Multiple Comparisons using Composite Likelihood in Clustered Data

Mahdis Azadbakhsh, Xin Gao and Hanna Jankowski

York University

May 29, 2015

Abstract

We study the problem of multiple hypothesis testing for correlated clustered data. As the existing multiple comparison procedures based on maximum likelihood estimation could be computationally intensive, we propose to construct multiple comparison procedures based on composite likelihood method. The new test statistics account for the correlation structure within the clusters and are computationally convenient to compute. Simulation studies show that the 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 erratic performance. Using data arising from a depression study, we show how our composite likelihood approach makes an otherwise intractable analysis possible.

1 Introduction

The prevalence of depression in seniors estimated by the World Health Organization varies between 10% to 20% (Barua, 2011). Understanding the relationship between depression and other health factors can help prevent the disease and alleviate the symptoms. The health and retirement study (HRS) conducted by the University of Michigan is a longitudinal study which measured various aspects of health, retirement and aging, including the subject's depression status. In this study, seniors were measured every two years from 1994 to 2012. The objective of our analysis is to estimate the effect of several health factors known to be associated with depression status and compare the effect sizes of different factors. Multiple comparisons on the effect sizes will clarify the relative importance of different factors to the

disease. For example, the factor of being sleepless and the factor of smoking both are shown to attribute to the occurrence rate of depression. One might question whether or not they are equally important or one factor is more important than the other for development of the disease. Therefore, to fully understand the effects of the health factors, we perform all pairwise comparisons on their effect sizes.

The repeated binary measurements of depression status observed in this data set are correlated within individuals. These repeated measurements can be viewed as clustered data since they are recorded from the same experimental unit multiple times. Clustered data examples arise in many other situations, including measurements coming from siblings or same pedigrees, or measurements taken in close proximity to each other in spatial data. Ignoring existing correlations within clusters leads to invalid individual or multiple inferences.

When performing multiple comparisons in clustered data, one should therefore, take into account the correlation structure within the clusters. However, full likelihood analyses on such data often encounter computational challenges. For a repeated binary measurement data, the distribution can be described by multivariate probit or quadratic exponential. Evaluating the full likelihood of a multivariate probit model involves multi-dimensional integration, which quickly becomes computationally prohibitive. For the quadratic exponential model, the normalizing constant has to be computed through summation of all possible configurations of the clustered data, and here again computational intensity increases with the cluster size. We can avoid this computational burden by using a composite likelihood approach.

Composite likelihood methods are extensions of the likelihood method that project high-dimensional likelihood functions to low-dimensional ones (Cox and Reid, 2004, Lindsay, 1988). This dimension reduction is achieved by compounding valid marginal or conditional densities. It has been shown that, under regularity conditions, the composite likelihood estimator has desirable properties, such as consistency and asymptotic normality (Cox and Reid, 2004, Lindsay, 1988, Varin, 2008, Varin et al., 2011). This makes it an appealing alternative in inferential procedures. Furthermore, composite likelihood is more computationally convenient than full likelihood at a cost of some loss of efficiency. The magnitude of this loss depends on the dimension of the multivariate vector and its dependency structure. Composite likelihood methodology has been applied to numerous statistical problems (Geys et al., 1997, Renard et al., 2004, Zhao and Joe, 2005), however, the potential of composite likelihood in multiple testing has yet to be explored. There is a great need to develop a procedure to integrate multiple hypothesis

testing procedures with the composite likelihood methodology.

Multiple testing procedures have been developed to control the overall type I error rate when the number of tests is greater than one (Bretz et al., 2010, Hochberg and Tamhane, 1987). The classical Bonferroni method is the simplest procedure to adjust the overall type I error rate, but it is well-known to be very conservative. The Dunn-Sidák procedure (Sidak, 1968) generalizes the Bonferroni procedure by using a slightly less conservative p-value threshold for each comparison. Schéffe (1959) 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. Several stage-wise procedures have also been proposed to improve the power. Simes (1986) modified the Bonferroni procedure based on ordered p-values. Holm (1979) proposed a multi-stage procedure that adjusts the family-wise error rate in each step using the number of remaining null hypotheses. Hommel (1988) 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 stage-wise procedures. As another alternative, Hothorn, Bretz and Westfall (2008a) proposed to use quantiles of the multivariate normal and multivariate t-distribution to perform multiple comparisons in parametric methods. This method takes into account the correlation structure of the test statistics and offers sharper 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 (Hothorn et al., 2008a, Konietschke et al., 2013, 2012). Recently, there has been considerable interest devoted to the problems of large-scale multiple testing applied on the analysis of high dimensional data (see, for example, Benjamini and Hochberg (1995), Meijer (2015)). New decision-analytic based multiple testing procedures (Lisovskaja, 2015) have also been proposed to design multiple testing procedure to minimize a predefined utility function.

In this paper, we propose a new procedure to handle multiple testing scenarios in computationally intensive or intractable likelihood scenarios. We do this by combining multiple testing methods with the dimension-reduction capabilities of inference based on composite likelihood. We explore in detail different multivariate models for correlated clustered data including the multivariate normal, multivariate probit, and quadratic exponential models to illustrate our multiple comparisons approach. Although the proposed composite likelihood methodology can be combined with many multiple testing

procedures, including the recent development in large scale multiple testing or decision-analytic based procedures, in this paper, we focus on combining the Bonferroni, Scheffé, Dunn-Sidák, Holm, and the multivariate normal quantile (MNQ) of Hothorn et al. (2008a) 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 paper is as follows: In Section 2, we develop our composite likelihood based test statistics for multiple inferences and establish their asymptotic properties. In Section 3, we provide details on how to apply the general approach on a variety of multivariate models. In Section 4, we conduct simulation studies to evaluate empirical performance of the proposed method. Finally, we analyze the depression data set to demonstrate the practical utility of the method. This is done in Section 5. We conclude the paper with a brief discussion of the results.

2 Multiple Comparisons Procedures based on Composite Likelihood

Let $\{f(Y;\theta), \theta \in \Theta\}$, where $\theta = (\theta_1, \ldots, \theta_p)^T$, be a parametric statistical model with parameter space $\Theta \subset \mathbb{R}^p$. 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 from different clusters are independent, whereas observations from the same cluster may be dependent. Note that each cluster is thus of size m_i, for an overall sample size of $\sum_{i=1}^n m_i$. In this work, it is assumed that the cluster size, m_i, is uniformly bounded.

Let

$$C = C_{c\times p} = (C^{(1)}, C^{(2)}, \cdots, C^{(c)})^T$$

denote the contrast matrix. A family of c linear combinations of the parameters can then be specified by $C\theta$. Let H_{0i} denote the hypothesis that $C^{(i)}\theta = 0$, for $i = 1, \ldots, c$. We focus here on jointly testing the family of hypotheses $H_{0i}, i = 1, \ldots, c$. In multiple testing, the family-wise type I error (FWER) rate is the probability of falsely rejecting at least one individual null hypothesis. It is said that one has weak control of FWER if the FWER $\leq \alpha$ when all of the null hypotheses are true, whereas one has strong control of FWER if the FWER $\leq \alpha$ under any combinations of null hypotheses and alternative hypotheses.

Most existing multiple comparisons methods estimate the parameters based on the full likelihood. However, for some multivariate distributions, maximizing the full likelihood function can be computationally challenging. Composite likelihood is an alternative method that has attracted much attention in recent years (Besag, 1974, Lindsay, 1988, Varin, 2008, Varin et al., 2011). A composite likelihood is a compounded form of marginal or conditional likelihoods, which is computationally easier to maximize. The general formulation for composite likelihood may be written as follows: Let $A_1, \ldots, A_K$ be a suitably chosen collection of index sets, with $A_k \subseteq \{(i,j), i = 1, \ldots, n, j = 1, \ldots, m_i\}$. For each A_k, a weight w_{A_k} is also chosen/specified. The composite likelihood function is then defined as

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

where $f(y_{A_k}; \theta)$ is the density for the subset vector y_{A_k}. For example, to obtain the so-called univariate composite likelihood $CL(\theta; Y) = \prod_{i=1}^{n} \prod_{j=1}^{m_j} f(y_{ij})$, one chooses $A_k = \{(i,j)\}$ as the single index pairs and weights $w_{A_k} \equiv 1$. (Note that the univariate composite likelihood is equivalent to the full likelihood if the y_{ij} are independent.) The so-called conditional composite likelihood is formulated as $CL(\theta; Y) = \prod_{i,j} f(y_{ij}|y_{i(-j)}) = \prod_{i,j} f(y_i)/f(y_{i(-j)})$, where $y_{i(-j)}$ denotes the sub-vector y_i with its jth element removed. This composite likelihood uses index sets $\{(i,1), \ldots, (i,m_i)\}$ with weight $w_{A_k} = 1$ and index sets $\{(i,1), \ldots, (i,j-1), (i,j+1), \ldots, (i,m_i)\}$ with weight $w_{A_k} = -1$. As this example shows, the index sets A_k need not be disjoint.

The maximum composite likelihood estimate (MCLE) is defined as

$$\widehat{\theta}_n^c \quad = \quad \text{argmax}_{\theta \in \Theta} CL(\theta; Y).$$

Xu and Reid (2011) give precise conditions under which $\widehat{\theta}_n^c$ is consistent for θ. Under appropriate assumptions, $\sqrt{n}(\widehat{\theta}_n^c - \theta)$ is asymptotically normally distributed with mean zero and limiting variance given by the inverse of the the Godambe information matrix (Lindsay, 1988, Varin and Vidoni, 2005), where

$$G^{-1}(\theta) \quad = \quad H^{-1}(\theta) J(\theta) H^{-1}(\theta), \tag{2.1}$$

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 $H(\theta)$ 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 $\widehat{H}_n$ and $\widehat{J}_n$ are consistent (Varin and Vidoni, 2005).

Consider the hypothesis test on a family of linear combinations of the parameters: $\{H_0 : C\theta = 0\}$. Denote by $\Gamma = G^{-1}(\theta)$ the inverse Godambe information matrix, and let $\widehat{\Gamma}_n$ denote a consistent estimator of Γ with $\widehat{\Gamma}_n = \widehat{H}_n^{-1}\widehat{J}_n\widehat{H}_n^{-1}$. We propose the following test statistics for our hypothesis test

$$T_{i,n} \;=\; \frac{C^{(i)^T}\widehat{\theta}_n^c}{\sqrt{\left(C^{(i)^T}\widehat{\Gamma}_n C^{(i)}\right)/n}}, \quad i = 1, \ldots, c. \tag{2.2}$$

Theorem 2.1. *Suppose that the following conditions hold*

1. $\sqrt{n}(\widehat{\theta}_n^c - \theta) \Rightarrow N(0, G^{-1}(\theta))$,
2. H_0 *is true, and*
3. $\widehat{\Gamma}_n \xrightarrow{p} G^{-1}(\theta)$.

Then the limiting distribution of $T_n = (T_{1,n}, \cdots, T_{c,n})^T$ *is multivariate normal* $N_c(0, V)$, *where*

$$V \;=\; \mathrm{diag}(D)^{-1/2} D\, \mathrm{diag}(D)^{-1/2}, \quad D \;=\; CG^{-1}(\theta)C^T. \tag{2.3}$$

Furthermore, since $V_{i,i} = 1$, *the marginal asymptotic distribution of each individual* $T_{i,n}$ *is standard normal.*

Proof of Theorem 2.1. Asymptotic normality of T_n is shown as in Hothorn et al. (2008a). Moreover, as the diagonal elements of the matrix V are equal to one, the individual test statistics $T_{i,n},\ i = 1, \ldots, c$, are standard normal. Therefore, the V matrix is the correlation matrix for $C\widehat{\theta}_n^c$. □

In practice, we estimate V by plugging $\widehat{\Gamma}_n$ as a consistent estimator of $G^{-1}(\theta)$ into (2.3). This results in a consistent estimator of V. It is worthy to point out that the test statistics we propose here are Wald-type of statistics which are not invariant to reparametrization. Under reparametrization, 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 associated with Wald-type statistics.

To apply various multiple testing procedures, we propose to apply the corresponding rejection criterions based on the composite likelihood test

statistics $\{T_{i,n}\}$ derived above. In the numerical analysis, we examine the performance of composite likelihood test statistics with four popular multiple testing methods: Bonferroni, Dunn-Sidák (Sidak, 1968), Holm (Holm, 1979), and the simultaneous multiple comparison test based on multivariate normal quantiles (MNQ) of Hothorn et al. (2008a). The first three methods are applied to the marginal distributions of $T_{i,n}$ based on the asymptotic theory in Theorem 2.1, whereas MNQ uses a cutoff based on the multivariate quantile based on the full variance matrix V in (2.3).

3 Three multivariate models

To showcase our methodology, we consider three different multivariate distributions: The multivariate normal, multivariate probit, and quadratic exponential distributions. A further, fourth, distribution (the skewed multivariate gamma) is considered within the supplementary material. For the first two distributions, the composite likelihood is constructed as sum of univariate likelihoods, whereas for the third distribution, the composite likelihood is constructed as conditional likelihood. All details for the gamma example are given in the supplementary files. Naturally, our methodology is not limited to these distributions and can be applied to other distributions, given that the composite likelihood is available and that the conditions of Theorem 2.1 hold.

In order to include covariates into our modelling scheme, 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).

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

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

where the σ_{jj}'s are nuisance parameters. The Hessian matrix and variability matrix are, respectively, $H(\beta) = n^{-1}\left(\sum_{i=1}^{n} X_i^T W X_i\right)$ and $J(\beta) = n^{-1}\left(\sum_{i=1}^{n} X_i^T W\,\Sigma\,W X_i\right)$, with $W = \mathrm{diag}(\Sigma)^{-1}$. To estimate the regression

coefficients, we employ an iterative algorithm: Given the current estimate for the nuisance parameters σ_{jj}'s, we maximize the composite likelihood to obtain an estimate of $\widehat{\beta}_n^c = (\sum_{i=1}^n X_i^T W X_i)^{-1} \sum_{i=1}^n X_i^T W Y_i$, and given a current estimate for β, we use the sample covariance matrix of residuals to estimate Σ. Based on the estimates $\widehat{\beta}_n^c$ and $\widehat{\Sigma}$, we obtain estimates for $H(\beta)$ and $J(\beta)$ with W being replaced by its estimate $\widehat{W} = \text{diag}(\widehat{\Sigma})$.

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 $y_{ij} = I(y_{ij}^* > 0), \quad j = 1, \cdots, m$ yield the multivariate probit model. We therefore have that $P(y_{ij} = 1|X_i) = \Phi(\vec{x}_{ij}\beta/\sigma)$ 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

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

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

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

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

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

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

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

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

3.3 Quadratic Exponential Model

The quadratic exponential model is a popular tool used to model clustered binary data with intra-cluster interactions (Geys et al., 1997). In this model, the binary observations take values $y_{ij} \in \{-1, 1\}$ and the joint distribution is given by

$$f_Y(y_i) \quad \propto \quad \exp\left\{\sum_{j=1}^{m_i} \mu_{ij}^* y_{ij} + \sum_{j<j'} w_{ijj'}^* y_{ij} y_{ij'}\right\}, \tag{3.4}$$

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 (1999), showed that the joint distribution can be equivalently written in terms of $z_i = \sum_{j=1}^{m_i} 1(y_{ij} = 1)$ (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 (3.4) is famously difficult, but also necessary to compute the full likelihood function. It is therefore desirable to use an alternative approach, one which does not involve such an intensive calculation. Replacing the joint distribution 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.

We now define two conditional probabilities

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

Heuristically, 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 $cl(\mu, w; Y) = \sum_{i=1}^{n} (z_i \log p_{is} + (m_i - z_i) \log p_{if})$.

This special form of the composite likelihood means 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 CMLE 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,

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. (1997).

4 Simulation Results

We evaluate the validity of our proposed approach on three different multivariate models from Section 3 using simulations. We test two different global null hypotheses on the regression coefficients $\beta_1, \cdots, \beta_p$: many-to-one comparisons, $H_0 : \cap_{i=2}^{p}\{\beta_1 = \beta_i\}$; and all pairwise comparisons $H_0 : \cap_{1\leq i,j\leq p, i\neq j}\{\beta_i = \beta_j\}$. The results for many-to-one comparisons are summarized here while the results for all pairwise comparisons are provided in the supplementary material. We choose a collection of different types of multiple testing methods including one-step methods (Bonferroni and Dunn-Sidák), a stagewise (Holm), a projection method (Scheffé), and the MNQ method based on the multivariate distribution of test statistics. For the MNQ method, the critical values can be obtained using the R package `mvtnorm` (Hothorn et al., 2008b).

Part of our goal is to show practitioners what happens if the correlation structure in the clustered data is ignored. To this end, we also include a "misspecified" scenario, where independence is erroneously assumed within the clusters. Due to the specific composite likelihood methods we use (uni-

Table 1: Multiple comparison methods considered:

CASE	multiple comparison method	$\widehat{\Gamma}_n$
(a)	MNQ	$\widehat{H}_n^{-1}\widehat{J}_n\widehat{H}_n^{-1}$
(b)	Bonferroni	
(c)	Dunn-Sidák	
(d)	Holm	
(e)	Scheffé	
(f)	MNQ "naive"	$\widehat{H}_n^{-1}$

variate marginals and univariate conditionals), such a misspecification is equivalent to $H(\theta) = J(\beta)$ in (2.1). This results in an estimate of $\widehat{\Gamma}_n = \widehat{H}_n^{-1}$ in Theorem 2.1. 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 $\widehat{\Gamma}_n = \widehat{H}_n^{-1}$). Throughout, it is referred to as the "naive" approach. Overall, we therefore consider six different approaches, and these are provided in Table 1.

In our simulations, we study the three 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, before we proceed with our more in-depth investigations. 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 2. We observe that $n = 200, 500$ and 700 are required for the multivariate normal, multivariate probit and quadratic exponential 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 2: 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 evaluate the power of each of the different methods, we consider two different alternative scenarios: one alternative configuration a_1 with only one non-zero parameter with a large effect size, and a second alternative configuration 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 the global null hypothesis, but also in the ability of the test to reject the 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). Note also that the true effects are purposefully chosen to be small under a_2, and therefore the typical empirical results are considerably smaller than 5, as expected. This is done so that the observed global power is not uniformly high, which allows us to detect more subtle differences among the various methods.

4.1 Multivariate Normal 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 (three exchangeable Σ scenarios) and Table 4 (general Σ) summarize the results of our simulations. Overall, it is shown that the method which utilizes MNQ and correctly accounts for the intra-cluster correlations, has the best performance. 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. Overall, MNQ exhibits the best performance among all of the multiple comparison procedures. 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.

Table 3: Simulations results for the multivariate normal model with exchangeable Σ

	ρ	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé	efficiency
FWER				0.0545	0.0553	0.0419	0.0427	0.0419	0.0007	0.9983
a_1			10	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		4		0.9080	0.9079	0.8417	0.8503	0.8848	0.2242	
FWER				0.0511	0.0502	0.0352	0.0363	0.0352	0.0000	0.9980
a_1			20	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			0.7698	0.7674	0.7081	0.7007	0.7518	0.0162	
FWER				0.0479	0.0471	0.0375	0.0378	0.0375	0.0001	0.9989
a_1			10	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		10		1.4822	1.4816	1.4219	1.4284	1.4896	0.8933	
FWER				0.0487	0.0485	0.0363	0.0373	0.0363	0.0000	0.9986
a_1			20	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				0.0494	0.0670	0.0389	0.0397	0.0389	0.0001	0.9453
a_1			10	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		4		0.8556	0.9268	0.7939	0.8032	0.8317	0.1845	
FWER				0.0533	0.0734	0.0390	0.0397	0.0390	0.0000	0.9430
a_1			20	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	0.2			0.7106	0.7777	0.6476	0.6438	0.6937	0.0102	
FWER				0.0467	0.1019	0.0357	0.0365	0.0357	0.0003	0.8685
a_1			10	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		10		1.3506	1.5795	1.2871	1.2968	1.3407	0.7374	
FWER				0.0468	0.1057	0.0320	0.0331	0.0320	0.0000	0.8636
a_1			20	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				0.0513	0.0977	0.0390	0.0398	0.0390	0.0007	0.7491
a_1			10	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		4		0.7570	0.9391	0.7208	0.7074	0.7533	0.1502	
FWER				0.0510	0.1029	0.0377	0.0385	0.0377	0.0000	0.7343
a_1			20	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	0.5			0.6369	0.8035	0.5677	0.5750	0.6109	0.0076	
FWER				0.0520	0.2079	0.0410	0.0417	0.0410	0.0000	0.6070
a_1			10	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		10		1.1914	1.6903	1.1367	1.1372	1.1877	0.5096	
FWER				0.0459	0.2271	0.0328	0.0337	0.0328	0.0000	0.5898
a_1			20	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	

Table 4: Simulations results for the multivariate normal model 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			0.5499	0.6707	0.5094	0.5112	0.5357	0.0615
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			0.4057	0.5090	0.3436	0.3508	0.3743	0.0011
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			0.5442	0.5142	0.4870	0.4911	0.5158	0.0537
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			0.3711	0.3490	0.3186	0.3204	0.3462	0.0010

4.2 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 5. 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.

4.3 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 6. Overall, MNQ has clearly the best performance.

Table 5: Simulation results for the probit model

	ρ	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé
FWER				0.0530	0.0506	0.0413	0.0424	0.0413	0.0001
a_1			10	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		4		1.0828	1.0779	1.0193	1.0305	1.0682	0.3590
FWER				0.0528	0.0503	0.0389	0.0395	0.0389	0.0000
a_1			20	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			0.9436	0.9389	0.8847	0.8825	0.9308	0.0410
FWER				0.0526	0.0515	0.0423	0.0428	0.0423	0.0005
a_1			10	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		10		1.6649	1.6594	1.5839	1.5939	1.6658	1.0024
FWER				0.0527	0.0508	0.0364	0.0375	0.0364	0.0000
a_1			20	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				0.0508	0.0793	0.0393	0.0404	0.0393	0.0003
a_1			10	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		4		0.9852	1.1028	0.9321	0.9334	0.9768	0.2708
FWER				0.0585	0.0915	0.0406	0.0415	0.0406	0.0000
a_1			20	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	0.5			0.8637	0.9712	0.7855	0.7963	0.8330	0.0247
FWER				0.0513	0.1437	0.0402	0.0412	0.0402	0.0005
a_1			10	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		10		1.4075	1.7926	1.3520	1.3552	1.4154	0.8147
FWER				0.0543	0.1622	0.0382	0.0389	0.0382	0.0000
a_1			20	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

Table 6: Simulation results for the quadratic exponential model

	w	m	p	MNQ	naive	Bonf	S-D	Holm	Scheffé
FWER				0.0514	0.0562	0.0400	0.0403	0.0400	0.0001
a_1			10	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		4		0.9779	1.0283	0.8826	0.8888	0.9373	0.0986
FWER				0.0561	0.0767	0.0404	0.0412	0.0404	0.0000
a_1			20	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			0.7731	0.8347	0.6514	0.6679	0.7029	0.0027
FWER				0.0491	0.0549	0.0381	0.0384	0.0381	0.0001
a_1			10	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		10		0.9780	1.0284	0.8826	0.8890	0.9373	0.0985
FWER				0.0561	0.0767	0.0404	0.0412	0.0404	0.0000
a_1			20	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				0.0521	0.0000	0.0417	0.0424	0.0417	0.0002
a_1			10	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		4		1.5136	0.0452	1.4102	1.4089	1.4915	0.2753
FWER				0.0509	0.0000	0.0377	0.0383	0.0377	0.0000
a_1			20	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	0.5			1.2902	0.0150	1.1532	1.1612	1.2141	0.0134
FWER				0.0521	0.0000	0.0417	0.0424	0.0417	0.0002
a_1			10	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		10		1.5326	0.0416	1.4102	1.4261	1.4915	0.2746
FWER				0.0509	0.0000	0.0378	0.0384	0.0378	0.0000
a			20	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

5 Analysis of Depression Data

Table 7: Composite likelihood estimates of the health factors' regression coefficients

	estimate	SE	p-value
sleeplessness	1.3330	0.0290	$< 2e-16$
smoking	0.2826	0.0439	$< 2e-16$
high blood pressure	0.0764	0.0219	$2.07e-11$
diabetes	0.0710	0.0296	$8.96e-07$
difficulty in walking	0.0695	0.0054	$< 2e-16$
age	0.0007	0.00003	$< 2e-16$
activity	-0.0156	0.0064	$2.35e-05$
w	0.2877	0.0094	$< 2e-16$

We apply our proposed method to the health and retirement study (HRS) dataset. Information about health, financial situation, family structure, and health factors were collected by the RAND center at the University of Michigan. We perform multiple comparisons on the effects of seven health factors on depression status of seniors. Depression status is recorded as a binary response variable, whereas the seven health factors include age (in months), smoking, restless sleep, diabetes, high blood pressure, frequent vigorous physical activity, and difficulty in walking. For each individual we include only the years for which all of the factors were recorded. In total, there are 33 636 people included in the analysis and the number of repeated measurements vary across individuals. As the response variable is binary and the cluster sizes vary, the quadratic exponential model is a natural choice to model this data set.

The full likelihood approach is very computationally challenging for this model, and hence we use the proposed composite likelihood based method to perform inference. The w parameter in the quadratic exponential model allows us to account for the interaction effect among the repeated measurements for the same individuals. The MCLE estimates and the associated standard errors are reported in Table 7.

To compare the effect sizes of all the seven health factors, we perform all

pairwise comparisons on the seven parameters with $H_{0,i,j} = \{\beta_i = \beta_j\}$ for a total of 21 null hypotheses. Of the six methods described in Table 1, we choose the MNQ approach based on its superior performance in Section 4. To show how different the results will be if the within-patient correlations are ignored, we also compare the result of the MNQ with the naive MNQ method. Both the MNQ and the naive MNQ reject the global null hypothesis that all pairs of health factors have equal effect on the depression status. The results for the individual hypotheses are given in Table 10.

The MNQ method rejects 15 hypotheses, whereas the naive MNQ method rejects 18 out of the total 21 hypotheses. Based on the estimates of the effect sizes (Table 7), we note that restless sleep and smoking are the two health factors with the largest effect sizes. Both MNQ and naive MNQ reject the pairwise comparisons between restless sleep with all other health factors and smoking with all other factors. This shows that restless sleep and smoking are the two leading health factors for the occurrence of depression. High blood pressure, diabetes, and difficulty in walking have the third, fourth and fifth largest estimated effect sizes. When we examine the three pairwise comparisons among these three factors, both MNQ and naive MNQ accept the three null hypotheses, indicating that these three health factors have similar effects and importance to the disease. Furthermore, when we compare high blood pressure with age and activity, both methods reject the two comparisons, indicating that high blood pressure is more important than age and activity with regard to the disease development.

MNQ and naive MNQ are in agreement in all the aforementioned comparisons. However, when we compare the effect sizes between age and diabetes, diabetes and activity, age and activity, the MNQ method accepts these three null hypotheses while the naive method rejects all three. The difference between the two methods is due to the correlation among the repeated measurements, which is estimated as $\widehat{w} = 0.285$. By ignoring this correlation, as in the naive method, the standard errors are underestimated leading to more rejections.

6 Discussion

In many correlated multivariate models, it is often difficult to perform multiple comparisons based on the full likelihood. In this paper, we propose to use the composite likelihood method to construct multiple comparison procedures to overcome this computational difficulty. Theory is developed based on the asymptotic properties of the composite likelihood test statistic

Table 8: Results of MNQ and naive MNQ in testing individual null hypotheses in the depression study data set. A: fail to reject, R: reject H_0

H_0	MNQ	naive	H_0	MNQ	naive
$\beta_{sleep} = \beta_{smoke}$	R	R	$\beta_{hbp} = \beta_{diabet}$	A	A
$\beta_{sleep} = \beta_{hbp}$	R	R	$\beta_{hbp} = \beta_{diff\ walk}$	A	A
$\beta_{sleep} = \beta_{diabet}$	R	R	$\beta_{hbp} = \beta_{age}$	R	R
$\beta_{sleep} = \beta_{diff\ walk}$	R	R	$\beta_{hbp} = \beta_{activity}$	R	R
$\beta_{sleep} = \beta_{age}$	R	R	$\beta_{diabet} = \beta_{diff\ walk}$	A	A
$\beta_{sleep} = \beta_{activity}$	R	R	$\beta_{diabet} = \beta_{age}$	A	R
$\beta_{smoke} = \beta_{hbp}$	R	R	$\beta_{diabet} = \beta_{activity}$	A	R
$\beta_{smoke} = \beta_{diabet}$	R	R	$\beta_{diff\ walk} = \beta_{age}$	R	R
$\beta_{smoke} = \beta_{diff\ walk}$	R	R	$\beta_{diff\ walk} = \beta_{activity}$	R	R
$\beta_{smoke} = \beta_{age}$	R	R	$\beta_{age} = \beta_{activity}$	A	R
$\beta_{smoke} = \beta_{activity}$	R	R			

and illustrated for three different models: multivariate normal, multivariate probit and quadratic exponential. The simultaneous quantile of multivariate normal is used as a threshold for test statistics compared to some well-known traditional thresholds. This 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.

7 Acknowledgment

All authors are grateful to the anonymous referees, Editor, and Associate Editor for their careful reading of the manuscript and valuable comments, which greatly improved the manuscript. This work is supported by NSERC grants held by Gao and Jankowski.

Appendix

7.1 Simulation result on all pairwise comparisons

In Section 4 of the main manuscript, we provide simulation results under various settings for many-to-one comparisons. Here, we provide additional simulations for all pairwise comparisons. We simulate multivariate normal, multiviate probit and quadratic exponential models as described in Section 4 of the paper. The global null hypothesis, sample size and the two alternative configurations are the same as those used in many-to-one comparisons. We perform all pairwise comparisons with $m = 4$ and $p = 10$. For the multivariate normal, multivariate probit, and quadratic exponential models, we consider $\rho = 0$, or 0.5. For the quadratic exponential, we consider $w = 0, 0.5$. The results are summarized in Table 1. We again observe that the MNQ approach has the best performance. MNQ maintains good control of the FWER except for the case of quadratic exponential model with $\rho = 0.5$, where it is slightly above 0.05. The naive MNQ either has either very large FWER or very small FWER, indicating its poor control of the error rate. Among all the methods which maintain good control of FWER, the MNQ method achieves the highest power. In addition, we consider the Tukey approach, as it is a commonly used testing procedure in all pairwise comparisons.

Table 9: Simulations results for 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

7.2 A skewed distribution example

Here, we consider a multivariate gamma distribution which has marginal univariate gamma distribution and a covariance structure. 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 (Ronning, 1997). (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.

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

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

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 $\widehat{\mu}_i = \exp\{X_i\widehat{\beta}_n^c\}$.

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 $\beta_{a_2}^T = (0.75, 0.80, 0.68, 0.70, 0.79, 0.69, 0.75, \ldots, 0.75)$. 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 2.

Table 10: FWER and power for 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

7.3 Some technical details

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.

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 ∂^α 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

$$\left|(\partial^\alpha \partial^{\theta_i} \log f)(y;\theta)\right| \leq M(y),$$

for all $||\theta - \theta_0||_2 < c$, all $|\alpha| = 2$, and any $i = 1, \ldots, p$. Here, $||\cdot||_2$ 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 7.1. *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.

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 CMLE 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

$$\begin{aligned} \xi^T \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\nabla cl_m)(\theta_0, y_i) \right\} + \xi^T \left\{ \frac{1}{n} \sum_{i=1}^{n} (Hcl_m)(\theta^*, y_i) \right\} \xi \\ \equiv \quad \xi^T b_n M + \xi^T B_n \xi M^2, \end{aligned} \tag{7.5}$$

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 $\epsilon > 0$, there exists an M_ϵ, such that

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

Then, by (7.5), 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

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

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

Let $g(\theta; y) = cl_m^{(1)}(\theta; y) = \nabla cl_m(\theta; y)$ (this is the vector of first deriva-

tives), then using a multivariate Taylor expansion, we have that

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

again using the multi-index notation. We take $\widehat{\theta}_n^c$ to be the local maximizer found above. This time, for fixed y, g 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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (D cl_m)(\theta; y_i)^T (\widehat{\theta}_n^c - \theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n R_{n,i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(\theta_0; y_i). \quad (7.6)$$

By condition (A4), we have that

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

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 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 (7.6) 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. □

References

Barua, G. M. N. B. M., A. (2011). Prevalence of depressive disorders in the elderly **31** 620–624.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. 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.

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

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

Geys, H., Molenberghs, G. and Ryan, L. M. (1997). Pseudo-likelihood inference for clustered binary data. *Comm. Statist. Theory Methods* **26** 2743–2767.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test 383–386.

Hothorn, T., Bretz, F. and Westfall, P. (2008a). Simultaneous inference in general parametric models. *Biom. J.* **50** 346–363.

Hothorn, T., Bretz, F., Westfall, P. and Heiberger, R. M. (2008b). multcomp: Simultaneous inference in general parametric models .

Konietschke, F., Bosiger, S., Brunner, E. and Hothorn, L. A. (2013). Are multiple contrast tests superior to the anova? *Int. J. Biostat.* **9** 11.

Konietschke, F., Hothorn, L. A. and Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electron. J. Stat.* **6** 738–759.

Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical inference from stochastic processes (Ithaca, NY, 1987)*, vol. 80 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 221–239.

Lisovskaja, B. C. F., V. (2015). A decision theoretic approach to optimization of multiple testing procedures **57** 64–75.

Meijer, G. J. J., R. J. (2015). A multiple testing method for hypotheses structured in a directed acyclic graph **57** 123–143.

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

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Statist. Data Anal.* **44** 649–667.

Ronning, G. (1997). A simple scheme for generating multivariate gamma distributions with non-negative covariance matrix **19** 179–183.

Schéffe (1959). *The analysis of variance*. Wiley, New York.

Sidak, Z. (1968). On multivariate normal probabilities of rectangles: Their dependence on correlations. *Ann. Math. Statist.* **39** 1425–1434.

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

Varin, C. (2008). On composite marginal likelihoods. *AStA Adv. Stat. Anal.* **92** 1–28.

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

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

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *J. Statist. Plann. Inference* **141** 3047–3054.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33** 335–356.