} \cap \tilde{\mathbb{P}}_{02}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2) \in \mathrm{ri}(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\right)$ and the underlying distribution of the sum of weighted probabilities are also needed to be studied.

For the transformed tangent cone $\tilde{T}_{\Theta}(\theta) = \{\tilde{\theta} \in \mathbb{R}^r : \mathcal{A}^T U^{-1}\theta \geq 0\}$, define $(\mathcal{A}^T U^{-1})_{\mathcal{I}} = \{(\mathcal{A}^T U^{-1})_i, i \in \mathcal{I}\}$,the set of the rows of $\mathcal{A}^T U^{-1}$ which correspond to the elements of $\mathcal{I}$ in the relative interior set of $\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2$. From (4.4), we get

$$\|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}}) \cap \mathbb{P}_2)\|^2 = \tilde{z}^T \ddot{U}_{\mathcal{I}}^T \left(\ddot{U}_{\mathcal{I}} \ddot{U}_{\mathcal{I}}^T\right)^{-1} \ddot{U}_{\mathcal{I}} \tilde{z} \qquad (4.16)$$

where $\ddot{U} = \mathcal{A}^T U^{-1}$

Define $\ddot{U}_{0\mathcal{I}}$ similar to $\ddot{U}_{\mathcal{I}}$ for null tangent cone $\tilde{T}_{\Theta_0}$. Then

$$\begin{aligned}
S_{ij} &= \|\tilde{z} - \Pi(\tilde{z}|\tilde{T}_{\Theta_0}(\theta))\|^2 - \|\tilde{z} - \Pi(\tilde{z}|\tilde{T}_{\Theta}(\theta))\|^2 \\
&= \|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02})\|^2 - \|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2)\|^2 \\
&= \tilde{z}^T \left(\ddot{U}_{\mathcal{I}_{0j}}^T \left(\ddot{U}_{\mathcal{I}_{0j}} \ddot{U}_{\mathcal{I}_{0j}}^T \right)^{-1} \ddot{U}_{\mathcal{I}_{0j}} \right) \tilde{z} - \tilde{z}^T \left(\ddot{U}_{\mathcal{I}_i}^T \left(\ddot{U}_{\mathcal{I}_i} \ddot{U}_{\mathcal{I}_i}^T \right)^{-1} \ddot{U}_{\mathcal{I}_i} \right) \tilde{z} \\
&= \tilde{z}^T \left(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i} \right) \tilde{z}
\end{aligned}$$

Therefore (4.15) becomes

$$P(\bar{\chi}^2 \leq c) = \sum_{j=1}^{m_0} \sum_{i=1}^{m} P\left(\tilde{z}^T \left(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i} \right) \tilde{z} \leq c \right) w_{ij} \tag{4.17}$$

Remark 3. Let $\mathcal{I}$ be any subset of $\{1, \ldots, r\}$,

- If $\mathcal{I} = 2^{\{1,\ldots,q\}}$, the power set, then $\|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}})) \cap \tilde{\mathbb{P}}_2\|^2 = \|\tilde{z}\|^2$,
- If $\mathcal{I} = \emptyset$ then $||\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}}))\|^2 = 0$,
- As we consider only the rows corresponds to the set $\mathcal{I}$, the matrix $U_{\mathcal{I}}^{-1}$ is not necessarily a square matrix and therefore we cannot say that $(U_{\mathcal{I}}^{-T} U_{\mathcal{I}}^{-1})^{-1} = U_{\mathcal{I}} U_{\mathcal{I}}^T$.

Some specific cases

- **Case 1:** Consider the case $\Theta_0 = \{\theta : g_i(\theta) = 0, i = 1, \ldots, r\}$ and $\Theta = \{\theta : g_i(\theta) \geq 0, i = 1, \ldots, r\}$ therefore the tangent cones are $T_{\Theta_0}(\theta) = \cap_{i=1}^r \{\theta : a_i^T \theta = 0\} = \{\theta : \mathcal{A}^T \theta = 0\}$ and $T_{\Theta}(\theta) = \cap_{i=1}^q \{\theta : a_i^T \theta \geq 0\} \cap \{\cap_{i=q+1}^r \{\theta_i \in \mathbb{R}\}\} = \{\theta \in \mathbb{R}^r : \mathcal{A}^T \theta \geq 0\}$.

Under the null hypothesis,

$$\begin{aligned}\inf_{\theta\in T_{\Theta_0}}(z-\theta)^T H(z-\theta) &= \inf_{\theta\in T_{\Theta_0}}(Uz-U\theta)^T(Uz-U\theta)\\ &= \inf_{\theta\in T_{\Theta_0}}(U\mathcal{A}^{-T}\mathcal{A}^T z-U\mathcal{A}^{-T}\mathcal{A}^T\theta)^T(U\mathcal{A}^{-T}\mathcal{A}^T z-U\mathcal{A}^{-T}\mathcal{A}^T\theta)\\ &= (Uz)^T(Uz)=\tilde{z}^T\tilde{z}\end{aligned}$$

Hence for $\tilde{\theta}=U\theta$ and $\tilde{T}_\Theta=\{\tilde{\theta}:\mathcal{A}^T U^{-1}\tilde{\theta}\geq 0\}$.

the test statistic can be written as

$$\begin{aligned}\bar{\chi}^2 &= \tilde{z}^T\tilde{z}-\inf_{\theta\in\tilde{T}_\Theta}(\tilde{z}-\tilde{\theta})^T(\tilde{z}-\tilde{\theta})\\ &= \sum_{i=1}^{m}\tilde{z}^T\left(I_{r\times r}-(U^{-T}\mathcal{A}^T)_{\mathcal{I}_i}\left[(\mathcal{A}U^{-1})_{\mathcal{I}_i}(U^{-T}\mathcal{A}^T)_{\mathcal{I}_i}\right]^{-T}(\mathcal{A}U^{-1})_{\mathcal{I}_i}\right)\tilde{z}\\ &\times\ \mathbb{I}(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in ri(F_{\mathcal{I}_i})\cap\tilde{\mathbb{P}}_2)\end{aligned}$$

and the distribution (4.17) is written as

$$P(\bar{\chi}^2\leq c) = \sum_{i=1}^{m}P\Big(\tilde{z}^T\left(I_{r\times r}-\mathcal{Q}_{\mathcal{I}_i}\right)\tilde{z}\leq c\Big)\,P\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1\cap\tilde{\mathbb{P}}_2)\in ri(F_{\mathcal{I}_i})\cap\tilde{\mathbb{P}}_2\Big) \tag{4.18}$$

- **Case 2:** If in the parameter space Θ, $g_i(\theta)=\theta_i, i=1,\ldots,r$, then the matrix $\mathcal{A}=$

$[\mathcal{A}_1^T \;\; \mathcal{A}_2^T]$ becomes an $r \times r$ identity matrix. Then (4.17) is simplified to

$$\begin{aligned}
& \|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02})\|^2 - \|\tilde{z} - \Pi(\tilde{z}|ri(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2)\|^2 \\
= \;& \tilde{z}^T \left(U_{0\mathcal{I}_j}^{-T} (U_{\mathcal{I}_{0j}}^{-1} U_{\mathcal{I}_{0j}}^{-T})^{-1} U_{\mathcal{I}_{0j}}^{-1} - U_{\mathcal{I}_i}^{-T} (U_{\mathcal{I}_i}^{-1} U_{\mathcal{I}_i}^{-T})^{-1} U_{\mathcal{I}_i}^{-1} \right) \tilde{z} \qquad (4.19) \\
= \;& \tilde{z}^T \Big(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i} \Big) \tilde{z}
\end{aligned}$$

Therefore, distribution of $\bar{\chi}^2$ is similar to (4.17), with $\ddot{U} = U$.

- **Case 3:** Assume $\Theta_0 = \{\theta : \theta_1 = \ldots = \theta_r = 0\}$ and $\Theta = \{\theta : \theta_i \geq 0, i = 1, \ldots q, \theta_j \in \mathbb{R}, j = q+1, \ldots, r\}$, that is q parameters lie on the boundary and $r - q$ parameters are interior. By computing the Jacobian matrix, it is clear that the tangent cone $T_{\Theta_0}(\theta) = \Theta_0 = \{0\}^r$, and $T_\Theta(\theta) = \{\theta \in \mathbb{R}^k : \theta_i \geq 0, i = 1, \ldots, q\} = [0, \infty)^q$. The null cone has only one face with $\mathcal{I} = \{1, \ldots, r\}$. Therefore, $U_{0\mathcal{I}}$ is a $r \times r$ matrix and $\mathcal{Q}_{0\mathcal{I}_{j,0}}$ in (4.19) becomes an identity matrix. Similar to (4.18) the distribution of $\bar{\chi}^2$ becomes

$$\begin{aligned}
P(\bar{\chi}^2 \leq c) \;=\; & \sum_{i=1}^{m} P\Big(\tilde{z}^T \left(I_{r\times r} - U_{\mathcal{I}_i}^{-T} (U_{\mathcal{I}_i}^{-1} U_{\mathcal{I}_i}^{-T})^{-1} U_{\mathcal{I}_i}^{-1} \right) \tilde{z} \leq c \Big) \\
\times \;& P\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2) \in ri(F_{\mathcal{I}_i}) \Big) \\
=\; & \sum_{i=1}^{m} P\Big(\tilde{z}^T \left(I_{r\times r} - \mathcal{Q}_{\mathcal{I}_i} \right) \tilde{z} \leq c \Big) \; P\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2) \in ri(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2 \Big)
\end{aligned} \qquad (4.20)$$

Example 4.6.1. In a $3-$dimensional parameter space, $\Theta = \{\theta \in \mathbb{R}^3 : \theta_i \geq 0, i = 1, 2, 3\} = [0, \infty)^3$. One is interested to test $H_0 : \theta_1 = \theta_2 = 0$, and θ_3 may lie on the boundary. Then

$\Theta_0 = \{\theta \in \mathbb{R}^3 : \theta_1 = \theta_2 = 0\} \cap \{\theta \in \mathbb{R}^3 : \theta_3 \geq 0\} = \{0\}^2 \times [0, \infty)$. The tangent cones $T_\Theta(\theta) = \mathbb{P}_1$ and $T_{\Theta_0} = \mathbb{P}_{02} \cap \mathbb{P}_{01}$ are equal to Θ and Θ_0, respectively with $m = 2^3$ and $m_0 = 2$ relative interior sets in each one.

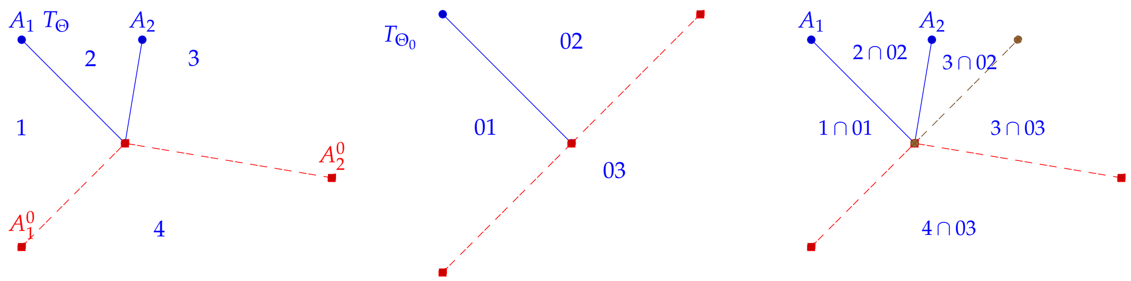
After estimating parameters through composite likelihood and estimating $H = U^T U$ and covariance $\widehat{\Sigma}$ matrix, we can compute the 3×3 matrix $\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i}$ in the test statistics, similar to the case 2,

$$\bar{\chi}^2 = \sum_{j=1}^{2}\sum_{i=1}^{2^3} \tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z}\; P\Big(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01} \cap \tilde{\mathbb{P}}_{02}) \in ri(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{\mathbb{P}}_1) \in ri(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\Big) \tag{4.21}$$

where $\mathcal{I}_{j,0}$ and $\mathcal{I}_i$ belong to $2^{\{3\}} \cup \{1,2\}$ and $2^{\{1,2,3\}}$, respectively.

4.6.2 Computing the weight

Recall $w_{ij} = P\left(\Pi(\tilde{z}|\tilde{\mathbb{P}}_{01} \cap \tilde{\mathbb{P}}_{02}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \tilde{\mathbb{P}}_{02}, \Pi(\tilde{z}|\tilde{\mathbb{P}}_1 \cap \tilde{\mathbb{P}}_2) \in ri(F_{\mathcal{I}_i}) \cap \tilde{\mathbb{P}}_2\right)$. The exact computation of the weight w_{ij} could be very complicated specially in dimensions larger than three. To explain why computation of weight can be complicated in practice, let us look at the two dimensional example. Assume in hypothesis testing there is one parameter of interest and one which is not tested and both lie on the boundary. For simplicity assume $g_i(\theta) = \theta_i,\ i = 1, 2$. Under the null $T_{\Theta_0} = \{0\} \times [0, \infty)$ and under the alternative $T_\Theta = [0, \infty) \times [0, \infty)$. The transferred tangent cones are $\tilde{T}_{\Theta_0} = \{\tilde{\theta} : U_1^{-1}\tilde{\theta} = 0, U_2^{-1}\tilde{\theta} \geq 0\}$ and $\tilde{T}_\Theta = \{\tilde{\theta} : U_1^{-1}\tilde{\theta} \geq 0, U_2^{-1}\tilde{\theta} \geq 0\}$, where U_1 and U_2 denote the rows of matrix U. The figure (4.4c) shows that there are five non-empty intersections which are the weights in the



(a) Under H_a

(b) Under H_0

(c) intersection of two plots

Figure 4.4: Tangent cones for $H_0 : a_1^T\theta = 0$ vs $H_0 : a_1^T\theta \geq 0$, while $a_2^T\theta_2 \geq 0$

test statistic. For higher dimensions, the number of weights increases dramatically.

We propose two algorithm for computing the weight; one based on quadratic optimization and the second one by looking at the regions between each relative interior set of the cone and its corresponding set on the polar cone.

First Algorithm

In this method, we use quadratic programming to minimize equation $(z-\theta)^T H(z-\theta)$, subject to $\left\{\theta \in \mathbb{R}^k : \mathcal{A}^T\theta \geq 0\right\}$, where $\mathcal{A} = [\mathcal{A}_1^T \ \ \mathcal{A}_2^T]$ That is,

$$\begin{array}{lrcl} \text{minimize} & Q(x) & = & x^T\Sigma^{-1}x, \quad x = z - \theta, \\ & -\mathcal{A}_1 x & \geq & -\mathcal{A}_1 z \\ & \mathcal{A}_2 x & = & \mathcal{A}_2 z \end{array}$$

when $z \sim N(0,\Sigma)$. The constraint $\mathcal{A}_2 x = \mathcal{A}_2 z$ represents the parameters which lie on the boundary under the null hypothesis and show up in $\mathbb{P}_2$. Suppose x^* is the minimizer of $Q(x)$, then the point on the cone with minimum distance from z is $z - x^*$. We solved the quadratic equations using the ″quadprog″ R-package. The results are the projection points for any given samples on the polyhedral cone. The coordinate of the projection points is either zero or greater than zero that the number of zeros shows $k-$ dimension of the face and the location of the zeros specifies which relative interior set in $T_{\Theta_0}(\theta)$ contains $z - x^*$.

The steps for computing the weight are

1. generate n data point $z_1, z_2, \ldots, z_n$ from $N_k(0, \widehat{\Sigma})$. The matrix $\widehat{\Sigma}$ is the estimated covariance structure estimated by from the full or composite likelihood approach.

2. minimizing over the total tangent cone ; for each $z_i,\ i = 1, \ldots, n$, finding the projection point θ_{01} that minimizes $(z_i - \theta)^T \widehat{\Sigma}(z_i - \theta)$ over $T_\Theta(\theta)$, using quadratic programming.

3. minimizing over the null cone; for each $z_i,\ i = 1, \ldots, n$, finding the projection point θ_{00} that minimizes $(z_i - \theta)^T \widehat{\Sigma}(z_i - \theta)$ over $T_{\Theta_0}(\theta)$. (In cases that there is no nuisance parameters on the boundary, $inf_{\theta \in T_{\Theta_0}(\theta)}(z_i - \theta)^T \widehat{\Sigma}(z_i - \theta) = z_i^T \widehat{\Sigma} z_i$. So dismiss this step and go to the next one).

4. Define $n \times k$ matrix P_1, where $P_1 = [\theta_{00}\ \ \theta_{01}]$.

5. allocating zero and one to elements of P_1 such that if the projection is not zero and 1 if that is zero. Then, finding the proportion of identical rows which shows the proportion of z_i's which their projection lie simultaneously in a relative interior set of $T_{\Theta_0}(\theta)$ and one of $T_\Theta(\theta)$. The proportion of identical rows gives the empirical weights.

To clarify the algorithm, we explain the steps using an example.

Example 4.6.2. (continued) In example 4.6.1, the test statistic for hypothesis testing in a 3-dimensional space with one nuisance parameter one the boundary was defined. To find the weights, first we generate a sample $z_1, \ldots, z_{100}$ from $N_3(0, \widehat{\Sigma})$, where $\widehat{\Sigma}$ is estimated full or composite maximum likelihood covariance structure. To compute the weight, we use quadratic programming. Under the null hypothesis, the constraints matrices of the quadratic programming are $A_1 = diag(1,1,0)$ and $A_2 = diag(0,0,1)$ and under the alternative are $A_1 = 0$ and $A_2 = diag(1,1,1)$. The two left columns of the Table 4.3 shows the projection of the first 20 generated observations on the null and alternative cones.

Table 4.3: $\inf_\theta (z-\theta)^T H (z-\theta)$ under H_0 and H_a

under H_0			under H_a			under H_0			under H_a		
0	0	0	0	0.058	0	1	1	1	1	0	1
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0.020	0	0	1	1	1	0	1	1
0	0	0	0.036	0	0	1	1	1	0	1	1
0	0	0.022	0.077	0	0.007	1	1	0	0	1	0
0	0	0.040	0	0	0.040	1	1	0	1	1	0
0	0	0.018	0	0.056	0	1	1	0	1	0	1
0	0	0.041	0	0.031	0.026	1	1	0	1	0	0
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0.008	0	0	0.008	1	1	0	1	1	0
0	0	0	0	0.021	0	1	1	1	1	0	1
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0.006	0.082	0	0	1	1	0	0	1	1
0	0	0.025	0	0	0.025	1	1	0	1	1	0
0	0	0	0.011	0	0	1	1	1	0	1	1
0	0	0	0.028	0	0	1	1	1	0	1	1
0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1
⋮											

As it can be seen the optimal values of θ on the cone are either zero or greater than zero. The two right columns of the Table 4.3 is made by allocating 0 and 1 to interior coordinates and the boundary ones of the projection points in the right side, which each row uniquely specifies the relative interior set that projection point belongs to it. As an example some identical series in the table 4.3 are shadowed. Finding the proportion of identical series of 0 and 1 gives an estimation of the weights for each relative interior set.

In total, there are 16 possible series of 0 and 1's. In this example, observations place such that 7 of the relative interior sets contains at least a projection point and weight of the rest of the sets are zero. The proportion of identical rows gives an estimation of the weights for the two relative interior set simultaneously. Weights are shown in table 4.4.

□

Second Algorithm

Define the sets

$$\begin{aligned} W_{0,j} &= \left\{ z \in \mathbb{R}^r : \Pi(z|\mathbb{P}_{01} \cap \mathbb{P}_{02}) \in \mathrm{ri}(F_{\mathcal{I}_{0j}}) \cap \mathbb{P}_{02} \right\}, \\ W_i &= \left\{ z \in \mathbb{R}^r : \Pi(z|\mathbb{P}_1 \cap \mathbb{P}_2) \in \mathrm{ri}(F_{\mathcal{I}_i}) \cap \mathbb{P}_2 \right\}. \end{aligned}$$

The weight $w_{ij} = P\Big(\Pi(z|\mathbb{P}_{01} \cap \mathbb{P}_{02}) \in \mathrm{ri}(F_{\mathcal{I}_0}) \cap \mathbb{P}_{02}, \Pi(z|\mathbb{P}_1 \cap \mathbb{P}_2) \in \mathrm{ri}(F_{\mathcal{I}}) \cap \mathbb{P}_2\Big)$ is equivalent to $P\Big(z \in W_{0,j} \cap W_i\Big)$ probability that an observation z locates in the region that its projection onto $T_\Theta(\theta)$ and $T_{\Theta_0}(\theta)$ lie on $ri(F_{\mathcal{I}_i}) \cap \mathbb{P}_2$ and $ri(F_{\mathcal{I}_{0j}}) \cap \mathbb{P}_{02}$, simultaneously.

First we find the region W_i. Let $\mathcal{A}^T = [a_1^T \; a_2^T \; \ldots \; a_r^T]$, the tangent cone is $\mathbb{P} = \{\theta : \mathcal{A}^T U^{-1}\theta \geq 0\}$ and the polar cone is $\mathbb{P}^0 = \{y \; : (\mathcal{A}^T U^{-1})^{-T} y \leq 0\}$. Let b_i and $\dot{b}_i$ denote the rows of $\mathcal{A}^T U^{-1}$ and $(\mathcal{A}^T U^{-1})^{-T}$, respectively. The relative interior sets of $\mathbb{P}$ and $\mathbb{P}^0$ are ri $(F_{\mathcal{I}}) = \{\cap_{i \in \mathcal{I}}\{\theta : b_i\theta = 0\}\} \cap \{\cap_{i \in \mathcal{I}^c}\{\theta : b_i\theta > 0\}\}$ and ri $\left(F_{\mathcal{I}}^0\right) = \{\cap_{i \in \mathcal{I}}\{\theta : \dot{b}_i\theta < 0\}\} \cap$

$\left\{\cap_{i\in\mathcal{I}^c}\{\theta : \dot{b}_i\theta = 0\}\right\}$. Then

$$W_i = \{y : B^{-T}y \leq 0\},$$

where $B = \begin{pmatrix} b_{\mathcal{I}} \\ \\ -\dot{b}_{\mathcal{I}^c} \end{pmatrix}$. If $\mathcal{I} = \emptyset$, then W_i is inside the cone $\mathbb{P}$ and if $\mathcal{I} = \{1, \dots, r\}$, then W_i represents inside the polar cone $\mathbb{P}^0$. For $\mathcal{I} \subset \{1, \dots, r\}|\emptyset$ the region is constructed between the relative interior of the cone and its corresponding relative interior on the polar cone which is perpendicular to it. It means the area between $\{b_i : i \in \mathcal{I}\}$ and $\{\dot{b}_i : i \in \mathcal{I}^c\}$, which is a polyhedral cone.

Then, to find $W_{0,j}$ let $\mathcal{A}_{01} = [a_1^T \; a_2^T \; \dots \; a_{q_0}^T]$ and $\mathcal{A}_{02} = [a_{q_0+1}^T \; \dots \; a_r^T]$, the tangent cone is $\mathbb{P}_0 = \{\theta : \mathcal{A}_{01}U^{-1}\theta \geq 0\} \cap \{\theta : \mathcal{A}_{02}U^{-1}\theta = 0\}$ and the polar cone is $\mathbb{P}_0^0 = \{y : (\mathcal{A}_{01}U^{-1})^{-T}y \leq 0\} \cap \mathbb{R}$. The relative interior sets of $\mathbb{P}_{01} = \{\theta : \mathcal{A}_{01}U^{-1}\theta \geq 0\}$ is $\mathrm{ri}\left(F_{\mathcal{I}_0}\right) = \left\{\cap_{i\in\mathcal{I}_0}\{\theta : b_i\theta = 0\}\right\} \cap \left\{\cap_{i\in\mathcal{I}_0^c}\{\theta : b_{0i}\theta > 0\}\right\}$, where b_{0i} denote the rows of $\mathcal{A}_{01}U^{-1}$. Similarly, for $\mathbb{P}_0^0$ define $\mathrm{ri}\left(F_{\mathcal{I}_0}^0\right) = \left\{\cap_{i\in\mathcal{I}_0}\{\theta : \dot{b}_{0i}\theta < 0\}\right\} \cap \left\{\cap_{i\in\mathcal{I}_0^c}\{\theta : \dot{b}_{0i}\theta = 0\}\right\}$ and $\dot{b}_{0i}$ denote the rows of $(\mathcal{A}_{01}U^{-1})^{-T}$. Then $W_{0,j} = \{y : B_0^{-T}y \leq 0\}$, where $B_0 = \begin{pmatrix} b_{\mathcal{I}_0'} \\ \\ -\dot{b}_{\mathcal{I}_0^c} \end{pmatrix}$ and $\mathcal{I}_0' = \mathcal{I}_0 \cup \{q_0 + 1, \dots, r\}$.

The steps of this algorithm are

1. generating data $y_1, \dots, y_n$ from $N_r(0, \widehat{\Sigma})$,

   (total cone)

2. $\mathcal{A} =$ the Jacobian matrix under the alternative hypothesis and $\mathcal{I} \subseteq 2^{\{1,\dots,r\}}$

3. computing $\mathcal{A}^T U^{-1}$ and $(\mathcal{A}^T U^{-1})^{-T}$. If the latter is not square we use a pseudo-

inverse that is $B^{-1} = (B^TB)^{-1}B^T$.

4. defining $b_{\mathcal{I}}$ for the cone and $-\dot{b}_{\mathcal{I}^c}^{-T}$ for the polar cone.

5. constructing the matrix $\begin{pmatrix} b_{\mathcal{I}} \\ \dot{b}_{\mathcal{I}^c} \end{pmatrix}$ by by combining orthogonal rows from $\mathbb{P}$ and $\mathbb{P}^0$.

6. computing $Y = \begin{pmatrix} b_{\mathcal{I}} \\ \dot{b}_{\mathcal{I}^c} \end{pmatrix}^{-T} \tilde{y}$ for the generated sample and $\tilde{y} = Uy$. For each observation, Y is a $r \times 2^r$ matrix. Each column corresponds to one relative interior set. If all elements of a specific column are negative then it results that the observation lies inside the side polyhedral cone made by relative interior set corresponding to that column and the one on the polar cone. Let $ind =$ # of relative interior set specified by the column.

   (null cone)

7. computing $\mathcal{A}_0^T U^{-1}$ and $(\mathcal{A}_0^T U^{-1})^{-T}$, where $\mathcal{A}_0^T = [\mathcal{A}_{01} \;\; \mathcal{A}_{02}]$.

8. $q_0 =$ number of constraints and defining $\mathcal{I}_0 \subseteq 2^{\{1,\ldots,q_0\}}$ and $\mathcal{I}_0' = \mathcal{I}_0 \cup \{q_0 + 1, \ldots, r\}$.

9. defining $\dot{b}_{\mathcal{I}_0'}$ for the cone and $-\dot{b}_{\mathcal{I}_0^c}^{-T}$ for the polar cone.

10. constructing $q_0 \times r$ matrix $\begin{pmatrix} b_{\mathcal{I}_0'} \\ \dot{b}_{\mathcal{I}_0^c} \end{pmatrix}$ by by combining orthogonal rows from $\mathbb{P}_0$ and $\mathbb{P}_0^0$.

11. computing $Y_0 = \begin{pmatrix} b_{\mathcal{I}_0'} \\ \dot{b}_{\mathcal{I}_0^c} \end{pmatrix}^{-T} \tilde{y}$ for the generated sample and $\tilde{y} = Uy$. If elements 1 to q_0 are negative then $ind_0 =$ # of relative interior set specified by the column of Y_0.

12. merging ind and ind_0 into one matrix $[ind_0 \;\; ind]$ and finding the proportion of the similar rows, that gives the weight for simultaneous relative interior sets from null and total cone.

4.6.3 Computing the coefficients

In equation (4.17), it is shown that to find the distribution of $\bar{\chi}^2$ we need to compute the distribution of the quadratic form $\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z}$. Using lemma 3.2 in Kudo [28], it is shown that the distribution of $\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z}$ is independent of w_{ij}, therefore $P(\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z} \leq c|\tilde{z}) = P(\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z} \leq c)$. Then the distribution of $\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z}$ is the same as $\sum_{l=1}^{q} \lambda_l V_l$ that $V_l,\ l = 1, \ldots, q$ are independent non-central chi-square distribution with one degrees of freedom.

Hence (4.17) with the form $P(\bar{\chi}^2 \leq c) = \sum_{j=1}^{m_0} \sum_{i=1}^{m} P\Big(\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z} \leq c\Big)\, w_{ij}$ can be expressed in this form

$$P(\bar{\chi}^2 \leq c) \;=\; \sum_{k=1}^{s} P\left(\sum_{l=1}^{l_{ijk}} \lambda_{kl}\chi_1^2 \leq c\right) w_k \qquad (4.22)$$

where s is the number of non-zero weights.

In composite likelihood, the weights λ_{kl} are the eigenvalues of $(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})(U^{-T}JU^{-1})$ and l_{ijk} is the number of eigenvalues for the relative interiors that are corresponded to the weight w_k.

In full likelihood case, the weights λ_{kl} become equal to one. Therefore $\tilde{z}^T(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})\tilde{z}$ is distributed as $\chi^2_{l_{ijk}}$, where l_{ijk} is rank of $(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})$. That is $P(\bar{\chi}^2 \leq c) = \sum_{k=1}^{s} P\left(\chi^2_{l_{ijk}} \leq c\right) w_k$.

Approximation

The obtained formula for $P(\bar{\chi}^2 \leq c)$ can become very large even for a slightly large number of parameters. We employed Satterthwaite [39] approach that approximate the mixture of independent χ_1^2 as $c\chi_{df}^2$, where c and df are chosen so that the first two moments of the two distributions are equal. The expectation and variance of $\sum \lambda_{kl}\chi_1^2$ are defined as $E_k = \sum_l \lambda_{kl}$ and $V_k = 2\sum_l \lambda_{kl}^2$. Then the coefficient and degree of freedom are $g_k = V_k/(2E_k)$ $df_k = 2E_k^2/V_k$, respectively. Finally the distribution of $\bar{\chi}^2$ becomes

$$P(\bar{\chi}^2 \le c) = \sum_{k=1}^{s} P\left(g_k\chi_{df_k}^2 \le c\right) w_k$$

The algorithm for approximating the weighted sum of χ_1^2 into one chi-square term is:

1. if length of non-zero elements in vector $(\lambda_{k1}, \ldots) > 1$ then go to step 2, otherwise $g_k = \lambda_{k1}$ and $df_k = 1$,

2. $E_k = \sum_l \lambda_{kl}$ and $V = 2\sum_l \lambda_{kl}^2$,

3. if $(E_k = 0 \& V_k = 0)$ then $g_k = 0$ and $df_k = 0$ otherwise $g_k = V_k/(2E_k)$ and $df_k = 2E_k^2/V_k$,

4.6.4 Computing quantile

After computing the weights, to find the quantile of the distribution, define $G(c) = P(\bar{\chi}^2 \le c)$. Newton's method give the estimate of the c through

$$c_{n+1} = c_n - \frac{G(c_n)}{G'(c_n)},$$

where G' is the derivative of g. The algorithm for finding the quantile in composite likelihood is:

Example 4.6.3. (continued) In example 4.6.1, the matrix $(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})$ is computed. To find the coefficient of the mixture of the chi-squares, we find the eigenvalues of the $(\mathcal{Q}_{0\mathcal{I}_{j,0}} - \mathcal{Q}_{\mathcal{I}_i})U^{-T}JU^{-1}$, for every relative interior set with non-zero probability. In total there are $2^3 \times 2$ cases with with at most 3 eigenvalues, but in our example only 7 cases have non-zero weight. Let $l_{ijk} \in \{1, 2, 3\}$. Table 4.4 shows the coefficients $\lambda_1, \lambda_2, \lambda_3$ in

$$P(\bar{\chi}^2 \le c) = \sum_{k=1}^{7} P(\sum_{l=1}^{l_{ijk}} \lambda_{kl}\chi_1^2 \le c)w_k,$$

Table 4.4: All coefficients and weights. There are only 7 non-zero weights

	λ_1	λ_2	λ_3	weight
1	2.461	1.895	0	0.140
2	2.455	0	0	0.190
3	1.944	0	0	0.080
4	2.149	1.808	-1.303	0
5	0	0	0	0.150
6	2.118	-1.303	0	0
7	1.808	-1.614	0	0
8	-1.648	0	0	0
9	2.540	1.997	1.467	0
10	2.540	1.563	0	0
11	1.998	1.594	0	0
12	2.478	1.824	0	0
13	1.648	0	0	0
14	2.463	0	0	0.120
15	1.843	0	0	0.170
16	0	0	0	0.150

Approximating the coefficients into one term, we can estimate the distribution by a mixture of chi-squares

$$P(\bar{\chi}^2 \leq c) = \sum_{k=1}^{7} P(g_k \chi^2_{df_k} \leq c) w_k.$$

	g	df
1	2.21	1.97
2	2.45	1.00
3	1.94	1.00
4	0.00	0.00
5	2.46	1.00
6	1.84	1.00
7	0.00	0.00

Then using Newton's algorithm the quantile c is computed as 8.323.

Chapter 5

Simulation of Likelihood Ratio Testing in Non-standard Condition

In this chapter, first, different steps in finding the distribution of $\bar{\chi}^2$ is demonstrated by some examples. Then the methods and algorithms are evaluated through simulation in section (5.2).

5.1 Examples

Example 5.1.1. Assume $y = x\beta + \epsilon$, where $y_{4\times1}$ has multivariate normal distribution and the parameter $\beta_{p\times1}$ is restricted to be non-negative. The matrix x is a $4 \times p$ matrix where in parts A and B, $p = 3$ and in C, $p = 2$. Three different scenarios are considered in this example. The suggested composite likelihood ratio test is applied in this case.

A. Test in 3-dimensional β, testing $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \geq 0, \beta_2 \geq 0$, and β_3

Table 5.1: testing if $\beta_1 = \beta_2 = 0$ and β_3 lies on the boundary.

λ_1	λ_2	λ_3	$\mathcal{I}_{0j}$			$\mathcal{I}_i$			weights
1.16	1.09	0	0	1	2	0	0	0	0.129
1.15	0	0	0	1	2	0	1	0	0.111
1.09	0	0	0	1	2	0	0	2	0.127
0	0	0	0	1	2	0	1	2	0.140
1.16	1.09	1.03	1	2	3	0	0	0	0.006
1.09	1.04	0	1	2	3	0	0	2	0.001
1.16	1.09	0	1	2	3	0	0	3	0.127
1.15	0	0	1	2	3	1	0	3	0.112
1.09	0	0	1	2	3	1	2	3	0.120
0	0	0	1	2	3	1	2	3	0.127

may lie on the boundary.

Assume β is a three dimensional vector, that β_3 is a nuisance boundary point. For this case, $\Theta_0 = \{\beta \in \mathbb{R}^3 : \mathcal{A}_1\beta = 0, \mathcal{A}_2 \geq 0\}$ and $\Theta = \{\beta \in \mathbb{R}^3 : I_{3\times 3}\beta \geq 0\}$, where $\mathcal{A}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, and $\mathcal{A}_2 = (0\ 0\ 1\ )$ matrix.

Then it is concluded that $T_{\Theta_0}(\theta) = \Theta_0$ and $T_{\Theta}(\theta) = \Theta$. The tangent cone $T_{\Theta}(\theta)$ and $T_{\Theta_0}(\theta)$ have 2^3 and 2 faces, respectively and the set $\mathcal{I}_i \in 2^{\{1,2,3\}}$ and $\mathcal{I}_{0j} \in 2^{\{3\}} \cup \{1,2\}$. That is, before the transformation the tangent cone $T_{\Theta}(\theta)$ is the first octant and the null tangent cone is positive side of the z axis plus zero. Then $\bar{\chi}^2$ is the same as (4.21). The results of composite likelihood ratio test suggest that 11 cases out of 16 possible intersection between the faces of two cones have non-zero probability. The distribution can be written as $P(\bar{\chi}^2 \leq c) = \sum_{k=1}^{10} P\left(\sum_{l=1}^{l_{ijk}} \lambda_{kl}\chi_1^2 \leq c\right) w_k$. In Table 5.1, λ_{k1}, λ_{k2} and λ_{k3} are shown. Also different possible matches of $\mathcal{I}_i$ and $\mathcal{I}_{0j}$ is available in the Table 5.1.

Then the test statistic can be written as

$$\bar{\chi}^2 = \begin{cases} 1.16\chi_1^2 + 1.09\chi_1^2 & \text{, w.p. } 0.129 \\ 1.15\chi_1^2 & \text{, w.p. } 0.111 \\ 1.09\chi_1^2 & \text{, w.p. } 0.127 \\ 1.16\chi_1^2 + 1.09\chi_1^2 + 1.03 & \text{, w.p. } 0.006 \\ 1.09\chi_1^2 + 1.04\chi_1^2 & \text{, w.p. } 0.001 \\ 1.16\chi_1^2 + 1.09\chi_1^2 & \text{, w.p. } 0.127 \\ 1.15\chi_1^2 & \text{, w.p. } 0.112 \\ 1.09\chi_1^2 & \text{, w.p. } 0.120 \\ 0 & \text{, w.p. } 0.267 \end{cases} \tag{5.1}$$

The obtained result so far is a weighted sum of mixture of chi-squares. It is possible to approximate each term into one, and it results in a mixture of chi-squares,

$$\begin{aligned} P(\bar{\chi}^2 \le c) &= 0.129P(1.12\ \chi_{1.99}^2 \le c) + 0.111P(1.15\ \chi_1^2 \le c) + 0.127P(1.09\ \chi_1^2 \le c) \\ &+ 0.006P(1.09\ \chi_{2.99}^2 \le c) + 0.001P(1.06\ \chi_{1.99}^2 \le c) + 0.127P(1.12\ \chi_{1.99}^2 \le c) \\ &+ 0.112P(1.15\ \chi_1^2 \le c) + 0.120P(1.09\ \chi_1^2 \le c) + 0.267 \end{aligned} \tag{5.2}$$

Applying newton's method, the %95 quantile becomes 4.78.

To assess the accuracy of the approximation, the cdf of (5.1) and the approximated one in (5.2) and their Q-Q plot, are shown through simulation in figure 5.1.

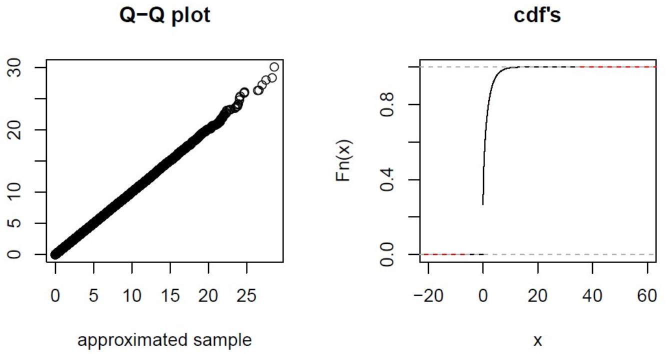


Figure 5.1: Q-Q plot and cdf's of $P(\bar{\chi}^2 \leq c)$ befor and after approximation

B. Test in 3-dimensional $\beta = (\beta_1, \beta_2, \beta_3)$, testing $H_0 : \begin{cases} \beta_1 + 2\beta_2 = 0 \\ 3\beta_1 - 5\beta_2 = 0 \end{cases}$ versus $H_a : \begin{cases} \beta_1 + 2\beta_2 \geq 0 \\ 3\beta_1 - 5\beta_2 \geq 0 \end{cases}$ and β_3 may lie on the boundary.

Assume β is a three dimensional vector, that β_3 could be a boundary point that is not tested. The null parameter space is $\Theta_0 = \{\beta \in \mathbb{R}^3 : \mathcal{A}_1\beta = 0, \mathcal{A}_2\beta \geq 0\}$ and $\Theta = \{\beta \in \mathbb{R}^3 : \mathcal{A}\beta \geq 0\}$, where

$$\mathcal{A}_1 = \begin{pmatrix} 1 & 2 & 0 \\ 3 & -5 & 0 \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 1 & 2 & 0 \\ 3 & -5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Due to linearity of constraints, the Jacobian matix is equal to matrix $\mathcal{A}$ and the tangent cones are the same as parameter spaces. Table 5.2 shows the coefficient and non-zero weights and their corresponding relative interior set in each cone.

Table 5.2: testing $H_0 : \beta_1 + 2\beta_2 = 0,\ \ 3\beta_1 - 5\beta_2 = 0$ and β_3 lies on the boundary.

λ_1	λ_2	λ_3	$\mathcal{I}_{0j}$			$\mathcal{I}_i$			weights
1.156	1.085	0	1	2	0	0	0	0	0.350
0	0	0	1	2	0	1	2	0	0.180
1.157	1.092	1.035	1	2	3	0	0	0	0.005
1.156	1.086	0	1	2	3	0	0	3	0.355
0	0	0	1	2	3	1	2	3	0.110

then the test statistic can be written as

$$\bar{\chi}^2 = \begin{cases} 1.156\chi_1^2 + 1.085\chi_1^2 & \text{, w.p. } 0.350 \\ 1.157\chi_1^2 + 1.092\chi_1^2 + 1.035\chi_1^2 & \text{, w.p. } 0.005 \\ 1.156\chi_1^2 + 1.086\chi_1^2 & \text{, w.p. } 0.355 \\ 0 & \text{, w.p. } 0.290 \end{cases}$$

Also we can approximate the distribution function (5.1.1) as

$$P(\bar{\chi}^2 \leq c) = 0.35P(1.12\chi^2_{1.99} \leq c) + 0.005P(1.10\chi^2_{2.99} \leq c) + 0.355P(1.12\chi^2_{1.99} \leq c) + 0.290$$

C. Test in 2-dimensional β, testing $H_0 : \beta_1^3 = \beta_2^2$ versus $H_a : \beta_1^3 \geq \beta_2^2$.

The limitation of the proposed approach can be illustrated for hypotheses about the mean vector of a bivariate normal distribution in the following example.

Let $\Theta_0 = \{\beta \in \mathbb{R}^2 : \beta_1^3 - \beta_2^2 = 0\}$ that is shown in figure 5.2 and $\Theta = \{\beta \in \mathbb{R}^2 : \beta_1^3 - \beta_2^2 \geq 0\}$.

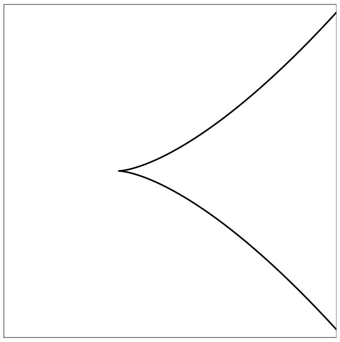


Figure 5.2: $\beta_1^3 - \beta_2^2 = 0$

Then $k = 2,\ r = 1$ and the Jacobian matrix becomes $\mathbb{J}(\beta) = (3\beta_1^2 \quad -2\beta_2)^T$. If the true parameter point is $\beta_0 = (0, 0)$, the cusp, then $\text{rank}(\mathbb{J}(\beta_0)) = 0 < r$. One of our assumptions is violated and our approach does not give the tangent cone at this point. This point is a singularity point which is another non-standard case that we have not studied here. □

Now assume that the true parameter point is $\beta_0 = (1, 1)$, a boundary point. Then $\mathcal{A} = (3 \quad -2)^T$ and $T_{\Theta_0}(\theta) = \{\beta \in \mathbb{R}^2 : 3\beta_1 - 2\beta_2 = 0\}$ and $T_{\Theta}(\theta) = \{\beta \in \mathbb{R}^2 : 3\beta_1 - 2\beta_2 \geq 0\}$

The suggested composite likelihood ratio test is applied on this case. The distribution of $\bar{\chi}^2$ is obtained as

$$P(\bar{\chi}^2 \leq c) = 0.5P(0.8232\chi_1^2 + 0.8185\chi_1^2 \leq c) + 0.5P(0.8203\chi_1^2 \leq c)$$

approximating into one term results in

$$P(\bar{\chi}^2 \leq c) = 0.5P(0.8209\chi_{1.999}^2 \leq c) + 0.5P(0.8203\chi_1^2 \leq c)$$

The %95 quantile is 4.217.

5.2 Simulation: Mixed Models

Next, we examined the performance of the proposed approach for testing some different hypothesis tests through simulation studies with regard to the maintenance of the nominal level of type I error $\alpha = 1\%, 5\%$ under the null hypothesis and the powers under specific alternatives. In total, the results of four approaches are compared: the asymptotic full likelihood (LRT), the asymptotic composite likelihood (CLRT), the naive likelihood ratio tests (F-naive , CL-naive), and a conditional method (cond-LRT, cond-CLRT). In the naive likelihood test, the boundary problem is ignored and it is treated as classical hypothesis testing with interior parameters. The threshold of the naive method is considered as $\chi^2_{k-r,\alpha}$. The conditional method is based on Susko [46] work, that threshold is $\chi^2_{d,\alpha}$, where d is the number of parameters which lie on the interior of the parameter space. Since conditional method is basically developed for the cases with interior nuisance parameter, a modification is applied for the tests with nuisance parameters on the boundary. For these cases the threshold is $\chi^2_{d',\alpha}$, where d' is the difference between number the interior parameters of the whole parameter space and the null parameter space.

Data is generated from linear mixed model $y_i = X_i\beta + zb + \epsilon$, $i = 1, \dots, n$, where y_i 's are vectors of length m, β is a $p \times 1$ fixed effect parameter, z is a $m \times s$ matrix and b denote the random effect vector where $b \sim N_s(0, \Sigma_b)$ and $\Sigma_b = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_s^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I_s)$. Then $\Sigma_y = z\Sigma_b z^T + \sigma_\epsilon^2 I_m$. For estimating the parameters we used a reparametrization suggested by Hartley and Rao [18] such that $\Sigma_y = \sigma_\epsilon^2(\sum_{i=1}^s \lambda_i z_i z_i^T + I_s)$, where $\lambda_i = \sigma_i^2/\sigma_\epsilon^2$ and $z_i, i = 1, \dots, s$ are the columns of z. Then using Newton-Raphson's method parameters through full likelihood and bivariate composite likelihood are estimated.

The bivariate composite likelihood function is

$$\begin{aligned}
CL &= \prod_{i=1}^{n} \prod_{1\leq j<l\leq m} f(y_{ij}, y_{il}) \\
\log CL &= cl = \frac{-nm(m-1)}{2}\log 2\pi - \frac{nm(m-1)}{2}\log \sigma_\epsilon^2 - \frac{n}{2}\sum_{j<l}\log|\Phi_{jk}| \\
&- \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\sum_{j<l}\begin{pmatrix} y_{ij}-\mu_{ij} & y_{il}-\mu_{il}\end{pmatrix}\Phi_{jl}^{-1}\begin{pmatrix} y_{ij}-\mu_{ij} \\ y_{il}-\mu_{il}\end{pmatrix}
\end{aligned}$$

where $\Phi = \sum_{i=1}^{s} \lambda_i z_i z_i^T + I_s$

To compute a bivariate composite likelihood, it is helpful to define a $2 \times m$ elimination matrix e_{jl} that the j element of the first row and l element of its second row is one and the rest are zero. Let

$$\begin{aligned}
\Phi_{jl} &= e_{jl}\Phi e_{jl}^T \\
Z_{jl}^m &= e_{jl}Z_m Z_m^T e_{jl}^T \\
E_{ijl} &= e_{jl}(y_i - \mu_i)
\end{aligned}$$

then the score vector can be written as

$$\begin{aligned}
\frac{\partial cl}{\partial \lambda_m} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j<l}\left[-tr(\Phi_{jl}^{-1}Z_{jl}^m) + \frac{1}{\sigma_\epsilon^2}E_{ijl}^T\Phi_{jl}^{-1}Z_{jl}^m\Phi_{jl}^{-1}E_{ijl}\right] \\
\frac{\partial cl}{\partial \sigma_\epsilon^2} &= \sum_{i=1}^{n}\sum_{j<l}\left[\frac{-1}{\sigma_\epsilon^2} + \frac{1}{2\sigma_\epsilon^4}E_{ijl}^T\Phi_{jl}^{-1}E_{ijl}\right]
\end{aligned}$$

and hessian matrix is obtained by

$$
\begin{aligned}
\frac{\partial^2 cl}{\partial\lambda_n\partial\lambda_m} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j<l}\big[tr(\Phi_{jl}^{-1}Z_{jl}^{n}Z_{jl}^{m}\Phi_{jl}^{-1}) \\
&- \frac{1}{\sigma_\epsilon^2}E_{ijl}^{T}\Phi_{jl}^{-1}Z_{jl}^{m}\Phi_{jl}^{-1}Z_{jl}^{n}\Phi_{jl}^{-1}E_{ijl} - \frac{1}{\sigma_\epsilon^2}E_{ijl}^{T}\Phi_{jl}^{-1}Z_{jl}^{n}\Phi_{jl}^{-1}Z_{jl}^{m}\Phi_{jl}^{-1}E_{ijl}\big] \\
\frac{\partial^2 cl}{\partial\sigma_\epsilon^2\partial\lambda_l} &= \frac{-1}{2\sigma_\epsilon^4}\sum_{i=1}^{n}\sum_{j<l}\big[E_{ijl}^{T}\Phi_{jl}^{-1}Z_{jl}^{m}\Phi_{jl}^{-1}E_{ijl}\big] \\
\frac{\partial^2 cl}{\partial\sigma_\epsilon^4} &= \sum_{i=1}^{n}\sum_{j<l}\big[\frac{1}{2\sigma_\epsilon^4} - \frac{1}{\sigma_\epsilon^6}E_{ijl}^{T}\Phi_{jl}^{-1}E_{ijl}\big]
\end{aligned}
$$

Four different hypothesis tests are designed. All simulations were performed using R (R Development Core Team (2013)). In each setting, the family size $m = 5$, $p = 2$, $\beta = (1\ 1)^T$ and $\sigma_\epsilon = 1$. The parameter β is treated as known. The type I error is calculated for number of families $n = 70, 100$ and power test is done for $n = 100$. In the first two tests, to evaluate the power of each method, we consider two different alternative scenarios: in first configuration only two non-zero parameters are available with five different effect sizes $0.05, 0.1, 0.2, 0.3, 0.5$ and a second alternative configuration with eight non-zero parameters and five other effect sizes $0.01, 0.03, 0.06, 0.09, 0.1$. In the third test, power is computed when four non-zero parameters are available with effect sizes $0.01, 0.03, 0.06, 0.09$ and 0.1. Power of the Test 4, is computed when the first three element of the parameters are equal and non-zero and with a value of $0.05, 0.1, 0.2, 0.3, 0.5$.

- **Test 1:** only parameters of interest lie on the boundary (results in 5.4),
- **Test 2:** there is one parameter that is not tested but may lie on the boundary (results in 5.5),

- **Test 3:** there are three parameters that is not tested but may lie on the boundary (results in 5.6),
- **Test 4:** hypothesis testing with nonlinear constraints (results in 5.7).

Test 1: Let $k = 8$ and $H_0 : \sigma_i^2 = 0$ versus $H_a : \sigma_i^2 \geq 0, i = 1, \ldots, 7$, and parameter σ_ϵ^2 is an interior parameters. Then $\Theta_0 = \{\sigma \in \mathbb{R}^8 : g_i(\sigma) = \sigma_i = 0, i = 1, \ldots, 7, g_8(\sigma) = \sigma_\epsilon \geq 0\}$ and $\Theta = \{\sigma \in \mathbb{R}^8 : g_i(\sigma) = \sigma_i \geq 0, i = 1, \ldots, 7, g_8(\sigma) = \sigma_\epsilon \geq 0\}$. The tangent cone can be shown as $T_{\Theta_0} = \{0\}^7$ and $T_\Theta = [0, \infty)^7$. The test statistic is similar to (4.20) in the Case 3. Simulation result is shown in Table 5.4.

Test 2: Let $k = 11$, and $H_0 : \sigma_i^2 = 0$ versus $H_0 : \sigma_i^2 \geq 0, i = 1, \ldots, 8$, while σ_9^2 may lie on the boundary and parameters σ_{10}^2 and σ_ϵ^2 are interiors. Therefore $\Theta_0 = \{\sigma^2 \in \mathbb{R}^{11} : g_i(\sigma^2) = \sigma_i^2 = 0, i = 1, \ldots, 8, g_j(\sigma^2) = \sigma_j^2 \geq 0, j = 9, 10, \epsilon\}$ and $\Theta = \{\sigma^2 \in \mathbb{R}^{11} : g_i(\sigma^2) = \sigma_i^2 \geq 0, i = 1, \ldots, 10, \epsilon\}$.

Then the null tangent cone becomes $T_{\Theta_0} = \{\sigma^2 \in \mathbb{R}^{11} : \mathcal{A}_{01}\sigma = 0, \mathcal{A}_{02}\sigma \geq 0\} = \{0\}^8 \times [0, \infty)$, where $\mathcal{A}_{01}$ is a 8×11 zero matrix with an 8×8 identity matrix on its left side and $\mathcal{A}_{02}$ is a 1×11 zero matrix that its 9th element is one. The cone $T_\Theta = \{\sigma^2 \in \mathbb{R}^{11} : \mathcal{A}_1\sigma \geq 0\} = [0, \infty)^9$ is the total cone where $\mathcal{A}_1$ is a 9×11 zero matrix with an 9×9 identity matrix on its left side.

Similar to the Case 2, the test statistic becomes

$$\bar{\chi}^2 = \sum_{j=1}^{2} \sum_{i=1}^{2^9} \tilde{z}^T \left(U_{\mathcal{I}_{0j}}^{-T} (U_{\mathcal{I}_{0j}}^{-1} U_{\mathcal{I}_{0j}}^{-T})^{-1} U_{\mathcal{I}_{0j}}^{-1} - U_{\mathcal{I}_i}^{-T} (U_{\mathcal{I}_i}^{-1} U_{\mathcal{I}_i}^{-T})^{-1} U_{\mathcal{I}_i}^{-1} \right) \tilde{z} w_{ij} \tag{5.3}$$

Table 5.5 shows the type I error and power for this test.

Table 5.3: Type I error for two different sample sizes

	n	α	LRT	CLRT	F-naive	CL-naive	cond-LRT	cond-CLRT
				type I error				
Test 1	70	0.01	0.012	0.012	0.002	0.045	0.009	0.129
	100	0.01	0.010	0.014	0.003	0.030	0.010	0.138
	70	0.05	0.050	0.045	0.008	0.045	0.053	0.259
	100	0.05	0.052	0.051	0.004	0.096	0.047	0.267
Test 2	70	0.01	0.005	0.0113	0.002	0.028	0.01	0.086
	100	0.01	0.0048	0.011	0.0015	0.033	0.013	0.079
	70	0.05	0.051	0.046	0.001	0.051	0.053	0.168
	100	0.05	0.050	0.047	0.005	0.072	0.053	0.189

Table 5.4: Type I error and power for Test 1, $n = 100$, $\alpha = 0.05$, $B = 1000$ simulation

	σ_i	LRT	CLRT	F-naive	CL-naive	cond-LRT	cond-CLRT
2 nonzero variance	0	0.052	0.051	0.004	0.096	0.047	0.267
	0.05	0.178	0.099	0.054	0.142	0.126	0.383
	0.1	0.405	0.247	0.140	0.304	0.365	0.612
	0.2	0.879	0.684	0.681	0.792	0.857	0.927
	0.3	0.992	0.973	0.960	0.985	0.988	0.998
	0.5	1	1	1	1	1	1
8 nonzero variance	0	0.052	0.051	0.004	0.096	0.047	0.267
	0.01	0.153	0.085	0.016	0.164	0.089	0.346
	0.03	0.480	0.274	0.181	0.401	0.302	0.594
	0.06	0.898	0.716	0.692	0.825	0.758	0.830
	0.09	0.990	0.945	0.941	0.962	0.953	0.974
	0.1	0.996	0.947	0.977	0.967	0.981	0.991

Table 5.5: Type I error and power for Test 2, $n = 100$, $\alpha = 0.05$, $B = 1000$ simulation

	σ_i	LRT	CLRT	F-naive	CL-naive	cond-LRT	cond-CLRT
2 nonzero variance	0	0.050	0.047	0.005	0.072	0.053	0.189
	0.05	0.102	0.070	0.015	0.089	0.091	0.269
	0.1	0.204	0.163	0.037	0.181	0.186	0.421
	0.2	0.668	0.502	0.272	0.525	0.603	0.750
	0.3	0.912	0.856	0.691	0.878	0.897	0.946
	0.5	1	1	1	1	1	1
8 nonzero variance	0	0.050	0.047	0.005	0.072	0.053	0.189
	0.01	0.077	0.055	0.003	0.071	0.067	0.222
	0.03	0.183	0.162	0.029	0.177	0.141	0.349
	0.06	0.523	0.364	0.144	0.392	0.328	0.570
	0.09	0.772	0.675	0.420	0.695	0.636	0.787
	0.1	0.858	0.753	0.589	0.766	0.738	0.844

Test 3: Assume $k = 8$, and $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 0$ versus $H_a : \sigma_i^2 \geq 0, i = 1, \ldots, 4$. Parameters σ_5^2, σ_6^2 and σ_7^2 may lie on the boundary. $\Theta_0 = \{\sigma^2 \in \mathbb{R}^8 : g_i(\sigma^2) = \sigma_i^2 = 0, i = 1, \ldots, 4, g_j(\sigma^2) = \sigma_j^2 \geq 0, j = 5, 6, 7, \epsilon\}$ and $\Theta = \{\sigma^2 \in \mathbb{R}^8 : g_i(\sigma^2) = \sigma_i^2 \geq 0, i = 1, \ldots, 7, \epsilon\}$

Then tangent cone can be expressed as $T_{\Theta_0} = \{\sigma^2 \in \mathbb{R}^8 : \mathcal{A}_{01}\sigma^2 = 0, \mathcal{A}_{02}\sigma^2 \geq 0\} = \{0\}^4 \times [0, \infty)^3$, where $\mathcal{A}_{01}$ is a 3×8 zero matrix with an 3×3 identity matrix on its left side and $\mathcal{A}_{02}$ is a 4×8 zero matrix with an 4×4 identity matrix on its right side. $T_{\Theta_0} = \{\sigma^2 \in \mathbb{R}^8 : I_{7\times 7}\, \sigma^2 \geq 0\} = [0, \infty)^7$.

This test can be performed using the statistic in Case 2, while $i = 1, \ldots, 2^7$ and $j = 1, \ldots, 2^3$. Table 5.6 contains the simulation results related to this test.

Table 5.6: Type I error and power for Test 3, $n = 100$, $\alpha = 0.05$, $B = 1000$ simulation

4 nonzero variance	LRT	CLRT	F-naive	CL-naive	cond-LRT	cond-CLRT
0	0.0511	0.0507	0.001	0.270	0.047	0.156
0.01	0.143	0.130	0.006	0.101	0.120	0.296
0.03	0.587	0.482	0.147	0.379	0.466	0.576
0.06	0.960	0.955	0.781	0.959	0.916	0.962
0.09	1	1	0.977	1	0.995	0.997
0.1	1	1	1	1	1	1

Test 4 : Let $\sigma_b^2 = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$.One is interested to test

$$H_0 : \begin{cases} \sigma_1^2 - 2\sigma_2^2 = 0 \\ \sigma_1^2\sigma_3^2 - \sigma_4^4 = 0 \end{cases} \text{versus } H_a : \begin{cases} \sigma_1^2 - 2\sigma_2^2 \geq 0 \\ \sigma_1^2\sigma_3^2 - \sigma_4^4 \geq 0 \end{cases}$$

Then the parameter space for $\sigma^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$ is given by

$\Theta = \{\sigma^2 \in \mathbb{R}^k : g_1(\sigma^2) = \sigma_1^2 - 2\sigma_2^2 \geq 0,\ g_2(\sigma^2) = \sigma_1^2\sigma_3^2 - \sigma_4^4 \geq 0\}$,

and under the null hypothesis

$\Theta_0 = \{\sigma^2 \in \mathbb{R}^k : g_1(\sigma^2) = \sigma_1^2 - 2\sigma_2^2 = 0,\ g_2(\sigma^2) = \sigma_1^2\sigma_3^2 - \sigma_4^4 = 0\}$.

Assume the true point be $\sigma_0^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (2, 1, 2, 2)$. Let $\mathbb{J}$ be the Jacobian of $g_i(\sigma^2), i = 1, 2$ then

$$\mathbb{J}^T = \begin{pmatrix} 1 & -2 & 0 & 0 \\ \sigma_3^2 & 0 & \sigma_1^2 & -2\sigma_2^2 \end{pmatrix}$$

Therefore, the tangent cones can be written as $T_{\Theta_0}(\sigma_0^2) = \left\{\sigma^2 \in \mathbb{R}^4 : \mathcal{A}\sigma^2 = 0\right\}$, and $T_{\Theta}(\sigma_0^2) =$

$\{\sigma^2 \in \mathbb{R}^4 : \mathcal{A}\sigma^2 \geq 0\}$, where

$$\mathcal{A} = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 2 & 0 & 2 & -4 \end{pmatrix}$$

Then the test statistic for Test 4 becomes similar to (4.18). Table 5.7 shows the type I error rate and power of the test when three elements are non-zero.

Table 5.7: Type I error and power for Test 4, $n = 100$, $\alpha = 0.05$, $B = 1500$ simulation

3 nonzero variance	LRT	CLRT	F-naive	CL-naive	cond-LRT	cond-CLRT
0	0.050	0.051	0.136	0.575	0.819	0.592
0.05	0.199	0.074	0.337	0.588	0.875	0.615
0.1	0.639	0.144	0.783	0.698	0.930	0.692
0.2	0.995	0.457	0.997	0.927	0.996	0.874
0.3	1	0.815	1	0.989	0.999	0.956
0.5	1	0.995	1	0.999	1	0.997

□

Time of the algorithm depends on the number of parameters of interest, existing of the boundary parameters even if are not included in the hypothesis testing, and sample size. In general, average times of full likelihood ratio testing is less than composite likelihood version. In the whole process, the time is mostly spent on the step of estimating the parameters. This is because full likelihood requires less computation compared with a bivariate composite likelihood, even though it might need more iteration to converge sometimes.

In Table 5.8, the run time of the algorithm in for some different case discussed in simulation part is reported. The reported times are for mixed model. This time is shorter for a simple normal regression model.

Table 5.8: run time in seconds

	k	n	LRT	CLRT
Test 1	8	70	20.05	67.42
	8	100	32.11	97.79
Test 2	11	70	36.69	76.50
	10	100	48.86	102.90
Test 3	7	70	26.82	128.39
	7	100	38.24	182.21
Test 4	4	70	3.67	9.25
	4	100	5.85	12.22

In addition, the algorithm converges in a reasonable number of iterations. The Figure 5.3 show the convergence for different number of iterations for the four types of test in the simulation studies.

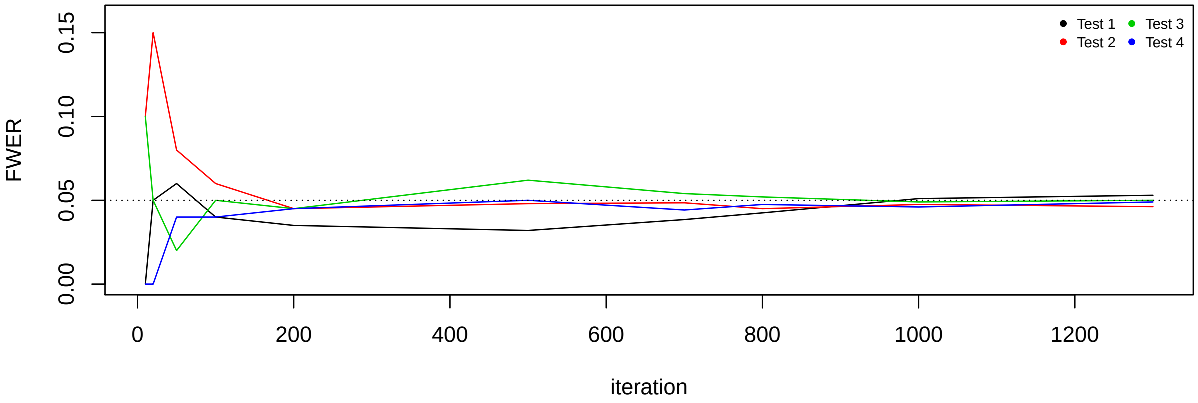


Figure 5.3: convergence of algorithm for different number of iterations

The Tables 5.4, 5.5 and 5.6 show that the three approaches likelihood ratio, composite likelihood ratio and the conditional tests maitain the nominal level for type I error rate. As it is expected full likelihood has the highest power and there is a bit of loss of power in composite likelihood approach. The naive methods which ignore the boundary problem cannot provide a reasonable error rate and also the conditional approach for composite likelihood is not a suitable choice.

The simulation result suggests that the proposed method works well for all cases of likelihood ratio tests under study. Although Susko's conditional method works well for the full likelihood case when the nuisance parameters are interior point, with only a little power loss and it is an easier approach, it does not work properly in the situations that composite likelihood is the more suitable option. Because in full likelihood case, the distribution of $\inf_{\theta}(\tilde{z} - \tilde{\theta})^T I(\theta_0)(\tilde{z} - \tilde{\theta})$ conditioned on the location of $\tilde{z}$, is exactly chi-square where the degree of freedom is the dimension of the region that $\tilde{z}$ is located on it, while in composite likelihood, $\tilde{z}$ is not a standard normal random variable and the conditional distribution of $\inf_{\theta}(\tilde{z} - \tilde{\theta})^T H(\theta_0)(\tilde{z} - \tilde{\theta})$ follows a mixture of chi-squares and not only a chi-square. The misspecified covariance $H(\theta_0)$ in composite likelihood causes the test statistic becomes larger. Also, the naive method does not provide an accurate result. Full likelihood ratio test with the naive threshold is very conservative with low power. However, composite likelihood version of the test leads to high type I error rate, which is again due to the larger value of the test statistic with the $H(\theta_0)$ matrix.

5.3 Discussion

Composite likelihood ratio test (CLRT) is an alternative to the full likelihood approach and follows a mixture of chi-square distribution with one degree of freedom. But this result is not valid when the true parameter lies on the boundary of the parameter space which is not a rare situation. Ignoring the boundary problem and treating it as standard cases leads in false inferences.

The limiting distribution of CLRT in this situation is studied and it is shown that CLRT at the boundary points follow a mixture of weighted sum of chi-square variables. Empirical study provides the results that the proposed approach gives a good estimation of distribution for different dimensions and linear and nonlinear combination of parameters in hypothesis tests.

Chapter 6

Future Work

Based on the assumption of the work and as it is shown in example 5.1.1, part (C), the proposed approach does not work when the boundary point is also a singularity. Drton [9] works on likelihood ratio test on singularity points. Regarding limitation of using full likelihood in many situations, it is useful to extend this study to composite likelihood ratio tests as well.

Here are two applied field that it is known that due to the nature of the response variable, usually composite likelihood is a more suitable option to estimate the parameters. Moreover, despite having higher dimension parameters in reality, for hypothesis testing in literature, usually no more than $2-$dimensional cases are considered, or the parameters that are not involved in hypothesis testing directly, are ignored due to computational limitation. Using our proposed approach, the hypothesis test can be performed in higher dimensions.

6.1 Simultaneous linkage and heritability analysis in pedigree data

Suppose that the data contain N families or general pedigrees, with m_i relatives in the i^{th} pedigree and $y_i = (y_{i1}, y_{i2}, \dots, y_{im_i})$ where y_{ij} is the trait value of the j^{th} individual in a family i. Traits could follow different kinds of distributions such as normal, gamma, binary and so on. Let μ denote the mean of the traits. For a link function $g(.)$,

$$g(\mu_{ij}) = x_{ik}\beta + z_{ijg}G + z_{iju}U + z_{ijp}P + e_{ij}$$

the vector $\beta = (\beta_1, \dots, \beta_r)$ is the coefficient vector of fixed effects at the individual, and the rest of the coefficients are the vector of random effects at the pedigree level. The constants G and P are the genetic effect of one disease allele and the polygenic effect, respectively. $U_{1\times s}$ shows the effect of s covariates at the pedigree level on the trait Under the trait model, we have r coefficients of fixed effects β at the first level, s variances $\sigma_u^2 = (\sigma_{u1}^2, \sigma_{u2}^2, \dots, \sigma_{us}^2)$ and the two variances σ_p^2 and σ_g^2 at the second level, and σ_e^2.

The parameters are $\theta = (\beta, G, U, P, \sigma_{u1}^2, \dots, \sigma_{us}^2, \sigma_p^2, \sigma_g^2, \sigma_e^2)$. Here it is assumed that the traits follow normal distribution and the link function is identity. Then it can be put in a generalized mixed model framework, where $y_i \sim N_{m_i}(x_i\beta, V_k)$. The variance-covariance matrix for pedigree $k(k = 1, 2, \dots, N)$ is given by $V_k = z_k\Omega z_k^T + \sigma_e^2 I$, where the entries in

z_k are defined based on linkage analysis [52], and I is an identity matrix and

$$\Omega = \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_p^2 & 0 \\ 0 & 0 & \sigma_g^2 \end{pmatrix}$$

We can perform likelihood ratio test for testing linkage and heritability$H_0 : \sigma_p^2 = 0,\ \sigma_g^2 = 0$ against $H_a : \sigma_p^2 \geq 0,\ \sigma_g^2 \geq 0$ while the first q elements of the nuisance parameter $(\sigma_{u1}^2, \ldots, \sigma_{uq}^2, \sigma_{u(q+1)}^2, \ldots, \sigma_{us}^2)$ may lie on the boundary. Considering $\theta = (\sigma_p^2,\ \sigma_g^2, \sigma_{u1}^2, \ldots, \sigma_{uq}^2)$, then $T_{\Theta_0} = \{0\}^2 \times [0, \infty)^q$ and $T_\Theta = [0, \infty)^{q+2}$. Applying Case 2 to find the test statistic, and finding the weights through finding possible intersections between the 2^q faces of T_{Θ_0} and 2^{q+2} faces of $T_\Theta(\theta)$.

6.2 Ferromagnetic Ising model

Here we focus on a type of this model developed in statistical physics to model ferromagnetism. Consider n atoms in the presence of a 2-directed magnetic field of strength h (that forms a 2-dimensional lattice. In general, we could have any dimension). The local magnetic moment of each atom is represented by a spin, and the model supposes that spin of each atom of a ferromagnet, interact with its neighbors. The spins has just two possible states described by $y_i = \pm 1,\ i = 1, \ldots, n$, which means spin is either pointing up or pointing down. Let J_{ij} denote a coefficient giving the interaction strength and h_i be the effect of the magnetic field on each spin i, then the association between spins is modelled by

$$P(\mathbf{y}) = \frac{1}{Z}\exp\{-\beta\left(\sum_{i=1}^{n} h_i y_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} J_{ij} y_i y_j\right)\}$$

which has a quadratic exponential form.

The case $J_{ij} \geq 0$ is called the ferromagnetic Ising model. The parameter $\beta \geq 0$ is the inverse of the temperature. The normalizing constant Z is called partition function. Derivation of the partition function is difficult. As the model is quadratic exponential, using a conditional composite likelihood can help to get rid of the partition function and estimate the rest of the parameters.

Let define the set of parameters $\theta = \{h, J, \beta\}$, where $h = \{h_1, \ldots, h_n\}$, $J = \{J_{ij}, 1 \leq i < j \leq n\}$. We wish to test $H_0 : J = 0$ versus $H_a : J \geq 0$. The null hypothesis means that the model in non-interacting. The parameters $h \in \mathbb{R}$ and $\beta \geq 0$ are the nuisance, where the latter may lie on the boundary. Therefore, there is boundary issue in both parameters of interest and the nuisance ones. The tangent cones are $T_{\Theta_0} = \{0\}^{\frac{n(n-1)}{2}}$ and $T_{\Theta} = [0, \infty)^{\frac{n(n-1)}{2}}$, and the CLRT test statistic becomes

$$\|\tilde{y}\|^2 - \inf_{\tilde{h} \in \tilde{T}_{\Theta}} \|\tilde{y} - \tilde{h}\|^2$$

where $\tilde{T}_{\Theta} = UT_{\Theta} = \{UJ | J \in [0, \infty)^3\}$.

We can use the result of estimating parameters in quadratic exponential model by composite likelihood and then the suggested approach for composite likelihood ratio test enables us to draw the distribution of different kind of hypothesis tests on elements of J. Details of computation for $n = 3$ case is provided in the appendix E.

Appendix A

Proof of asymptotic distribution of composite likelihood estimator for cluster data

Proof of theorem 1.3.1

Proof. The proof is divided into two main steps. We first show that there exists a $\widehat{\theta}_n^c$ which is of order $O(n^{-1/2})$, and then we derive its asymptotic normality.

Let $h(\theta; y) = cl(\theta; y)$. Note that for fixed y, h maps $\mathbb{R}^p$ into $\mathbb{R}$. Then, by a Taylor expansion, we have that

$$h(\theta; y) - h(\theta_0; y) \quad = \quad (\nabla h)(\theta_0; y)^T(\theta - \theta_0) + (\theta - \theta_0)^T (Dh)(\theta^*; y)(\theta - \theta_0),$$

where θ^* lies on a line joining θ and θ_0. We use ∇, D to denote the gradient and Hessian operators, respectively. Our goal will be to show that there exists a θ in a $n^{-1/2}$ ball of θ_0, the left hand side of the above equation is negative. This in turn will imply that there exists a MCLE which satisfies $\sqrt{n}(\widehat{\theta}_n^c - \theta_0) = O_p(1)$.

To this end, let $\theta - \theta_0 = \xi M / \sqrt{n}$, with $||\xi||_2 = 1$. Assume also that $||\theta - \theta_0||_2 < c$, that is, $M < c\sqrt{n}$. Then, by the above, we have

$$\begin{aligned} & \xi^T \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\nabla cl_m)(\theta_0, y_i) \right\} + \xi^T \left\{ \frac{1}{n} \sum_{i=1}^{n} (Dcl_m)(\theta^*, y_i) \right\} \xi \\ \equiv \quad & \xi^T b_n M + \xi^T B_n \xi M^2, \end{aligned} \tag{A.1}$$

where b_n is a random vector converging to a mean-zero Gaussian RV, and B_n is the random matrix converging to the negative definite matrix $-H(\theta_0)$. The first of these follows by the central limit theorem, along with assumption (A5). The second follows by applying the law of large numbers, along with assumptions (A4) and (A6). Note that the second fact implies also that the eigenvalues of B_n converge almost surely to the eigenvalues of $-H(\theta_0)$.

Let $\lambda_n^{(p)}$ denote the largest eigenvalue of $-B_n$, and let $S = \{\xi : ||\xi||_2 = 1\}$. Since b_n converges as a random Gaussian vector (with mean zero), and $\xi^T b_n$ is uniformly continuous on S, it follows that $\xi^T b_n$ converges to a mean-zero Gaussian process in $C(S)$, the space of continuous functions on S endowed with the uniform metric. This implies that $\xi^T b_n$ is tight in $C(S)$, and hence for all $\varepsilon > 0$, there exists an M_ε, such that

$$\limsup_n P\left(\sup_{\xi \in S} \xi^T b_n / \lambda_n^{(p)} < M_\varepsilon \right) \quad \geq \quad 1 - \varepsilon.$$

Then, by (A.1), if $\xi^T b_n/\lambda_n^{(p)} < M$, then $\xi^T b_n M + \xi^T B_n \xi M^2 < 0$, which in turn implies that

$$\limsup_n P\left(\xi^T b_n M_\varepsilon + \xi^T B_n \xi M_\varepsilon^2 < 0 \;\; \forall \xi \in S\right) \;\; \geq \;\; 1-\varepsilon.$$

Note that if $\xi^T b_n M_\varepsilon + \xi^T B_n \xi M_\varepsilon^2 < 0 \;\; \forall \xi \in S$, then, by the above and continuity of cl_m, this implies that for sufficiently large n, (with a probability of at least $1-\varepsilon$) there exists at least one local maximum on the set $B_{M_\varepsilon/\sqrt{n}}(\theta_0) \cap B_c(\theta_0)$. This implies that there exists a $\widehat{\theta}_n^c$ which satisfies $\sqrt{n}(\widehat{\theta}_n^c - \theta_0) = O_p(1)$.

Using a multivariate Taylor expansion, we have that

$$\begin{aligned}\nabla cl_m(\widehat{\theta}_n^c; y) \;\; = \;\; & \nabla cl_m(\theta_0; y) + \sum_{|\alpha|=1} (\partial\alpha cl_m^{(1)})(\theta_0; y)(\widehat{\theta}_n^c - \theta_0)^\alpha \\ & + \sum_{|\alpha|=2} \frac{2}{\alpha!}(\widehat{\theta}_n^c - \theta_0)^\alpha \int_0^1 (1-t)(\partial\alpha cl_m^{(1)})(\theta_0 + t(\widehat{\theta}_n^c - \theta_0); y)dt,\end{aligned}$$

again using the multi-index notation. We take $\widehat{\theta}_n^c$ to be the local maximizer found above. This time, for fixed y, ∇cl_m maps $\mathbb{R}^p$ into $\mathbb{R}^p$, so we have chosen to bound the error term a little differently than above. We let $R_{n,i}$ denote the third term on the right hand side of this equation when y is replaced with y_i. Next, as by definition $\sum_{i=1}^n cl_m^{(1)}(\widehat{\theta}_n^c; y_i) = 0$, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Dcl_m)(\theta; y_i)^T(\widehat{\theta}_n^c - \theta_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^n R_{n,i} \;\; = \;\; \frac{1}{\sqrt{n}}\sum_{i=1}^n f(\theta_0; y_i). \qquad \text{(A.2)}$$

By condition (A4), we have that

$$\begin{aligned}
&\left| \sum_{|\alpha|=2} \frac{2}{\alpha!} (\widehat{\theta}_n^c - \theta_0)^\alpha \int_0^1 (1-t)(D^\alpha cl_m^{(1)})(\theta_0 + t(\widehat{\theta}_n^c - \theta_0); y) dt \right| \\
&\qquad \leq \sum_{|\alpha|=2} \frac{1}{\alpha!} |\widehat{\theta}_n^c - \theta_0|^\alpha |M(y)|,
\end{aligned}$$

from which it follows that,

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n R_{n,i} \right| \leq \left\{ \sqrt{n} || \widehat{\theta}_n^c - \theta_0 ||_2^2 \right\} \left\{ \frac{1}{n} \sum_{i=1}^n |M(y_i)| \right\}.$$

The first term is then $o_p(1)$ by the first part of this proof, and by the law of large numbers (since $M(y)$ is integrable), the second term is $O_p(1)$. Next, consider

$$\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n cl_m^{(2)}(\theta; y_i) - H(\theta_0) \right\} (\widehat{\theta}_n^c - \theta_0).$$

By similar argument to that above, this is also $o_p(1)$. This allows us to re-write (A.2) as

$$\sqrt{n} H(\theta_0)(\widehat{\theta}_n^c - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(\theta_0; y_i) + o_p(1)$$

A straightforward application of the central limit theorem shows that the term on the right hand side has a Gaussian limiting distribution with mean zero and variance $J(\theta_0)$. The full result follows. □

Appendix B

Some useful definitions and theorems

Theorem B.0.1. *(Graybill [17] theorem 4.4.4) Let $y \sim N_p(\mu, \Sigma)$ where Σ has rank n. The limiting distribution of the random variable $y^T Ay$ is the same as $V = \sum_{i=1}^{n} \lambda_i V_i$, that $\lambda_i$are the eigenvalue of the matrix $A\Sigma$ and $V_1, \ldots, V_n$ are independent non-central chi-square variables with one degree of freedom.*

hyperplane: a hyperplane H in $\mathbb{R}^n$ is defined by $H = \{y; a^T Y = b\},\ a \in \mathbb{R}^n,\ a \neq 0, b \in \mathbb{R}$.

interior point: A point $\theta \in \Theta$ is called an interior point if there is a small neighbourhood centred at θ that lies entirely in Θ.

boundary point: A point $\theta \in \Theta$ is called a boundary point if any small neighbourhood centred at θ has non-empty intersection with both Θ and its complement.

open and closed space: the space Θ is said to be open if any point in Θ is an interior point and it is closed if its boundary is contained in Θ.

closure :The closure of a set is the smallest closed set containing that set. In other words, the closure of Θ is the union of Θ and its boundary.

limiting point : A point x is a limit point of a set A if $\forall \delta > 0, (x-\delta, x+\delta) \cup A\{x\} \neq \emptyset$. In other words, a limit point is a point that has points around it of arbitrary closeness, so one can make a sequence of distinct points that converges to limit point. Statistically speaking, x is a limit point of a sequence x_n of distributions if there exist a subsequence x_{n_j} that converges in distribution to x.

tight: The family Π is said to be tight, if for all $\epsilon > 0$, there is a compact set K such that $\mu(K) > 1 - \epsilon$ for all $\mu \in \Pi$.

relatively compact: A family of probability measures Π is relatively compact if every sequence $\mu_n \in \Pi$ has a subsequence μ_{n_j} such that $\mu_{n_j} \to \mu$.

Prohorov's theorem: A sequence $\{\mu_n\}$ of probability measures on $(\mathbb{R}^n, \mathbb{B})$ is tight if and only if it is relatively compact.

Mahalanobis distance: The Mahalanobis distance of a point $x = (x_1, x_2, \ldots, x_r)^T$ from a set of points with mean $\mu = (\mu_1, \mu_2, \ldots, \mu_r)^T$ and covariance matrix Σ is defined as:

$$\Big((x-\mu)^T \Sigma^{-1} (x-\mu)\Big)^{1/2}$$

chi-bar squared distribution: In non-standard conditions, a distribution that may raise naturally is a mixture of chi-squares that is called chi-bar square distribution, which is in fact the convex combination (or wighted mean) of tail probabilities of chi-square random variables with possible degrees of freedom.

Let $C_\theta \subseteq \mathbb{R}^p$ be a closed convex cone and $z \sim N_p(0, \Sigma)$. Then the random variable

$\bar{\chi}^2(\Sigma, C)$ has chi-bar squared distribution, which has the same distribution as $z^T\Sigma^{-1}z - \inf_{\theta\in C}(z-\theta)^T\Sigma^{-1}(z-\theta)$. Then the chi-bar random variable is written as

$$\bar{\chi}^2(\Sigma, C) = z^T\Sigma^{-1}z - \inf_{\theta\in C}(z-\theta)^T\Sigma^{-1}(z-\theta)$$

Appendix C

Proofs of theorems and lemmas

Proof of lemma (4.4.1)

Proof. By the regularity condition, there exist a point-wise Taylor expansion of the log composite likelihood function of f_θ around the point θ_0,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\log f_\theta &= \frac{1}{n}\sum_{i=1}^{n}\log f_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n}(\theta-\theta_0)^T\frac{\partial \log f_{\theta_0}}{\partial\theta} \\
&+ \frac{1}{2n}\sum_{i=1}^{n}(\theta-\theta_0)^T\frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}(\theta-\theta_0) \\
&+ \frac{1}{2n}\sum_{i=1}^{n}(\theta-\theta_0)^T\left(\frac{\partial^2 \log f_{\theta^*}}{\partial\theta^2}-\frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}\right)(\theta-\theta_0)
\end{aligned}$$

that θ^* lies on the line connecting θ and θ_0 in ϕ.

Assume $|\theta-\theta_0|<\delta$, then by $(B3)$,

$$\sup_{|\theta-\theta_0|<\delta}|\frac{\partial^2 \log f_{\theta^*}}{\partial\theta^2}-\frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}| \le |\frac{\partial^2 \log f_{\theta^*}}{\partial\theta^2}|+|\frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}| \le 2M(y)$$

By dominated convergence theorem, $\left(\frac{\partial^2 \log f_{\theta *}}{\partial\theta^2} - \frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}\right)$ converges to its expectation and $E(|\frac{\partial^2 \log f_{\theta *}}{\partial\theta^2} - \frac{\partial^2 \log f_{\theta_0}}{\partial\theta^2}|)$ converges to zero as $\delta \to 0$. Then by $(B4)$ and $(B5)$ the proof is completed. □

Proof of lemma (4.4.2)

Proof. Since θ_0 is a limiting point of Θ, for each $\epsilon > 0$ and $c_\epsilon \to 0$, there exist $\theta \in \phi$ such that $\|\theta_0 - \theta\|_2 < c_\epsilon$.

For $\delta > 0$ let $N_\delta(\theta_0) = \{\theta \in \Theta | \ \|\theta - \theta_0\|_2 < \delta\}$ be a neighbourhood around θ_0. Let $cl(N_\delta(\theta_0))$ denote the closure of $N_\delta(\theta_0)$. It is assumed that $\phi = cl(N_\delta(\theta_0)) \cap \Theta$ is a closed set. Therefore for each n, a local maximum, $\widehat{\theta}_n$, exists in this closed set.

We need to show that there exist a sequence $\widehat{\theta}_n^c$, converging to θ_0, in the intersection of the ball around θ_0 with radius $\frac{M}{\sqrt{n}}$ and Θ. Assume $c_\epsilon = \psi\frac{M_\epsilon}{\sqrt{n}}$ and $\theta - \theta_0 = \psi\frac{M_\epsilon}{\sqrt{n}} = c$ with $\|\psi\|_2 = 1$.

Taylor expansion of the log composite likelihood around θ_0, is

$$\begin{aligned}\sum_{i=1}^{n} \log f(\theta, y_i) - \sum_{i=1}^{n} \log f(\theta_0, y_i) &= \sqrt{n}(\theta - \theta_0)^T \sqrt{n} A_{n,\theta_0} \\ &+ \sqrt{n}(\theta - \theta_0)^T B_{n,\theta^*} \sqrt{n}(\theta - \theta_0) \qquad \text{(C.1)}\end{aligned}$$

As the sequence $\sqrt{n}A_{n,\theta_0}$ converging to normal distribution with zero mean and covariance $J(\theta_0)$, by Prohorov's theorem $\sqrt{n}A_{n,\theta_0}$ is uniformly tight. That is, for every $\epsilon > 0$, there exist an K_ϵ for which $\sup_n P(\sqrt{n}|A_{n,\theta_0}| \leq K_\epsilon) > 1 - \epsilon$.

And θ^* is a point between θ and θ_0. By (B4) and LLN, $B_{n,\theta^*} \to -H(\theta_0)$. So there exist a

K'_ϵ such that $|B_{n,\theta^*} + H(\theta_0)| < K'_\epsilon$.

Let λ_1 be the smallest eigenvalue of $H(\theta^*)$. Then , if $|A_{n,\theta_0}| \leq \frac{K_\epsilon}{\sqrt{n}}$

$$\begin{aligned}
\sqrt{n}(\theta-\theta_0)^T\sqrt{n}A_{n,\theta_0} &+ \sqrt{n}(\theta-\theta_0)^T B_{n,\theta^*}\sqrt{n}(\theta-\theta_0) \\
&= \psi^T M\sqrt{n}A_{n,\theta_0} + \psi^T(B_{n,\theta^*} + H(\theta^*))\psi M^2 - \psi^T H(\theta^*)\psi M^2 \\
&\leq \psi^T M K_\epsilon + K'_\epsilon M^2 - \lambda_1 M^2
\end{aligned}$$

So the left side of (C.1) that is $cl(\theta) - cl(\theta_0)$ is negative if $\psi^T K_\epsilon/(K'_\epsilon - \lambda_p) < M$. So for a proper amount of M and sufficiently large n, there exist a local maximum $\widehat{\theta}^c_n$ in $N_{\frac{M_n}{\sqrt{n}}}(\theta_0) \cap N_\delta(\theta_0) \cap \Theta_0$ that satisfies $\sqrt{n}(\widehat{\theta}^c_n - \theta_0) = O_p(1)$.

□

Proof of lemma (4.4.5)

Proof. Let $z_n = \sqrt{n}J^{-1/2}A_{n,\theta_0}$, where z_n is a standard normal random variable, then

$$\begin{aligned}
cl(\theta) - cl(\theta_0) &= \sum_{i=1}^{n}\left(\log f_\theta - \log f_{\theta_0}\right) \\
&= n(\theta-\theta_0)^T A_{n,\theta_0} - \frac{n}{2}(\theta-\theta_0)^T H(\theta-\theta_0) + o_p(1) \\
&= \sqrt{n}(\theta-\theta_0)^T J^{1/2} z_n + \frac{1}{2}\sqrt{n}(\theta-\theta_0)^T H\sqrt{n}(\theta-\theta_0) + o_p(1) \\
&= \frac{1}{2}\left(\sqrt{n}(\theta-\theta_0)^T - H^{-1}J^{1/2}z_n\right)^T H\left(\sqrt{n}(\theta-\theta_0)^T - H^{-1}J^{1/2}z_n\right) \\
&- z_n J^{1/2}H^{-1}J^{1/2}z_n + o_p(1) \qquad \text{(C.2)}
\end{aligned}$$

Let $k(\theta)$ is the first term of (C.2) and $r(\theta)$ be the last term of (C.2), then

$$0 \leq l(\widehat{\theta}_n^c) - l(\tilde{\theta}_n) = k(\widehat{\theta}_n^c) - k(\tilde{\theta}_n) + r(\widehat{\theta}_n^c) - r(\tilde{\theta}_n)$$

Since $k(\widehat{\theta}_n^c) - k(\tilde{\theta}_n) \leq 0$ then it is concluded that $|k(\widehat{\theta}_n^c) - k(\tilde{\theta}_n)| \leq r(\widehat{\theta}_n^c) - r(\tilde{\theta}_n)$,

$k(\widehat{\theta}_n^c) - k(\tilde{\theta}_n) = \frac{1}{2}n(\widehat{\theta}_n^c - \tilde{\theta}_n)^T H(\widehat{\theta}_n^c - \tilde{\theta}_n)^T \leq o_p(1)$

Since H is positive definite,

$$|\sqrt{n}(\widehat{\theta}_n - \tilde{\theta}_n)^T| = o_p(1)$$

and the proof is completed.

□

Proof of theorem (4.4.4)

Proof. Let $h_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ and $h_{n,0} = \sqrt{n}(\widehat{\theta}_{n,0} - \theta_0)$ be converging sequences in Θ and Θ_0, that converge to h and h_0 in $T_\Theta(\theta)$ and $T_{\Theta_0}(\theta)$, consequently. Then the composite likelihood ratio test statistic is

$$\begin{aligned}
\tilde{\lambda}_n &= -2\log\tilde{\Lambda}_n = -2\left(\sup_{\theta\in\Theta_0}\sum_{i=1}^{n}\log\frac{f_\theta}{f_{\theta_0}} - \sup_{\theta\in\Theta}\sum_{i=1}^{n}\log\frac{f_\theta}{f_{\theta_0}}\right)\\
&= 2\inf_{h_0\in T_{\Theta_0}}\sum_{i=1}^{n}\log\frac{f_{\theta_0+\frac{h_0}{\sqrt{n}}}}{f_{\theta_0}} - 2\inf_{h\in T_\Theta}\sum_{i=1}^{n}\log\frac{f_{\theta_0+\frac{h}{\sqrt{n}}}}{f_{\theta_0}}\\
&= 2\inf_{h_0\in T_{\Theta_0}}(\sqrt{n}h_0^T A_{n,\theta_0} - \frac{1}{2}h^T H h_0) - 2\inf_{h\in T_\Theta}(\sqrt{n}h^T A_{n,\theta_0} - \frac{1}{2}h^T H h) + o_p(1)\\
&= \inf_{h_0\in T_{\Theta_0}}\|\sqrt{n}H^{-1/2}A_{n,\theta_0} - H^{1/2}h_0\|^2 - \inf_{h\in T_\Theta}\|\sqrt{n}H^{-1/2}A_{n,\theta_0} - H^{1/2}h\|^2 + o_p(1)\\
&= \inf_{h_0\in T_{\Theta_0}}(z_n-h_0)^T H(z_n-h_0) - \inf_{h\in T_\Theta}(z_n-h)^T H(z_n-h) + o_p(1)
\end{aligned}$$

the third equality is written using (4.6). The random sequence $z_n = \sqrt{n}H^{-1}A_{n,\theta_0}$ that converges to a random variable z with normal distribution with zero mean and covariance matrix $H^{-1}JH^{-1}$. And the forth equality is from the complete square

$$\sqrt{n}h^T A_{n,\theta_0} - \frac{1}{2}h^T H h = \|\sqrt{n}H^{-1/2}A_{n,\theta_0} - H^{1/2}h\|^2 - nA_{n,\theta_0}^T H^{-1}A_{n,\theta_0}$$

Therefore the distribution of composite likelihood ratio test converges to the distribution of

$$\inf_{h\in T_{\Theta_0}}(z-h)^T H(z-h) - \inf_{h\in T_\Theta}(z-h)^T H(z-h).$$

which gives the squared Mahalanobis distance when the covariance matrix is misspecified. That is $Q_{T_\Theta}(z)$ is the H^{-1}-distance between z and T_θ while the true covariance matrix is $H^{-1}JH^{-1}$. □

Proof of proposition (4.4.3)

Proof. Consider the Taylor expansion of the first derivative of log composite likelihood function around θ_0,

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial \log f(\widehat{\theta}_n^c, y_i)}{\partial\theta} - \frac{\partial \log f(\theta_0, y_i)}{\partial\theta}\right) \\
= \quad & \sqrt{n}(\widehat{\theta}_n^c - \theta_0)^T \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k}\frac{\partial^2 \log f(\theta_0, y_i)}{\partial\theta_j\theta_k} \qquad \text{(C.3)} \\
+ \quad & \sqrt{n}\sum_{j,k}(\widehat{\theta}_{nj}^c - \theta_{0j})(\widehat{\theta}_{nk}^c - \theta_{0k})\sum_{l}\frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^3 \log f(\theta^*, y_i)}{\partial\theta_j\theta_k\theta_l} \\
= \quad & \sqrt{n}(\widehat{\theta}_n^c - \theta_0)B_{n,\theta_0} \\
+ \quad & \sqrt{n}\sum_{l}\frac{1}{2n}\sum_{i=1}^{n}(\widehat{\theta}_n^c - \theta_0)^T\left(\frac{\partial^3 \log f(\theta^*, y_i)}{\partial\theta_j\theta_k\theta_l}\right)_{j,k}(\widehat{\theta}_n^c - \theta_0) \\
\leq \quad & \sqrt{n}(\widehat{\theta}_n^c - \theta_0)B_{n,\theta_0} + \frac{p}{2}\frac{1}{n}\sum_{i=1}^{n}|M(y_i)|\sqrt{n}\|\widehat{\theta}_n^c - \theta_0\|_2^2 \\
= \quad & o_p(1) - \sqrt{n}(\widehat{\theta}_n^c - \theta_0)H(\theta_0) + \frac{p}{2}O_p(1)o_p(1) \qquad \text{(C.4)}
\end{aligned}$$

where the inequality is witten by $(B3)$. As $\widehat{\theta}_n^c$ is the local maximizer of $cl(\theta)$ in ϕ, $\sum_{i=1}^{n}\frac{\partial \log f(\widehat{\theta}_n^c, y_i)}{\partial\theta} = 0$. By $(B5)$, H^{-1} exist, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(\theta_0, y_i)}{\partial\theta}H^{-1}(\theta_0) \geq \sqrt{n}(\widehat{\theta}_n^c - \theta_0) + o_p(1) \qquad \text{(C.5)}$$

Since $\sqrt{n}(\widehat{\theta}_n^c - \theta_0)$ is bounded in probability, $\sqrt{n}(\widehat{\theta}_n^c - \theta_0) = H^{-1}(\theta_0)\sqrt{n}A_{n,\theta_0} + o_p(1)$, and the proof is completed.

□

Appendix D

Projection matrix on a relative interior set

Assume for $A = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$, $\mathbb{P} = \{\tilde{\theta} : A^T\tilde{\theta} \geq 0\}$ that equivalently be written as $\{A^{-T}\theta : \theta \geq 0\}$. The relative interior set of $\mathbb{P}$ is $ri\,(F_{\mathcal{I}}) = \{\tilde{\theta} : A_{\mathcal{I}}^T\tilde{\theta} = 0\} \cap \{\tilde{\theta} : A_{\mathcal{I}^c}^T\tilde{\theta} > 0\}$. Let $\Pi(z|\mathbb{P})$ denote the projection of point z onto the cone $\mathbb{P}$, then $\inf_{x\in\mathbb{P}} \|y - x\| = \|y - \Pi(y|\mathbb{P})\|$, that is the closest point on $\mathbb{P}$ to y. Note that $y - \Pi(y|ri(F_{\mathcal{I}}))$ is orthogonal to the cone that y is being projected onto and therefore to every point in $ri(F_{\mathcal{I}})$ as well as $\Pi(y|ri(F_{\mathcal{I}}))$. The relative interior set $ri\,(F_{\mathcal{I}})$ is spanned by the columns of $A_{\mathcal{I}'}^{-T}$, such that $A_{\mathcal{I}'}^{-T}\theta$ gives equivalent space as $A_{\mathcal{I}}^T\tilde{\theta}$ in relative interior set. Then

$$0 = A_{\mathcal{I}'}^{-1}(y - \Pi(y|ri(F_{\mathcal{I}}))) = A_{\mathcal{I}'}^{-1}(y - A_{\mathcal{I}'}^{-T}z))$$

Then $z = (A_{\mathcal{I}'}^{-1}A_{\mathcal{I}'}^{-T})^{-1}A_{\mathcal{I}'}^{-1}y$ and $\Pi(y|ri(F_{\mathcal{I}})) = A_{\mathcal{I}'}^{-T}(A_{\mathcal{I}'}^{-1}A_{\mathcal{I}'}^{-T})^{-1}A_{\mathcal{I}'}^{-1}y$. This gives a projection matrix onto a linear space spanned by the face $F_{\mathcal{I}}$. Therefore it can also be expressed as

$$\Pi(y|ri(F_{\mathcal{I}})) = A_{\mathcal{I}}^T (A_{\mathcal{I}} A_{\mathcal{I}}^T)^{-1} A_{\mathcal{I}} y. \tag{D.1}$$

Appendix E

Special case of ferromagnetic Ising model

fixed temperature case: Assume $n = 3$, so $y^T = (y_1, y_2, y_3)$ and the inverse temperature $\beta = 1$. In this case, nuisance are interior points,

$$f_{\mathbf{Y}}(\mathbf{y}) \;\; \propto \;\; \exp\left\{\sum_{i=1}^{3} h_i y_i + J_{12} y_1 y_2 + J_{13} y_1 y_3 + J_{23} y_2 y_3\right\}.$$

We wish to test $H_0 : J = 0$ versus $H_a : J \geq 0$. The parameters are $\theta = (h_1, h_2, h_3, J_{12}, J_{13}, J_{23})$ and $\Theta_0 = \mathbb{R}^3 \times \{0\}^3$ and $\Theta = \mathbb{R}^3 \times [0, \infty)^3$. Hence $T_{\Theta_0} = \{0\}^3$ and $T_\Theta(\theta) = [0, \infty)^3$. We need to find the distribution of $\bar{\chi}^2 = y^T H y - inf_{h \in T_\Theta} (y - h)^T H (y - h)$. This can be done using the Case 3.

The conditional probabilities are

$$p_{rs} = \frac{1}{1 + e^{-2[h_r + \sum_{i \neq r} J_{ri} y_i]}}, \qquad p_{rf} = \frac{1}{1 + e^{2[h_r + \sum_{i \neq r} J_{ri} y_i]}}$$

and the log-composite likelihood function is

$$cl = \log CL(h, \beta, J; \mathbf{y}) \;=\; \sum_{i=1}^{n} \left(I(y_i = 1) \log p_{is} + I(y_i = -1) \log p_{if} \right), \tag{E.1}$$

with score vector $U(\theta) = \frac{\partial cl}{\partial \theta}$ and the Hessian matrix $H(\theta) = \frac{\partial^2 cl}{\partial \theta^2}$ with elements,

$$\begin{aligned}
\frac{\partial cl}{\partial J_{ij}} &= 2[y_j I(y_i = 1)(1 - p_{is}) - y_j I(y_i = -1)(1 - p_{if}) \\
&+ y_i I(y_j = 1)(1 - p_{js}) - y_i I(y_j = -1)(1 - p_{jf})] \\
\frac{\partial^2 cl}{\partial J_{ij}^2} &= -4[I(y_i = 1) p_{is}(1 - p_{is}) + I(y_i = -1) p_{if}(1 - p_{if}) \\
&+ I(y_j = 1) p_{js}(1 - p_{js}) + I(y_j = -1) p_{jf}(1 - p_{jf})] \\
\frac{\partial^2 cl}{\partial J_{ij} \partial J_{ik}} &= -4 y_j y_k [I(y_i = 1) p_{is}(1 - p_{is}) + I(y_i = -1) p_{if}(1 - p_{if})]
\end{aligned}$$

In matrix H, for $i, j, k = 1, 2, 3$,

$$H_{i,i} = \frac{\partial^2 cl}{\partial J_{ij}^2} = -\frac{1}{\cosh\left(h_i + \sum_{k \neq i} \mathrm{J_{ik}\, y_k}\right)^2} - \frac{1}{\cosh\left(\mathrm{h_j} + \sum_{k \neq j} \mathrm{J_{jk}\, y_k}\right)^2}$$

$$H_{i,j} = \frac{\partial^2 cl}{\partial J_{ij} \partial J_{ik}} = -\frac{\mathrm{y_j\, y_k}}{\cosh\left(h_i + J_{ij}\, y_j + J_{ik}\, y_k\right)^2}$$

$$H_{j,k} = \frac{\partial^2 cl}{\partial J_{ij} \partial J_{kj}} = -\frac{y_i\, y_k}{\cosh\left(h_j + J_{ij}\, y_i + J_{kj}\, y_k\right)^2}$$

By Cholesky decomposition, we find the matrix U such that $H = U^T U$.

$$U = \begin{pmatrix} \sqrt{2}\,\sqrt{-\frac{1}{{T_1}^2} - \frac{1}{{T_2}^2}} & 0 & 0 \\ -\frac{\sqrt{2}\,y2\,y3}{{T_1}^2\,\sqrt{-\frac{1}{{T_1}^2}-\frac{1}{{T_2}^2}}} & \sqrt{-\frac{2\left({T_1}^2+{T_2}^2+{T_3}^2\right)}{{T_3}^2\left({T_1}^2+{T_2}^2\right)}} & 0 \\ -\frac{\sqrt{2}\,y1\,y3}{{T_2}^2\,\sqrt{-\frac{1}{{T_1}^2}-\frac{1}{{T_2}^2}}} & -\frac{\sqrt{2}\,y1\,y2\left({T_1}^2+{T_2}^2-{T_3}^2\right)}{{T_3}^2\left({T_1}^2+{T_2}^2\right)\sqrt{-\frac{{T_1}^2+{T_2}^2+{T_3}^2}{{T_3}^2\left({T_1}^2+{T_2}^2\right)}}} & 2\,\sqrt{2}\,\sqrt{-\frac{1}{{T_1}^2+{T_2}^2+{T_3}^2}} \end{pmatrix}$$

where

$$T_1 = \cosh(h_1 + J_{12}\,y_2 + J_{13}\,y_3)$$

$$T_2 = \cosh(h_2 + J_{12}\,y_1 + J_{23}\,y_3)$$

$$T_3 = \cosh(h_3 + J_{13}\,y_1 + J_{23}\,y_2)$$

Bibliography

[1] Andrews, D. W. K. (2000). Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space . *Econometrica* **68**, 2, 399–405.

[2] Azadbakhsh, M., Gao, X., Jankowski, H. (2015). Multiple Comparisons Using Composite Likelihood in Clustered Data. submitted.

[3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B* **36**, 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

[4] Bretz, F., Hothorn, T. and Westfall, P. (2010). *Multiple Comparisons Using R*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

[5] Chen, y. and Liang, K. y. (2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems *Biometrika* **97**, 603–620.

[6] Chernoff, H. (1954). On the distribution of the likelihood ratio *Annals of Mathematical Statistics* **25**, 573–578.

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

[8] Demidenko, E. (2001). Computational aspects of probit model. *Mathematical Communications* **6**, 233–247.

[9] Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics* **37**, 979–1012.

[10] Drton, M. (2011). Quantifying the failure of bootstrap likelihood ratio tests . *Biometrika* **98**,4, 919–934. https://doi.org/10.1093/biomet/asr033

[11] Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **64**, 657–680.

[12] Fisher, R. (1935). *The design of experiments*. Edinburgh and London: Oliver and Boyd.

[13] Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Annals of Mathematical Statistics* **40**, 224–250.

[14] Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association* **105**, 1531–1540. Supplementary materials available online.

[15] Geyer, C. (1994). On the Asymptotics of Constrained M-Estimation. *The Annals of Statistics* **22**,4 , 1993–2010.

[16] Geys, H., Molenberghs, G. and Ryan, L. M. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics - Theory and Methods* **26**, 2743–2767.

[17] Graybill, F.A. (1976). Theory and Application of the Linear Model. *Duxbury Press*.

[18] Hartley, H. o. and Rao, J.N.K. (1941). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93–108.

[19] Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1111.

[20] Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics. *Scandinavian Journal of Statistics* **21**, 289–357. With discussion and a reply by the authors.

[21] Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. New York: Willy.

[22] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

[23] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* **75**, 383–386.

[24] Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

[25] Hothorn, T., Bretz, F., Westfall, P. and Heiberger, R. M. (2008). multcomp: Simultaneous inference in general parametric models. *R package* .

[26] Konietschke, F., Bosiger, S., Brunner, E. and Hothorn, L. A. (2013). Are multiple contrast tests superior to the anova? *The International Journal of Biostatistics* **9**, 11.

[27] Konietschke, F., Hothorn, L. A. and Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics* **6**, 738–759.

[28] Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50**, 403–418.

[29] Li, Y. and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association* **101**, 591–603.

[30] Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical inference from stochastic processes (Ithaca, NY, 1987),Contemporary Mathematics* American Mathematical Society., Providence, RI, **80**, 221–239.

[31] Lokhorst, J., Venables, B., Turlach, B. and Maechler, M. (2012). r-cran-lasso2. `http://mloss.org/software/view/104/`.

[32] Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

[33] Molenberghs, G. and Ryan, L. M. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics* **10**, 279–300.

[34] Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics* **28** 295–302.

[35] Renal Association (2014). Information and Resources: Normal GFR. `http://www.renal.org/information-resources/the-uk-eckd-guide/normal-gfr`. [Online; accessed 7-Oct-2014].

[36] Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis* **44**, 649–667.

[37] Rockafellar,R.T. and Wets., R. J. (1998). Variational analysis . *New York, Springer*.

[38] Ronning, G.(1977). A simple scheme for generating multivariate gamma distributions with non-negative covariance matrix. *Technometrics* **19**, 179–183.

[39] Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika* 6, 309-316.

[40] Scheffe (1959). *The analysis of variance*. Wiley, New York.

[41] Self, S. G. and Liang, K. y.(1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.

[42] Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72**, 13344.

[43] Sidak, Z. (1968). On multivariate normal probabilities of rectangles: Their dependence on correlations. *Annals of Mathematical Statistics* **39**, 1425–1434.

[44] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

[45] Silvapulle, M.J., Sen, P.K. (2005). Constrained Statistical Inference: Order, Inequality, and Shape Constraints. *New York, Wiley*.

[46] Susko, E. (2013). Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika* **100**, 10191023.

[47] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.* **58**, 267–288.

[48] Tukey, J. W. (1953). *The problem of multiple comparisons*. Mimeoraphed monograph.

[49] Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis* **92**, 1–28.

[50] Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

[51] Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–528.

[52] Wang, T. and Elston, R.C. (2005). Two-level Haseman-Elston regression for general pedigree data analysis. *Genetic Epidemiology* 29, 12–22.

[53] Wilks, S. S. (1983). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* 9, 60–62.

[54] Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference* **141**, 3047–3054.

[55] Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics* **33**, 335–356.

[56] Zi. J(2010). Aspects of Composite Likelihood Inference. *PhD Thesis* http://hdl.handle.net/1807/26460