Multiple Comparisons using Composite Likelihood in Clustered Data

Mahdis Azadbakhsh, Xin Gao and Hanna Jankowski
York University
November 6, 2014

Abstract

We study the problem of multiple hypothesis testing for multidi-
mensional data when inter-correlations are present. The problem of
multiple comparisons is common in many applications. When the data
is multivariate and correlated, existing multiple comparisons proce-
dures based on maximum likelihood estimation could be prohibitively
computationally intensive. We propose to construct multiple compar-
isions procedures based on composite likelihood statistics. We focus
on data arising in three ubiquitous cases: multivariate Gaussian, pro-
bib, and quadratic exponential models. To help practitioners assess
the quality of our proposed methods, we assess their empirical per-
formance 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 erratic performance. Using data aris-
ing from a diabetic nephropathy study, we show how our composite
likelihood approach makes an otherwise intractable analysis possible.

1 Introduction

Correlated data often arise when sub-populations of the observed data are
related in some manner. These observations can be described as “clustered”
because they are recorded from the same experimental unit multiple times,
or come from the same group sharing common relationships. Thus, clus-
tered data examples arise in numerous situations, including sibling, pedi-
gree, longitudinal, or spatial data. Ignoring such correlations among the
subjects can lead to invalid inference for individual comparisons as well as
global decisions. Therefore, when performing multiple comparisons in clustered data, one should take into account the correlation structure within the clusters. However, full likelihood analyses on these multivariate distributions often encounter computational challenges. For example, 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 computational intensity increases with the cluster size. Here, we solve 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 \cite{Cox2004,Lindsay1988}. 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 \cite{Cox2004,Lindsay1988,Varin2008,Varin2011}. This makes it an appealing alternative in inferential procedures. \cite{Xu2011} 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.

Composite likelihood methodology has been applied to numerous statistical problems: \cite{Zhao2005} proposed composite likelihood methods for multivariate data analysis. \cite{Renard2004} used it in the generalized linear mixed model; \cite{Fearnhead2002} proposed to maximize the compounded marginal probabilities in genetics; \cite{Geys1997} 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 \cite{Heagerty1998,Hiort1994,Varin2005}, Markov random fields \cite{Besag1974}, and multivariate survival analysis \cite{Li2006,Parner2001}. However, the potential of composite likelihood in multiple testing has yet to be explored.

Multiplicity of hypothesis tests arises when the number of simultaneous comparisons is greater than one, leading to a family-wise type I error rate
larger than $\alpha$. The greater the number of comparisons, the more serious this effect becomes. There are various procedures available to adjust the overall type I error rate when performing multiple comparisons, cf. \cite{Bretz2010}, \cite{Hochberg1987}. 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-Sidak procedure \cite{Sidak1968} generalizes the Bonferroni procedure by using a slightly less conservative $p$-value threshold for each comparison. \cite{Scheffe1959}, \cite{Tukey1953} and \cite{Fisher1935} established methods for testing all possible linear comparisons among a set of normally distributed variables.

The above methods tend to be over-conservative for a finite family of multiple comparisons. To correct for this, several stage-wise procedures have been proposed. \cite{Simes1986} modified the Bonferroni procedure based on ordered $p$-values. \cite{Holm1979} proposed a multi-stage procedure that adjusts the family-wise error rate in each step using the number of remaining null hypotheses. \cite{Hommel1988} 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, \cite{Hothorn2008} 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. Later, \cite{Konietschke2012} proposed a rank based multiple comparison procedure for nonparametric models. Their method can accommodate non-normally distributed continuous data or discrete data. \cite{Konietschke2012} and \cite{Hothorn2008} discussed the construction of simultaneous confidence intervals based on multiple comparisons test statistics. \cite{Konietschke2012} compared their multiple contrast tests with the ANOVA F-test to test arbitrary linear hypotheses. They conclude that their multiple contrast tests can provide both local and global test decisions whereas the ANOVA F-test can only provide decision on the global null hypothesis. Therefore, when individual inference is needed, the multiple contrast test based on multivariate quantiles leads to sharp control of the family-wise type I error rate and provide both individual decisions and simultaneous confidence interval.

In this paper, we propose to use composite likelihood to estimate the parameters in multivariate models and construct multiple comparisons procedure based on the resulting composite likelihood statistics. We explore in
detail the multivariate normal, multivariate probit and quadratic exponential models to illustrate our multiple comparisons approach. Specifically, we compare the Bonferroni, Dunn-Sidak (Sidak, 1968), Holm (Holm, 1979), and the multivariate normal quantile (Hothorn et al., 2008a) methods in both a marginal and conditional composite likelihood. The multivariate normal quantile threshold suggested by Hothorn et al. (2008a) appears to have 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 describe our approach: we begin by reviewing multiple comparison approaches (Section 2.1) and composite likelihood (Section 2.2); In Section 2.3 we define our hypothesis testing approach based on composite likelihood, and then show how this is combined with multiple testing in Section 2.4. We then provide details on how to apply the general approach to the Gaussian, probit, and quadratic exponential settings in Section 3. For these three models, we provide a simulation study to provide some guidance on practical performance and best practice for small and medium sample sizes, see Section 4. Finally, the methodology is applied to a real data set. This is done in Section 5. The paper concludes with a brief discussion of the results.

2 Multiple Comparisons Procedures based on Composite Likelihood

2.1 Multiple Comparisons Procedures

In a parametric statistical model, suppose that a random vector \( y = y_{m \times 1} \) has a multivariate distribution density \( f(y; \theta) \). Let

\[
C = C_{exp} = (C^{(1)}, C^{(2)}, \ldots, C^{(c)})^T
\]

denote the contrast matrix and 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 false rejection of at least one individual null hypothesis when all of the null hypotheses are true,

\[
P \left( \text{reject at least one } H_{0i} \bigg| \bigcap_k H_{0k} \right).
\]

The above probability should be as close to a pre-specified significance level \( \alpha \), as possible, without going over. Numerous different procedures
have been suggested to thus control the FWER by adjusting the individual
test levels. These procedures fall into two main categories of single-stage
and multi-stage methods. The most commonly used Bonferroni procedure
compares a common threshold $\alpha/c$ with all observed individual p-values.
This approach is known to be very conservative, especially when the data
are highly correlated. Sidak (1968) proposed a slightly less conservative
threshold of $1 - (1 - \alpha)^{1/c}$ for individual p-values. Holm (1979) improved
the Bonferroni method by proposing a multistage procedure that adjusts
the FWER in each step using the number of remaining null hypotheses.
Simes (1986) proposed a p-value correction based on ordered p-values. In
his method, $H_0 = \cap_k H_{0k}$ is rejected if at least one $P(j) > \frac{j \alpha}{n}$, where $P(j)$ is the
$j^{th}$ smallest p-value. A stagewise rejective multiple test based on the prin-
ciple of closed test procedure was suggested by Hommel (1988). Although
these stepwise methods are simple and more powerful than the Bonferroni
correction, they do not explicitly take into account the dependency structure
between the test statistics. Hothorn et al. (2008a) proposed to use quantiles
of multivariate normal and multivariate t-distributions as the thresholds for
test statistics with such behaviours. Efficient software has been developed
by Hothorn et al. (2008b) to compute these quantiles. This method makes
full use of the correlation structure of the test statistics and the FWER
control is exact or asymptotically exact if the test statistics is exactly or
asymptotically distributed as a multivariate normal or t-distribution.

2.2 Composite Likelihood

In this work, observations are assumed to come from clustered data with
intra-cluster correlation. Examples of clustered data include longitudinal
and hierarchical data. Repeated measurements in clinical research, parent-
sibling data in genetic studies and spatial data are examples of clustered
data with intra-correlation. They share the common characteristic that the
subjects can be naturally partitioned into groups, such as measurements
from one patient, members from a single family, or observations from a
spatially defined neighbourhood, and the responses tend to be correlated
within the groups. In order to perform valid multiple comparisons, the
correlation structure has to be taken into account.

Most existing multiple comparisons methods require the joint density
to be known to estimate the parameters and covariance components based
on the full likelihood. However, as maximizing the full likelihood function
can be computationally difficult or even prohibitive, this approach may not
be feasible. Composite likelihood is an alternative method that has at-
tracted 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 often easier to specify and computationally much easier to maximize. Let \( \{f(Y; \theta), \theta \in \Theta\} \) be a parametric statistical model with parameter space \( \Theta \subset \mathbb{R}^p \). Let \( Y = (y_1^T, \ldots, y_n^T) \) denote the response variables, where \( y_i = (y_{i1}, \ldots, y_{im})^T \) is the vector of observations from cluster \( i, i = 1, \ldots, n \) from a study population. Observations from different clusters are independent, whereas observations from the same cluster are 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. Cox and Reid (2004) proved that the maximum composite likelihood estimator is not usually consistent due to the large intra-clusters correlation, when the number of clusters is fixed and cluster size increases. Zhao and Joe (2005) showed that in general, efficiency of univariate maximum composite likelihood estimator tends to decrease when \( m \) increases and cluster sizes are different.

To form the composite likelihood, we begin with some notation, in this case tuned specifically to the cluster data we focus on here. Let \( A \) denote a collection of indices: \( A \subset \Omega = \{(i, j), j = 1, \ldots, m_i, i = 1, \ldots, n\} \). For a fixed \( A \), let \( y_A \) denote the vector \( \{y_a, a \in A\} \). For example, if \( A = \{(1, j), j = 1, \ldots, m_1\} \), then \( y_A = \{y_{11}, \ldots, y_{1m_1}\} \), which is the vector of the response variables from the first cluster. For a collection of index subsets \( \mathcal{A} = \{A : A \subset \Omega\} \), the composite likelihood function is defined as

\[
CL(\theta; Y) = \prod_{A \in \mathcal{A}} f(y_A; \theta)^{w_A},
\]

where \( \{w_A, A \in \mathcal{A}\} \) is an index of suitable weights. The maximum composite likelihood estimate is then defined as

\[
\hat{\theta}_c = \arg\max_{\theta \in \Theta} CL(\theta; Y).
\]

Due to our cluster data structure, we can further define \( \mathcal{A}_i = \{A : A \subset \Omega_i\} \) with \( \Omega_i = \{(i, j), j = 1, \ldots, m_i\} \), so that \( \mathcal{A} = \{\mathcal{A}_i, i = 1, \ldots, n\} \). We then have

\[
CL(\theta; Y) = \prod_{A \in \mathcal{A}} f(y_A; \theta)^{w_A} = \prod_{i=1}^{n} \prod_{A \in \mathcal{A}_i} f(y_A; \theta)^{w_A}.
\]

For example, if \( \mathcal{A}_i = \{\Omega_i\} \) and \( w_A = 1 \) for all \( A \), then there is only one element in \( \mathcal{A}_i \) and the composite likelihood is equivalent to the full likelihood,

\[
CL(\theta; Y) = \prod_{i=1}^{n} f(y_{i1}, y_{i2}, \ldots, y_{im_i}; \theta)^{w_A}.
\]
On the other hand, if \( A_i = \{(i, 1), \ldots, (i, m_i)\} \) again with \( w_A = 1 \), then the formulation leads to composite likelihood of univariate margins,

\[
CL(\theta; Y) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} f(y_{ij}; \theta).
\]

The asymptotic properties of the maximum composite likelihood estimator (MCLE) are well-understood in various contexts, cf. Lindsay (1988). Xu and Reid (2011) give precise conditions under which \( \hat{\theta}_c \) is consistent for \( \theta \), even when the marginal densities are misspecified. It is also well-known that, under appropriate assumptions, \( \sqrt{n}(\hat{\theta}_c - \theta) \) is asymptotically normally distributed with mean zero and limiting variance given by the inverse of the the Godambe information matrix [Varin and Vidoni (2005), where

\[
G^{-1}(\theta) = H^{-1}(\theta)J(\theta)H^{-1}(\theta),
\]

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 \( CL(\theta; Y) \) with respect to \( \theta \). For convenience, we provide a sketch of the proof of this fact for \( m_i = m \) and marginal likelihood in the Appendix. The proof for varying \( m_i \) (but fixed and bounded), would be similar.

### 2.3 Defining the hypothesis tests

Let \( \theta = (\theta_1, \ldots, \theta_p) \) denote the vector of multiple parameters. Consider the hypothesis test on a family of linear combinations of the parameters:

\( \{H_0 : C\theta = 0\} \), recalling that \( C = (C^{(1)}, C^{(2)}, \ldots, C^{(c)})^T \) is the contrast matrix corresponding to the null hypothesis \( H_0 = \cap_{i=1,\ldots,c} H_{0i} \). Denote by \( \Gamma = G^{-1}(\theta) \) the inverse Godambe information matrix, and let \( \hat{\Gamma}_n \) denote any consistent estimator of \( \Gamma \).

We propose the following test statistics for our hypothesis test

\[
T_{l,n} = \frac{C^{(l)^T} \hat{\theta}_c}{\sqrt{(C^{(l)^T} \hat{\Gamma}_n C^{(l)})/n}}, \quad l = 1, \ldots, c.
\]

### Theorem 2.1

Suppose that the following conditions hold

1. \( \sqrt{n}(\hat{\theta}_c - \theta) \Rightarrow N(0, G^{-1}(\theta)) \),

2. \( H_0 \) is true, and

\[\sqrt{n}(\hat{\theta}_c - \theta) \Rightarrow N(0, G^{-1}(\theta))\)
Then the limiting distribution of $T_n = (T_{1,n}, \ldots, T_{c,n})^T$ is multivariate normal $N_c(0, V)$, where
\begin{equation}
V = \text{diag}(D)^{-1/2}D \text{diag}(D)^{-1/2}, \quad D = CG^{-1}(\Theta)C^T. \tag{2.3}
\end{equation}

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

In practice, we estimate $V$ by plugging in a consistent estimator of $G^{-1}(\theta)$, here denoted as $\hat{\Gamma}_n$, into (2.3). Clearly, this results in a consistent estimator of $V$.

**Proof of Theorem 2.1.** Asymptotic normality of $T_n$ is shown as in Hothorn et al. (2008a, Section 2). Moreover, as the diagonal elements of the matrix $V$ are equal to one, the individual test statistics $T_{l,n}$, $l = 1, \ldots, c$, are standard normal. Indeed, the $V$ matrix is the correlation matrix for $C\hat{\theta}_c$. Many well-known multiple comparison procedures are based on the closure principle which leads to a strong control over the family-wise type I error rate (FWER). Given the set of null hypotheses $H_0 = \{H_0, i = 1, \ldots, c\}$, let $w$ be any subset of $\{1, \ldots, c\}$ and $H_{0w} = \bigcap_{i \in w} H_0$. A multiple testing procedure strongly controls the family-wise type I error rate at level $\alpha$ if under any subset of the set of null hypotheses $H_{0w}$, $P($rejecting at least one $H_0, i \in w \mid \bigcap_{w} H_{0w}) \leq \alpha$. Using a closed test procedure, every single hypothesis $H_{0i}$ is only rejected at level $\alpha$ if its test, as well as the tests of all supersets that contain $H_{0i}$, are significant (Marcus et al., 1976). When a closed test procedure is applied on a joint testing family, the resulting procedure strongly controls the FWER.

**Theorem 2.2.** The family of hypotheses $\{H_{0i}, i = 1, \ldots, c\}$ and the corresponding test statistics $\{T_{i,n}, i = 1, \ldots, c\}$ constitute a joint testing family, asymptotically, in the sense of Gabriel (1969, p. 225).

**Proof.** For every $i$, the test statistic $T_{i,n}$ under null hypothesis $H_{0i}$ is asymptotically normally distributed with mean zero and variance one. Moreover, the limiting distribution of $T_{i,n}$ is completely specified, even if some/any of the null hypotheses do not hold. Therefore, for any intersection of hypotheses, $H_{0w} = \bigcap_{i \in w} H_{0i}$, the joint distribution of $T_{w,n}$ is completely specified under $H_{0w}$. It follows that the family of hypotheses $\{H_{0i}, i = 1, \ldots, c\}$ and the corresponding test statistics $\{T_{i,n}, i = 1, \ldots, c\}$ constitute a joint testing family. 

8
In light of the above result, any closed testing procedure applied to the test statistics $T_{i,n}$ will strongly control the FWER.

### 2.4 Multiple comparisons using composite likelihood in clustered data

To apply the composite likelihood approach in the context of multiple testing, we propose to use one of four popular methods: Bonferroni, Dunn-Sidak (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. On the other hand, MNQ uses a cutoff based on the multivariate quantile based on the full variance matrix $V$ in (2.3). These four multiple testing approaches are applied directly to the composite likelihood-based hypothesis tests defined in (2.2).

However, in order to fully define our procedure, we also need to choose the estimator $\hat{\Gamma}_n$. Here, we use

$$
\hat{\Gamma}_n = \hat{H}_n^{-1} \hat{J}_n \hat{H}_n^{-1},
$$

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

Let $L_A(\theta; y)$ denote the likelihood of $\theta$ when the data is restricted to the vector $y_A$. To estimate $H(\theta)$ and $J(\theta)$, it is proposed in Cox and Reid (2004) that

$$
\hat{H}_n = -\frac{1}{n} \sum_{i=1}^{n} \sum_{A \in A_i} \left. \frac{\partial^2 l_A(\theta, y)}{\partial \theta \partial \theta^T} \right|_{\theta = \hat{\theta}_n},
$$

where $l_A(\theta, y) = \log L_A(\theta; y)$ and $\hat{H}$ 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:

$$
\hat{J}_n = \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{A \in A_i} \left. \frac{\partial l_A(\theta, Y)}{\partial \theta} \right|_{\theta = \hat{\theta}_n} \right) \times \left( \sum_{A \in A_i} \left. \frac{\partial l_A(\theta, Y)}{\partial \theta} \right|_{\theta = \hat{\theta}_n} \right)^T.
$$

Both estimators $\hat{H}_n$ and $\hat{J}_n$ are consistent (Varin and Vidoni, 2005, page 523). For more details on the estimation of $H$ and $J$, we refer to Cox and Reid (2004) and Varin (2008).
Resampling methods such as bootstrap and jackknife are alternatives that can be employed to estimate the covariance matrix, particularly for smaller sample sizes. Zhao and Joe (2005) proposed a jackknife approach, however, computation in this method can be intensive and time consuming for high dimensional data. Varin et al. (2011, Section 5.1) discuss the estimation of the component of Godambe information matrix in more detail.

3 Three multivariate models

To model the dependency structure in clustered data, we consider three different multivariate distributions including the multivariate normal, multivariate probit, and quadratic exponential distributions. We note that the methodology proposed is not limited to these three distributions and can be applied to other distributions given that the composite likelihood is available, assuming 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 Normal distribution

Let \( \{(y_i, X_i), i = 1, \ldots, n\} \), denote the response and covariates arising from a multivariate normal model, with

\[
y_i = X_i \beta + \epsilon_i, i = 1, \ldots, n,
\]

with \( m_i = m, i = 1, \ldots, n \). We assume that \( \epsilon_i \sim N_m(0, \Sigma) \) where \( \Sigma = (\sigma_{ij}), i, j = 1, \ldots, m \), is an arbitrary covariance matrix, and we write \( \sigma_j^2 \) instead of \( \sigma_{jj} \). The univariate composite likelihood is formulated as

\[
CL(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left( -\frac{1}{2} \log(2\pi \sigma_j^2) - \frac{1}{2\sigma_j^2} (y_{ij} - \vec{x}_{ij}\beta)^2 \right).
\]

Note that this version of the composite likelihood can only provide information on the diagonal elements of \( \Sigma \), and we denote these as \( \text{diag}(\Sigma) = \).
\{\sigma_1^2, \ldots, \sigma_m^2\}. Let \( W = \text{diag}(\Sigma)^{-1} \). Then

\[ H(\beta) = n^{-1} \left( \sum_{i=1}^{n} X_i^T W X_i \right) \]
\[ J(\beta) = n^{-1} \left( \sum_{i=1}^{n} X_i^T W \Sigma W X_i \right) \]

We have removed the parameters \( \text{diag}(\Sigma) \) from our likelihood formulation, since we need to estimate the full covariance matrix empirically, without resorting to likelihood methods.

To estimate the regression coefficients and the covariance matrix, we employ a partial maximization algorithm. Ideally, given the current estimate for \( \Sigma \), we would maximize the composite likelihood to find an estimate for \( \beta \), and given a current estimate for \( \beta \), we would maximize the composite likelihood to find an estimate for \( \Sigma \). However, since the composite likelihood does not provide information on the full matrix \( \Sigma \), we replace that step with an empirical estimator instead. Thus, given the values of \( \text{diag}(\Sigma) \), we find that the maximizer of (3.4) in \( \beta \) is

\[ \hat{\beta}^c \bigg|_{\text{diag}(\Sigma)} = \left( \sum_{i=1}^{n} X_i^T W X_i \right)^{-1} \sum_{i=1}^{n} X_i^T W Y, \]

and we estimate \( \Sigma \) given the values of \( \beta \) as

\[ \hat{\sigma}_{jk} \big|_{\beta} = \frac{1}{n-p} \sum_{i=1}^{n} (y_{ij} - \bar{x}_{ij}\beta)(y_{ik} - \bar{x}_{ik}\beta), \quad 1 \leq j, k \leq m. \]

The procedure is initialized at \( \beta_0 = 0 \), and continues until the difference of the between the current and previous iteration of \( \beta \) is less than \( 10^{-5} \).

Having thus obtained the estimators, \( \hat{\beta}_n \) and \( \hat{\Sigma}_n \), with \( \hat{W}_n = \text{diag}(\hat{\Sigma}_n) \), we use the estimators for \( H(\beta) \) and \( J(\beta) \)

\[ \hat{H}_n = n^{-1} \left( \sum_{i=1}^{n} X_i^T \hat{W}_n X_i \right), \]
\[ \hat{J}_n = n^{-1} \left( \sum_{i=1}^{n} X_i^T \hat{W}_n \hat{\Sigma}_n \hat{W}_n X_i \right). \]
3.2 A Multivariate Probit Model

We again assume that the cluster sizes are the same, \( m_i = m, i = 1, \ldots, n \). To build our probit model, suppose that there exists an \( m \times 1 \) latent response variable \( y_i^* = X_i \beta + \epsilon_i \) with \( \epsilon_i \sim N_m(0, \Sigma) \) and \( \Sigma = \sigma R \), where \( R \) is an \( m \times m \) correlation matrix (all off-diagonal entries are equal to \( \rho \), with ones on the diagonal). The observed response variable is then the dichotomized version of the latent variable with \( y_{ij} = I(y_{ij}^* > 0) \), \( j = 1, \ldots, m \). Then

\[
P(y_{ij} = 1|X_i) = \Phi \left( \frac{\bar{X}_{ij} \beta}{\sigma} \right),
\]

where \( \Phi \) denotes the univariate standard normal cumulative distribution function. Clearly, the parameters \( \beta \) and \( \sigma \) are not fully identifiable in the model, and we can only estimate the ratio \( \beta/\sigma \). To simplify notation in what follows, \( \sigma \) is set equal to 1. The univariate composite log-likelihood function of our probit model is then formulated as

\[
CL(\beta; Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[ y_{ij} \log \Phi (\bar{X}_{ij} \beta) + (1 - y_{ij}) \log (1 - \Phi (\bar{X}_{ij} \beta)) \right].
\]

Writing \( \mu_{ij} = \Phi (\bar{X}_{ij} \beta) \), with \( \mu_i = (\mu_{i1}, \ldots, \mu_{im})^T \), we compute

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

where \( \Pi_i = \text{diag}(\text{var}(y_{i1}), \ldots, \text{var}(y_{im})) \), and \( \text{var}(y_{ij}) = \mu_{ij}(1 - \mu_{ij}) \). This leads to

\[
H(\beta) = n^{-1} \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T \Pi_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right),
\]

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

To find the \( \hat{\beta}_n^c \), we use the Newton-Raphson algorithm, initialized at \( \beta_0 = 0 \), and iterating until the difference of the between the current and previous value of \( \beta \) is less than \( 10^{-5} \). Occasionally, intermediate steps of the algorithm result in large values of \( |\bar{X}_{ij} \beta| \), which makes the computation unstable. To avoid this problem, we adopt the solution suggested by [Demidenko (2001)](https://link.springer.com/article/10.1023/A:1004155220635), based on the following limits for the standard normal density function

\[
\lim_{s \to -\infty} \frac{\phi(s)}{s \Phi(s)} = -1, \quad \lim_{s \to \infty} \frac{\phi(s)}{s(1 - \Phi(s))} = 1.
\]
For further details, we refer to Demidenko (2001). Then \( \hat{\mu}_{ijn} = \Phi(\vec{x}_{ij}\hat{\beta}_n^c) \).

Let \( \hat{\Pi}_{in} \) denote the estimator of \( \Pi_i \) obtained by substituting \( \hat{\mu}_{ijn} \) for \( \mu_{ijn} \).

We estimate \( H(\beta) \) and \( J(\beta) \) as

\[
\hat{H}_n = n^{-1} \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \beta} \right) T \hat{\Pi}_{in}^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right),
\]

\[
\hat{J}_n = n^{-1} \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \beta} \right) T \hat{\Pi}_{in}^{-1} \hat{\text{cov}}_n(y_i) \hat{\Pi}_{in}^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right),
\]

with empirical variance \( \hat{\text{cov}}_n(y_i) = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T \), where \( \hat{\mu}_i = \{\hat{\mu}_{i1n}, \hat{\mu}_{i2n}, \ldots, \hat{\mu}_{imn}\}^T \), where \( \hat{\mu}_i \) is the vector \( \hat{\mu}_i = \Phi(X_i\hat{\beta}_n^c) \).

### 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 approach, the binary observations take values \( y_{ij} \in \{-1, 1\} \) and the joint distribution is then

\[
f_Y(y_i) \propto \exp\left\{ \sum_{j=1}^{m_i} \mu^*_i y_{ij} + \sum_{j<j'} w^*_i y_{ij} y_{ij'} \right\},
\]

where \( \mu^*_i \) is a parameter which describes the main effect of the measurements and \( w^*_i \) describes the association between pairs of measurements within the cluster \( y_i \). Independence corresponds to the case that \( w^*_i = 0 \) and positive or negative correlation corresponds to \( w^*_i > 0 \) or \( w^*_i < 0 \), respectively.

For simplicity, we consider the case that \( \mu^*_i = \mu_i^* \) and \( w^*_i = w_i^* \), noting that the methodology can be readily applied to the arbitrary scenario.

Under this simplification, Molenberghs and Ryan (1999), showed that the joint distribution can be equivalently written in terms of \( z_i = \sum_{j=1}^{m_i} 1(\gamma_{ij} = 1) \) (the number of successes in the \( i \)th cluster) as follows

\[
f_Y(y_i) \propto \exp\{\mu_i z_i - w_i^* z_i (m_i - z_i)\}, \quad (3.5)
\]

where \( w_i = 2w_i^* \) and \( \mu_i = 2\mu_i^* \).

Specifying the normalizing constant in (3.5) is difficult, but also necessary to compute the full likelihood function. It is therefore desirable to use an alternative formulation, 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. For each cluster, there are \( m_i \) conditional probabilities to consider. We now define two conditional probabilities

\[ p_{is} = \frac{\exp\{\mu_i - w_i(m_i - 2z_i + 1)\}}{1 + \exp\{\mu_i - w_i(m_i - 2z_i + 1)\}}, \]

\[ p_{if} = \frac{\exp\{-\mu_i + w_i(m_i - 2z_i - 1)\}}{1 + \exp\{-\mu_i + w_i(m_i - 2z_i - 1)\}}. \]

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 also 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

\[ CL(\mu, w; Y) = \sum_{i=1}^{n} (z_i \log p_{is} + (m_i - z_i) \log p_{if}). \] (3.6)

The special form of the composite likelihood (3.6) means that a logistic regression approach can be used to estimate the parameters. Modelling a covariate effect is achieved 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 \( y_{ij} = -1 \). Let \( X_{new}^i \) denote the thus inflated covariate matrix, and \( \beta^\text{new} = \{\beta, w\} \). This allows us to obtain CMLE estimates of both \( \beta \) and \( w \) using iterative re-weighted least squares, commonly used to solve logistic regression maximization problems.

The maximum composite likelihood estimate of \( \beta \) can be obtained from the iterative re-weighted least squares algorithm, commonly used to solve logistic regression maximization problems. To estimate the covariance of \( \hat{\beta}_n \), we computed \( \hat{J}_n \) as the empirical variance of the score vector, \( \frac{\partial CL}{\partial \beta} \), plugging in estimates of \( \mu_1^*, w^* \) throughout. \( H \) 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 via simulations. In Section 3 we describe regression settings for each of the three models considered here. For this set-up, we test two different global null hypotheses on the regression coefficients $\beta_1, \cdots, \beta_p$ (a) many-to-one comparisons, $H_{01} : \cap_{i=2}^p \{ \beta_1 = \beta_i \}$, and (b) all pairwise comparisons $H_{02} : \cap_{1 \leq i,j \leq p, i \neq j} \{ \beta_i = \beta_j \}$. The two hypotheses are tested using the test statistic developed in (2.2), with an appropriate choice of contrast matrix. The CMLE $\hat{\beta}_n^c$ is obtained via composite likelihood and estimated for each model as described in Section 3.

To perform multiple testing, four different critical values were used for the test statistics. We compared the Bonferroni, the Dunn-Sidak (Sidak, 1968), Holm (Holm, 1979), and MNQ (Hothorn et al., 2008a) approaches. For MNQ, the critical values can be obtained using the R package mvtnorm (Hothorn et al., 2008b), based on the limiting multivariate distribution of the test statistics.

Table 1: Multiple comparison methods considered:

| CASE | multiple comparison method | $\hat{\Gamma}_n$ |
|------|-----------------------------|------------------|
| (a)  | MNQ                         |                  |
| (b)  | Bonferroni                  | $\hat{H}_n^{-1}$ |
| (c)  | Dunn-Sidak                  | $\hat{J}_n$ $\hat{H}_n^{-1}$ |
| (d)  | Holm                        |                  |
| (e)  | MNQ “naive”                 | $\hat{H}_n^{-1}$ |

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 assumed within the clusters. Due to the specific composite likelihood methods we use (independent marginals and univariate conditionals), such a misspecification is equivalent to erroneously believing that $H(\theta) = J(\beta)$ in (2.1). This results in an estimate of $\hat{\Gamma}_n = \hat{H}_n^{-1}$ 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 \( \hat{\Gamma}_n = \hat{H}_n^{-1} \). Throughout, it is referred to as the “naive” approach. Overall, we therefore consider five different approaches, and these are provided in Table 1.

For each setting, \( B = 1000 \) simulated data sets were generated. The family-wise type I error rate was set to be 0.05. Tables 2, 3, and 4 summarize the empirical FWER of the different methods. Empirical power in each scenario (for a specific alternative, as given below) is also provided. In the tables, the SE for the FWER is approximately 0.007. To make it easier to read, in each scenario, we highlight (using bold font) the simulation which of the FWER that are not statistically different from 0.05, has the largest power.

4.1 Multivariate Normal

We assume that we observe \( n = 200 \) clusters with \( m_i = m = 4 \) correlated observations within each cluster. The number of covariates \( p \) was set to \( p = 10 \). Overall, five different \( \Sigma \) were considered:

1. four exchangeable structures with \( \sigma^2 = 0.8 \) and four different values for \( \text{cov}(y_{ij}, y_{ik}) = \rho \). The different values for \( \rho \) are given by 0, 0.2, 0.5 and 0.8.

2. one arbitrary structure, where

\[
\Sigma = \begin{pmatrix}
1.3 & 0.9 & 0.5 & 0.3 \\
0.9 & 1.9 & 1.3 & 0.3 \\
0.5 & 1.3 & 1.3 & 0.1 \\
0.3 & 0.3 & 0.1 & 0.7
\end{pmatrix}.
\]

(4.7)

In each simulation, the \( m \times p \) covariate matrix \( X_i \) was obtained by randomly sampling each entry from a hierarchical normal distribution. That is, the normal distribution had mean randomly chosen from \( \{-3, -2.5, \ldots, 2, 2.5, 3\} \) and variance randomly chosen from \( \{3, 3.6, 4.2, 4.8, \ldots, 6\} \). We also assumed that under \( H_0 \) the true value of the regression parameters was \( \beta = 0 \), and power was calculated under alternative \( \hat{\beta}_a^T = (0.002, 0.008, 0, 0, -0.03, 0, 0.005, 0.02, -0.010, 0) \).

Table 2 summarizes the results of the simulations. Overall, it is easy to see that method (a) which utilizes MNQ and correctly accounts for the intra-cluster correlations, has the best performance. A comparison of columns (a) and (e) clearly shows the cost of ignoring these correlations: the FWER of (a) is superior to that of (e) for \( \rho \neq 0 \) (when \( \rho = 0 \) the methods in (a) and
Table 2: Normal example: empirical results for FWER and power (in brackets) for $n = 200$.

| $H_0$          | $\Sigma \sigma^2$ | (a)         | (b)         | (c)         | (d)         | (e)         |
|----------------|-------------------|-------------|-------------|-------------|-------------|-------------|
|                | $\rho$            | 0.053       | 0.037       | 0.037       | 0.037       | 0.054       |
| many to one    | 0                 | (0.894)     | (0.875)     | (0.877)     | (0.875)     | (0.894)     |
|                | 0.2               | 0.052       | 0.037       | 0.037       | 0.037       | 0.070       |
|                | (0.875)           | (0.840)     | (0.842)     | (0.840)     | (0.909)     |
|                | 0.5               | 0.056       | 0.045       | 0.047       | 0.045       | 0.109       |
|                | (0.852)           | (0.809)     | (0.811)     | (0.809)     | (0.932)     |
|                | 0.8               | 0.045       | 0.035       | 0.035       | 0.035       | 0.127       |
|                | (0.798)           | (0.755)     | (0.762)     | (0.755)     | (0.918)     |
| all pairwise   | 0                 | 0.055       | 0.046       | 0.047       | 0.046       | 0.056       |
|                | (0.766)           | (0.720)     | (0.726)     | (0.721)     | (0.669)     |
|                | 0.2               | 0.051       | 0.042       | 0.042       | 0.042       | 0.076       |
|                | (0.998)           | (0.995)     | (0.995)     | (0.995)     | (0.999)     |
|                | 0.5               | 0.050       | 0.036       | 0.036       | 0.036       | 0.107       |
|                | (0.990)           | (0.985)     | (0.986)     | (0.985)     | (0.997)     |
|                | 0.8               | 0.050       | 0.037       | 0.037       | 0.037       | 0.139       |
|                | (0.985)           | (0.979)     | (0.979)     | (0.979)     | (0.999)     |
|                | see (4.7)         | 0.050       | 0.039       | 0.040       | 0.039       | 0.029       |
|                | (0.988)           | (0.986)     | (0.986)     | (0.986)     | (0.980)     |

(e) are almost identical. Notably, the power in (e) is occasionally higher than that in (a), however, this is only due to the over-inflation of the FWER in (e); overall, (a) exhibits better test performance.
Table 3: Probit example: empirical results for FWER and power (in brackets) for $n = 200$.

| $H_0$ | $\rho$ | (a)   | (b)   | (c)   | (d)   | (e)   |
|-------|--------|-------|-------|-------|-------|-------|
|       |        |       |       |       |       |       |
|       | 0      | 0.055 | 0.043 | 0.039 | 0.048 | 0.911 |
|       |        | (0.911) | (0.885) | (0.888) | (0.885) | (0.911) |
|       | 0.2    | 0.056 | 0.045 | 0.045 | 0.045 | 0.062 |
|       |        | (0.907) | (0.885) | (0.886) | (0.886) | (0.916) |
|       | 0.5    | 0.052 | 0.043 | 0.044 | 0.043 | 0.079 |
|       |        | (0.862) | (0.843) | (0.844) | (0.843) | (0.912) |
|       | 0.8    | 0.064 | 0.043 | 0.043 | 0.044 | 0.107 |
|       |        | (0.807) | (0.780) | (0.783) | (0.780) | (0.899) |
|       | 0      | 0.059 | 0.042 | 0.042 | 0.042 | 0.054 |
|       |        | (0.995) | (0.992) | (0.992) | (0.992) | (0.994) |
|       | 0.2    | 0.069 | 0.047 | 0.049 | 0.047 | 0.071 |
|       |        | (0.989) | (0.988) | (0.988) | (0.988) | (0.991) |
|       | 0.5    | 0.069 | 0.049 | 0.050 | 0.049 | 0.095 |
|       |        | (0.986) | (0.985) | (0.982) | (0.985) | (0.992) |
|       | 0.8    | 0.054 | 0.044 | 0.044 | 0.044 | 0.116 |
|       |        | (0.972) | (0.963) | (0.963) | (0.963) | (0.988) |

4.2 Multivariate Probit Model

We next consider the multivariate probit model. For our simulations, bivariate response variables for $n = 200$ clusters were generated by dichotomizing multivariate normally distributed latent variables with a threshold of zero as discussed in Section 3.2. For each cluster, an $m \times p$ covariate matrix $X_i$, with $m = 4$ and $p = 10$, was obtained by randomly sampling each entry from a hierarchical normal distribution. That is, the normal distribution had mean randomly chosen from $\{-3, -2.5, \ldots, 2, 2.5, 3\}$ and variance randomly chosen from $\{3, 3.6, 4.2, 4.8, \ldots, 6\}$. The regression coefficient under the null was $\beta^T = 0$ with the alternative $\beta^T = (0, 0, 0.05, 0.01, -0.002, 0, 0.006, 0, 0, -0.05)$. The latent variable had mean $X_i\beta$ and correlation matrix with $\rho$ on the off-diagonals (recall that we set
\[ \sigma = 1 \]. Here, we considered four different values of \( \rho : 0, 0.2, 0.5, \) and 0.8.

The empirical results are given in Table 3. Here, the results show that although the MNQ method has good performance, it does not appear to be a clear front-runner. Again, note that when \( \rho = 0 \), methods (a) and (e) are almost identical. However, the MNQ approach still continues to have performance which is relatively close to optimal: For example, method (a) for the many-to-one test at \( \rho = 0.8 \) is not shown in bold only because the empirical significance is just slightly above 0.05 + 1.96 \( \times \) SE = 0.0935. We believe that this slight difference is largely due to slower convergence to the asymptotic distribution for the probit model.

### 4.3 Quadratic Exponential Model

Here, \( n = 500 \) clusters are generated from the model described in Section 3.3 and specifically \( \{3.5\} \). It is known that the quadratic exponential model requires a larger sample size to achieve its asymptotic properties, and for this reason we select \( n = 500 \) here. Each cluster contains \( m_i = m \) binary response variables and \( p = 10 \) predictors. The value of \( m \) varies between clusters and was uniformly sampled from \( \{4, 5, 6, 7, 8\} \). The \( m_i \times p \) covariate matrix \( X_i \) was then obtained in each case by randomly sampling each entry from a standard normal distribution. We consider four different values for the interaction parameter \( w: \{0, 0.2, 0.5, 0.8\} \). The null value of the regression coefficient is set to \( \beta^T = 0 \) with alternative set to \( \beta^T_a = (0.05, 0.07, -0.045, 0, -0.05, 0, 0.035, -0.07, -0.09, 0.06) \).

The empirical FWER and power are computed and summarized in Table 4. Overall, MNQ has clearly the best performance. It is interesting that for larger positive values of \( w \), the performance of the naive method drops sharply to zero, and we posit that this has to do with ferromagnetic nature of the model with these parameters. The antiferromagnetic setting, on the other hand, has similar behaviour for the FWER as the previous two models, but much lower power overall.

### 5 Analysis of Kidney Function Data

Diabetic nephropathy (DN) is damage to the kidneys, caused by the destruction of the kidney’s blood vessels by high blood sugar levels. DN is a serious complication of diabetes, and a common cause of chronic kidney failure and end-stage kidney disease. At the University of Michigan, a study was performed to determine if any biomarkers, from among 500 candidate
Table 4: Quadratic exponential example: empirical results for FWER and power (in brackets) for $n = 500$.

| $H_0$  | $w$  | (a)    | (b)    | (c)    | (d)    | (e)    |
|--------|------|--------|--------|--------|--------|--------|
|        |      | 0.049  | 0.041  | 0.041  | 0.371  |        |
|        |      | (0.271)| (0.236)| (0.237)| (0.236)| (0.743)|
|        | -0.2 | 0.048  | 0.042  | 0.042  | 0.193  |        |
|        |      | (0.464)| (0.420)| (0.425)| (0.421)| (0.727)|
|        | 0    | 0.058  | 0.047  | 0.048  | 0.062  |        |
|        |      | (0.662)| (0.620)| (0.624)| (0.620)| (0.680)|
|        | 0.2  | 0.052  | 0.042  | 0.043  | 0.042  | 0.003  |
|        |      | (0.887)| (0.859)| (0.862)| (0.859)| (0.514)|
|        | 0.5  | 0.062  | 0.045  | 0.047  | 0.047  | 0.000  |
|        |      | (0.857)| (0.838)| (0.838)| (0.839)| (0.037)|
|        | 0.8  | 0.061  | 0.050  | 0.052  | 0.050  | 0.000  |
|        |      | (0.398)| (0.363)| (0.366)| (0.365)| (0.001)|
|        |      | 0.046  | 0.034  | 0.034  | 0.034  | 0.513  |
|        |      | (0.399)| (0.337)| (0.341)| (0.337)| (0.941)|
|        | -0.2 | 0.051  | 0.044  | 0.044  | 0.044  | 0.263  |
|        |      | (0.613)| (0.556)| (0.558)| (0.556)| (0.904)|
|        | 0    | 0.052  | 0.035  | 0.036  | 0.035  | 0.064  |
|        |      | (0.843)| (0.807)| (0.808)| (0.807)| (0.854)|
|        | 0.2  | 0.047  | 0.037  | 0.037  | 0.037  | 0.000  |
|        |      | (0.983)| (0.975)| (0.977)| (0.975)| (0.664)|
|        | 0.5  | 0.051  | 0.041  | 0.042  | 0.041  | 0.000  |
|        |      | (0.963)| (0.952)| (0.953)| (0.952)| (0.019)|
|        | 0.8  | 0.061  | 0.046  | 0.047  | 0.046  | 0.000  |
|        |      | (0.555)| (0.502)| (0.505)| (0.502)| (0.000)|

genes, have important influence on the risk of DN, as part of a therapeutic program. This data set was also previously considered in ?.

The glomerular filtration rate (GFR) is a measurement of the flow rate
of filtered fluid through the kidney, and as such provides an indication of the health of the kidney. Normal results vary between 90-120 mL/min/1.73 m², with some sources indicating an ideal GFR of about 100 mL/min/1.73 m² (?). Abnormal GFR results correspond to lower values. In the study, the GFR of 35 DN abnormal patients was measured at multiple time points, and the number of these time measurements varied from patient to patient. The number of repeated measurements varied between 10 to 15 for each patient, and in total 402 measurements were collected. In addition, each patient’s renal tissue underwent a micro-array analysis providing expression levels of the 500 genes. A binary outcome variable indicating the treatment type was also recorded.

The quadratic exponential model is one natural choice to analyze this data. The original GFR was dichotomized by setting \( y_{ij} = 1 \) if the patient’s GFR at time point \( i \) was at least 100 mL/min/1.73 m², and \( y_{ij} = -1 \) otherwise, with \( j = 1, \ldots, m_i \) and \( i = 1, \ldots, 35 \). There were a few of missing in recording GFR in the original data set (about 12% of the observations). In such a case, we imputed the missing value as the median of the remaining GFR values for that patient. The observed covariates, \( X_i \), for each patient are composed of the 500 gene expression levels as well as the treatment type. Because we are using the reduced model (3.5), we use logistic regression and incorporate the \( w \) parameter as discussed in Section 3.3. In order to shrink the number of covariate and avoid over-fitting, we first run a generalized regression with an \( L_1 \) constraint (?) using the R package \texttt{lasso2} (?). This process reduces the number of covariates from the (augmented) 502 to 9. The shrinkage procedure unfortunately removes both the treatment and augmented \( w \) parameters.

Next, we compared the effect of these 9 genes on the dichotomized GFR response. Due to client’s interest, we re-introduced the treatment effect, as well as the augmented \( w \) parameter to help account for the within-patient correlations. Here, we compare all pair-wise gene effects for the 9 genes, \( H_{0,ij} = \{ \beta_i = \beta_j \} \), for a total of 36 null hypotheses. Of the five methods described in Table 1, we opt for the MNQ approach based on its superior performance in Section 4. To show how drastically different the results will be if the within-patient correlations are ignored, we also compare this with the MNQ “naive” version. The hypothesis test based on the MNQ method rejected three hypotheses:

\[ \beta_3 = \beta_4, \beta_3 = \beta_6, \beta_4 = \beta_7. \]

Each of these test were also rejected by the MNQ naive method (and the
three were in the top 5 most significant), but this approach rejected in total 29 hypotheses out of the possible 36.

In order to explain the drastic difference between the two methods, recall that there is correlation between the repeated measurements across the time points. The interaction is shown to be fairly strong with \( \hat{w}_n = -0.49 \). 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 (cf. Table 1), leading to erroneous results.

6 Discussion

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

Appendix

Xu and Reid (2011) provide 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
The marginal densities of $y_{ij}$ have common support for all $\theta$.

(A3). The true value $\theta_0$ is an interior point of $\Omega$, the space of possible values of the parameter $\theta$.

(A4). Let $\alpha$ 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 $\theta_0$. Moreover, there exists a constant $c$ and an integrable function $M(y)$ such that

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

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

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

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

**Theorem 6.1.** Under the regularity conditions (A1)-(A6), there exists a solution to the composite likelihood equation, $\widehat{\theta}_n^c$, which satisfies

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

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 \( h(\theta; y) = \text{cl}(\theta; y) \). Note that for fixed \( y \), \( h \) maps \( \mathbb{R}^p \) into \( \mathbb{R} \). Then, by a Taylor expansion, we have that
\[
h(\theta; y) - h(\theta_0; y) = (\nabla h)(\theta_0; y)^T (\theta - \theta_0) + (\theta - \theta_0)^T (Dh)(\theta^*; y)(\theta - \theta_0),
\]
where \( \theta^* \) lies on a line joining \( \theta \) and \( \theta_0 \). We use \( \nabla, D \) to denote the gradient and Hessian operators, respectively. Our goal will be to show that there exists a \( \theta \) in a \( n^{-1/2} \) ball of \( \theta_0 \), the left hand side of the above equation is negative. This in turn will imply that there exists a CMLE which satisfies \( \sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1) \).

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
\[
\xi^T \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\nabla \text{cl}_m)(\theta_0, y_i) + \frac{1}{n} \sum_{i=1}^{n} (H \text{cl}_m)(\theta^*, y_i) \right\} \xi \\
\equiv \xi^T b_n M + \xi^T B_n \xi M^2,
\] (6.8)
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 \( M_\epsilon \), such that
\[
\limsup_n P \left( \sup_{\xi \in S} \xi^T b_n / \lambda_n^{(p)} < M_\epsilon \right) \geq 1 - \epsilon.
\]
Then, by \( 6.8 \), 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
\[
\limsup_n P \left( \xi^T b_n M_\epsilon + \xi^T B_n \xi M^2 < 0 \quad \forall \xi \in S \right) \geq 1 - \epsilon.
\]

Note that if \( \xi^T b_n M_\epsilon + \xi^T B_n \xi M^2 < 0 \quad \forall \xi \in S \), then, by the above and continuity of \( \text{cl}_m \), 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 $B_{M, \sqrt{n}}(\theta_0) \cap B(\theta_0)$. This implies that there exists a $\hat{\theta}_n^c$ which satisfies
\[ \sqrt{n}(\hat{\theta}_n^c - \theta_0) = O_p(1). \]

Let $g(\theta; y) = cl_m^{(1)}(\theta; y) = \nabla cl_m(\theta; y)$ (this is the vector of first derivatives), then using a multivariate Taylor expansion, we have that
\[ g(\hat{\theta}_n^c; y) = g(\theta_0; y) + \sum_{|\alpha| \leq 1} (\partial^\alpha g)(\theta_0; y)(\hat{\theta}_n^c - \theta_0)^\alpha \]
\[ + \sum_{|\alpha| = 2} \frac{2}{\alpha!} (\hat{\theta}_n^c - \theta_0)^\alpha \int_0^1 (1 - t)(\partial^\alpha g)(\theta_0 + t(\hat{\theta}_n^c - \theta_0); y) dt, \]
again using the multi-index notation. We take $\hat{\theta}_n^c$ to be the local maximizer found above. This time, for fixed $y$, $g$ maps $\mathbb{R}^p$ into $\mathbb{R}$, 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 $\sum_{i=1}^n cl_m^{(1)}(\hat{\theta}_n^c; y_i) = 0$, we have that
\[ \frac{1}{\sqrt{n}} \sum_{i=1}^n (Dcl_m)(\theta; y_i)^T(\hat{\theta}_n^c - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n R_{n,i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(\theta_0; y_i). \quad (6.9) \]

By condition (A4), we have that
\[ \left| \sum_{|\alpha| = 2} \frac{2}{\alpha!} (\hat{\theta}_n^c - \theta_0)^\alpha \int_0^1 (1 - t)(D^\alpha g)(\theta_0 + t(\hat{\theta}_n^c - \theta_0); y) dt \right| \]
\[ \leq \sum_{|\alpha| = 2} \frac{1}{\alpha!} |\hat{\theta}_n^c - \theta_0|^\alpha |M(y)|, \]
from which it follows that,
\[ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n R_{n,i} \right| \leq \left\{ \sqrt{n}||\hat{\theta}_n^c - \theta_0||_2 \right\} \left\{ \frac{1}{n} \sum_{i=1}^n |M(y_i)| \right\} . \]

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
\[ \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n cl_m^{(2)}(\theta; y_i) - H(\theta_0) \right\} (\hat{\theta}_n^c - \theta_0). \]
By similar argument to that above, this is also $o_p(1)$. This allows us to re-write (6.9) as

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

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. \(\square\)

References

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.

DEMIDENKO, E. (2001). Computational aspects of probit model. Math. Commun. 6 233–247.

FEARNHEAD, P. and DONNELLY, P. (2002). Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 657–680.

FISHER, R. (1935). The design of experiments. Edinburgh and London: Oliver and Boyd.

GABRIEL, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. Ann. Math. Statist. 40 224–250.

GEYS, H., MOLENBERGHS, G. and RYAN, L. M. (1997). Pseudo-likelihood inference for clustered binary data. Comm. Statist. Theory Methods 26 2743–2767.

HEAGERTY, P. J. and LELE, S. R. (1998). A composite likelihood approach to binary spatial data. J. Amer. Statist. Assoc. 93 1099–1111.
Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics. Scand. J. Statist. 21 289–357. With discussion and a reply by the authors.

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.

Li, Y. and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. J. Amer. Statist. Assoc. 101 591–603.

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.

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.

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

Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. Scand. J. Statist. 28 295–302.
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.

Scheffe (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.

Tukey, J. W. (1953). The problem of multiple comparisons. Mimeoraphed monograph.

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.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica 50 1–25.

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.