Aspects of Composite Likelihood Inference

Zi Jin

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Statistics
University of Toronto

©Copyright 2010 Zi Jin
Aspects of Composite Likelihood Inference

Zi Jin

Doctor of Philosophy, Department of Statistics
University of Toronto, June 2010

Abstract

A composite likelihood consists of a combination of valid likelihood objects, and in particular it is of typical interest to adopt lower dimensional marginal likelihoods. Composite marginal likelihood appears to be an attractive alternative for modeling complex data, and has received increasing attention in handling high dimensional data sets when the joint distribution is computationally difficult to evaluate, or intractable due to complex structure of dependence.

We present some aspects of methodological development in composite likelihood inference. The resulting estimator enjoys desirable asymptotic properties such as consistency and asymptotic normality. Composite likelihood based test statistics and their asymptotic distributions are summarized. Higher order asymptotic properties of the signed composite likelihood root statistic are explored.

Moreover, we aim to compare accuracy and efficiency of composite likelihood estimation relative to estimation based on ordinary likelihood. Analytical and simulation results are presented for different models. The full efficiency with identical estimators compared to the full likelihood is shown for intraclass correlation normal and further extended to unrestricted multivariate normal distributions. For autoregressive models, we find that the composite likelihood method loses very little to the ordinary likelihood method, and appears to be large sample equivalent. When applied to correlated binary data, the pairwise likelihood
approach tends to outperform the full likelihood approach, as it provides more accurate estimates, demonstrates higher efficiency, and is less computational intensive. Furthermore the pairwise likelihood approach is more robust to model misspecifications.
Acknowledgements

I would foremost like to thank my supervisor Nancy Reid for her outstanding guidance, her noble knowledge, for the numerous inspiring discussions, for her prevision, and for her patience and encouragement. Her expertise and guidance will definitely have a deep impact in the way I approach statistics in my future career. I also thank her for providing me with various valuable opportunities and generous financial support.

My sincere thanks extend to the Department of Statistics for offering a friendly environment for my professional and personal time at Toronto. I am very grateful to the professors in the department who have taught me statistics and shared their wisdom in many occasions with me. The Department’s faculty, staff members, postdoctoral fellows and student colleagues have always been a source of friendships and good advice. It was a memorable fun time of growth and exploration.

I wish to thank Professor Nils Lid Hjort for kindly sharing with me his manuscript regarding the autoregressive model, and Professor Grace Y. Yi for helpful advice on model simulation in Chapter 5. I would also thank my external appraiser Peter Song for providing constructive comments and suggestions.

For my parents for being always so enthusiastic about research and statistics.

For my loving, encouraging, and patient Ben whose endless positive vibe and love during the final stages of this Ph.D are my eternal sunshine. Thank you.
Contents

1 Introduction

1.1 Background and motivation ........................................ 1

1.1.1 Motivation for composite likelihood .......................... 3

1.1.2 Some previous work and applications .......................... 4

1.2 Likelihood inference and asymptotics .............................. 8

1.2.1 Likelihood asymptotics ......................................... 11

1.3 Main results and Outline ........................................... 12

2 Composite Test Statistics and Higher Order Asymptotics .... 17

2.1 Definitions and derived quantities ................................. 17

2.2 Maximum composite likelihood estimator ....................... 20

2.3 Composite Likelihood Test Statistics ............................. 23

2.3.1 Composite likelihood test statistics, with no nuisance parameter 23

2.3.2 Composite test statistics for $\theta = (\psi, \lambda)$ ............... 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4 Higher accuracy inference in terms of $\tilde{r}$ under simple null hypothesis</td>
<td>28</td>
</tr>
<tr>
<td>2.4.1 Maximum composite likelihood estimation</td>
<td>29</td>
</tr>
<tr>
<td>2.4.2 Composite likelihood ratio statistic</td>
<td>34</td>
</tr>
<tr>
<td>2.4.3 Signed composite likelihood root statistic and its approximation</td>
<td>36</td>
</tr>
<tr>
<td>2.5 Example</td>
<td>43</td>
</tr>
<tr>
<td>3 Multivariate normal distribution</td>
<td>47</td>
</tr>
<tr>
<td>3.1 Intraclass Correlation</td>
<td>47</td>
</tr>
<tr>
<td>3.1.1 Maximum estimate of full likelihood</td>
<td>48</td>
</tr>
<tr>
<td>3.1.2 Expected Fisher information</td>
<td>49</td>
</tr>
<tr>
<td>3.1.3 Maximum estimate of pairwise likelihood</td>
<td>52</td>
</tr>
<tr>
<td>3.1.4 Godambe information</td>
<td>53</td>
</tr>
<tr>
<td>3.1.5 Maximum estimate of pseudolikelihood</td>
<td>59</td>
</tr>
<tr>
<td>3.2 General Unrestricted Multivariate Normal</td>
<td>60</td>
</tr>
<tr>
<td>3.2.1 MLE and Fisher information</td>
<td>61</td>
</tr>
<tr>
<td>3.2.2 MPLE and Godambe information</td>
<td>62</td>
</tr>
<tr>
<td>3.3 Discussions</td>
<td>66</td>
</tr>
<tr>
<td>4 AR model</td>
<td>70</td>
</tr>
<tr>
<td>4.1 The full likelihood</td>
<td>71</td>
</tr>
</tbody>
</table>
4.2 The composite likelihood .................................................. 73

4.3 Godambe matrix ............................................................... 75

4.3.1 Steps to derive $J(\theta)$ .............................................. 76

4.3.2 Relative efficiency .......................................................... 81

4.4 Analytical efficiency results ............................................... 82

4.5 Simulation study ............................................................... 84

4.5.1 MLE and MPLE .............................................................. 84

4.5.2 Simulation results ........................................................... 87

4.6 Discussion ........................................................................... 93

5 Correlated binary data ......................................................... 95

5.1 Bahadur representation ....................................................... 96

5.2 Model set-up and marginal distributions ............................... 98

5.3 Estimation procedures ....................................................... 101

5.3.1 Full likelihood .............................................................. 101

5.3.2 Pairwise likelihood ......................................................... 102

5.3.3 Independence case ........................................................ 104

5.4 Simulated data I – True model ........................................... 106

5.5 Simulated data II – Working model ..................................... 107
5.5.1 The algorithm ................................................. 109

5.6 Simulation study .................................................. 112

5.6.1 Simulation design .............................................. 112

5.6.2 Simulation results ............................................. 117

Bibliography ............................................................. 142
List of Figures

3.1 Likelihood functions of $\rho$ with $n = 30$ and $q = 3$. The graphs compare the standardized likelihood curves and the estimators of $\rho$; the full likelihood and MLE (in red), and the pairwise likelihood and MPLE (in green). Note that the scale of $\rho$ values is restricted on the same range suitable for both likelihoods. 68

3.2 Likelihood functions of $\rho$ with $n = 30$ and $q = 8$. The graphs compare the standardized likelihood curves and the estimators of $\rho$; the full likelihood and MLE (in red), and the pairwise likelihood and MPLE (in green). Note that the scale of $\rho$ values is restricted on the same range suitable for both likelihoods. 68

4.1 Efficiency of pairwise likelihood relative to full likelihood for AR(1) model $X_t = 0.55X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 0.1^2)$, as a function of $T$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

4.2 Given $T = 4$, efficiency of pairwise likelihood relative to full likelihood for AR(1) model $X_t = aX_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, \sigma^2)$, as a function of $(a, \sigma)$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83

4.3 Given $T = 20$, efficiency of pairwise likelihood relative to full likelihood for AR(1) model $X_t = aX_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, \sigma^2)$, as a function of $(a, \sigma)$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

4.4 Scatter plot of MLE and MPLE for $a$ (left panel) and $\sigma^2$ (right panel), from 100 simulations for AR(1) model with $(a, \sigma^2) = (0.55, 0.1^2)$ and length $T = 100$. 85
4.5 Histograms of the estimators of $a$, MLE (left panel) and MPLE (right panel), from the same 100 simulations, together with corresponding density curves and normal fits. Note that the scale of $a$ values is the same for both panels.

4.6 Histograms of the estimators of $\sigma^2$, MLE (left panel) and MPLE (right panel), from the same 100 simulations, together with corresponding density curves and normal fits. Note that the scale of $\sigma^2$ values is the same for both panels.
List of Tables

4.1 Relative efficiency of pairwise likelihood relative to full likelihood based on AR(1) model $X_t = 0.55X_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, 0.1^2)$ with different lengths $T$ . . . 82

4.2 AR(1) model, simulation results with $N = 1000, T = 100$ . . . . . . . . . . . . 91

4.3 AR(1) model, simulation results with $N = 1000, T = 20$ . . . . . . . . . . . . 92

4.4 AR(1) model, simulation results with $N = 3, T = 1000$ . . . . . . . . . . . . . 93

5.1 An example of simulated binary random sample with $x_i = (0, 0, 1, 1), \beta_0 = -1.4, \text{ and } \theta^T = (\beta_1, \phi) = (0.7, \log(1.2))$ . . . . . . . . . . . . . . . . . . . . . 107

5.2 Correlations corresponding to different log odds ratios . . . . . . . . . . . . . 113

5.3 Simulation results for $\beta_1$, for $N = 100, m = 4, \beta_1 = 0.7$ . . . . . . . . . 120

5.4 Simulation results for $\phi$, for $N = 100, m = 4, \beta_1 = 0.7$ . . . . . . . . . 121

5.5 Simulation results for $\beta_1$, for $N = 30, m = 4, \beta_1 = 0.7$ . . . . . . . . . 122

5.6 Simulation results for $\phi$, for $N = 30, m = 4, \beta_1 = 0.7$ . . . . . . . . . 123
Chapter 1

Introduction

1.1 Background and motivation

The likelihood function plays an important role in statistical inference, but there are a number of situations where some modification of the likelihood is needed. In the development of an approach to modeling that does not require specifying a full probability distribution for the data, variations of the likelihood function have been addressed.

Cox (1972b, 1975) proposed partial likelihood for fitting proportional hazards model for failure time data, which only models part of the observed data to make the problem tractable. A related idea was suggested by Besag (1974) for complex spatial models, using the product of conditional distributions as a pseudolikelihood function. The name composite likelihood was given in Lindsay (1988) to refer to a likelihood type object formed by multiplying together
individual component likelihoods, each of which corresponds to a marginal or conditional event, formulated, for a random variable $Y$ with density function $f(y; \theta)$, as

$$\text{CL}(\theta; y) = \prod_{k \in K} L_k(\theta; y)^{w_k},$$  \hfill (1.1)

where $L_k(\theta; y) = f(\{y \in A_k\}; \theta)$ is the likelihood function for an event $A_k$, $\{w_k, k \in K\}$ is a set of weights and $\{A_k, k \in K\}$ is a set of events. It groups a rich class of pseudolikelihoods based on likelihood objects, and the full likelihood function is a special case.

Cox and Reid (2004) investigated a version of this where the components are lower dimensional marginal densities, which is called composite marginal likelihood to distinguish it from the composite likelihood built by conditional distributions. For a $q$-dimensional variable $Y$, they considered first- and second-order log-likelihoods

$$l_1(\theta; Y) = \sum_{s=1}^{q} \log f(Y_s; \theta),$$  \hfill (1.2)

$$l_2(\theta; Y) = \sum_{s>t}^{q} \log f(Y_s, Y_t; \theta) - aql_1(\theta; Y),$$  \hfill (1.3)

where $a$ is to be specified, possibly according to some optimality criterion. When $a = 0$, $l_2$ is called the pairwise log-likelihood function; $a = 1/2q$ corresponds in effect to taking all possible conditional distributions of one component given another. Hence, the formulation of combining marginal distributions also allows the flexibility to represent both marginal and conditional likelihoods.
1.1.1 Motivation for composite likelihood

The composite likelihood method is statistically satisfactory in the sense that it is as generally applicable as the method of maximum likelihood. One motivation for composite likelihood estimation is to replace the ordinary likelihood by some other likelihood objects which are feasible or easier to evaluate, and hence to maximize. An example is the problem inherent in generalized linear mixed models (GLMMs) that the marginal likelihood function obtained after integrating over random effects involves intractable integrals. Renard et al. (2004) proposed to use pairwise likelihood estimation and compared it with maximum likelihood, and simulation results suggested that pairwise likelihood approach is a good compromise between computational burden and loss of efficiency. The ease of implementation and substantial computational gain of pairwise likelihood were also shown in Bellio and Varin (2005) for a general class of GLMs with crossed random effects.

Secondly, composite likelihood approach allows us a new tool to access to multivariate distributions when the joint distributions may be intractable or difficult to specify. Aside from mathematical or computational difficulties, it is hard to describe or verify the correct high-dimensional joint distribution of data, especially when complex dependencies are involved, whereas the bivariate marginal distributions are relatively easy to specify and obtain. In spite of the fact that the structure of dependence higher than the pairwise association is difficult to interpret, pairwise correlations are often of substantive interest and importance.

Some examples that illustrate the flexibility of the composite likelihood approach include: using copulas to model correlation of failure times within families in a two-stage procedure.
(Andersen, 2004); and proposing pairwise likelihood inference for the treatment of mixed continuous and non-continuous outcomes, such as mixtures of nominal, ordinal and continuous variables (de Leon, 2005; de Leon and Carriere, 2007).

Moreover, by only modeling lower dimensional distributions, we might expect the resulting estimation to be more robust.

### 1.1.2 Some previous work and applications

Composite likelihood method has received increasing attention in different areas and applications. Variations of composite likelihood have been proposed for the purpose of finding suitable alternatives to likelihood for complex models. The following exposition tries to give a quick and general view of the literature; mainly in the theme the forms of composite likelihood that has been proposed and the corresponding applications. Varin (2008) gave a comprehensive review on the applications of the method.

Varin and Vidoni (2006) defined the composite Kullback-Leibler divergence of a density \( g(y) \) relative to \( f(y) \) as

\[
\Delta c(g, f) = E_f(\log \text{CL}(g; Y)/\text{CL}(f; Y)),
\]

and the expectation, with respect to the true density \( f(y) \), is the difference between the composite log-likelihoods associated with \( g(y) \) and \( f(y) \), respectively. It is a linear combination of ordinary Kullback-Leibler divergences, corresponding to the likelihood objects forming
the composite likelihood function.

The model selection criterion is also discussed in Varin and Vidoni (2006) where they proposed a selection statistic similar to the Akaike (1973) criterion,

$$ AIC = -2 \log CL(\tilde{\theta}_{CL}; y) - 2 \dim(\theta), $$

where $\tilde{\theta}_{CL}$ is maximum composite likelihood estimator, and the effective number of parameters is determined by

$$ \dim(\theta) = \text{tr}\{H(\theta)G^{-1}(\theta)\}, $$

where $H(\theta)$ and $G(\theta)$ are information matrices corresponding to the composite likelihood.

A formal definition of $\tilde{\theta}_{CL}$, $H(\theta)$ and $G(\theta)$ is provided in Chapter 2. Recently, Gao et al. (2009) proposed to use the Bayesian information criterion

$$ BIC = -2 \log CL(\tilde{\theta}_{CL}; y) - \log n \dim(\theta) $$

for selection of tuning parameters.

A lot of exemplifications are available for clustered and longitudinal data. The book by Molenberghs and Verbeke (2005) addressed the methods using the name marginal pseudolikelihood in the analysis of discrete and categorical outcomes. A composite likelihood based on pairwise differences between observations was suggested in Lele and Taper (2002) for inference on large covariance components model for animal breeding and evolutionary
1.1. BACKGROUND AND MOTIVATION

biology,

$$\text{CL}(\theta; Y) = \prod_{i=1}^{K-1} \prod_{j>i} f(Y_i - Y_j; \theta),$$

where $Y$ denotes the vector of observations at $K$ locations, and $K$ may run into the millions in the analysis of large breeding pedigrees. Zhao and Joe (2005) proposed a two-stage procedure for multivariate models which iterates between estimation of marginal parameters from the first-order marginal likelihood and estimation of dependence parameters from the pairwise likelihood evaluated at the estimated marginal parameters.

Weighted composite likelihood has been proposed for clustered data, when cluster sizes are unequal. To model mortality and morbidity of preterm infants, Le Cessie and Van Houwelingen (1994) considered the weighted pairwise log-likelihood over $m$ clusters for the multivariate probit model,

$$pl(\theta; y) = \sum_{i=1}^{m} w_i pl_i(\theta; y) = \sum_{i=1}^{m} w_i \sum_{k>j}^{n_i} \log P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}; \theta),$$

where $w_i, i = 1, \ldots, m$ are cluster specific weights, and are chosen as inversely proportional to the cluster size, i.e. $w_i = 1/(n_i - 1)$. Kuk and Nott (2000) showed the influence of weights on estimating the regression parameters and the association parameters, and suggested that when cluster sizes are unequal a weighted pairwise likelihood should be used for the marginal regression parameters and the unweighted pairwise likelihood should be used for the association parameters. Their result was revisited in Renard et al. (2004) where the pairwise likelihood was applied to multilevel models, and they noted that the optimal score function
for the association parameter is also found to depend on the fixed effects. The choice of weights was carefully studied in Joe and Lee (2009) for clustered and longitudinal data.

A modification of the pairwise likelihood was examined in Kuk (2007), which replaces the composite score functions, for estimating the marginal parameters, by the optimal linear combinations of the marginal score functions.

In presence of missing data, Chen et al. (2009) developed a Markov model for the analysis of longitudinal categorical data which facilitates modelling both marginal and conditional structures. Based on simulation study, they concluded that by using weighted pairwise likelihood it is not necessary to model the missing data mechanism and the simulation results were promising.

In Hjort and Varin (2008), a composite likelihood formed by triplets was investigated in the context of classical Markov chains. Extensions to adopt pairs of observations at lagged distance $m$ are also considered. Their simulation results show that adding pairs separated by longer distance may reduce the efficiency of the estimators.

Varin et al. (2005) applied the pairwise likelihood method to spatial generalized linear mixed models, and in order to maximize the pairwise likelihood introduced an EM-type algorithm. Composite likelihood inference facilitated with an EM algorithm has also been considered by Gao et al. (2007) in gene networks to study the time-course data and reduce the complexity of the dependency structure. A composite conditional likelihood approach is proposed for gene mapping based on linkage disequilibrium in Larribe and Lessard (2008).
A composite likelihood, chosen as a product of likelihood functions for two or three sites at a time, has been applied to the estimation of the recombination rate, whereas some other statistical procedures to deal with missing data, that is, importance sampling and MCMC, require so intensive computing that they are intractable in estimating a full likelihood function based on all the information available in sampled DNA sequences.

1.2 Likelihood inference and asymptotics

To introduce an analogous inference framework for composite likelihood inference, we outline some of the main results in the theory of likelihood inference and asymptotics.

Consider a parametric model $f(y; \theta)$, which is the probability density function for a random variable $Y$. The parameter $\theta$ is assumed to be $d$-dimensional and $Y$ assumed to be $q$-dimensional. The likelihood function of $\theta$ is

$$L(\theta) = L(\theta; y) \propto f(y; \theta),$$

and the log-likelihood function is

$$l(\theta) = l(\theta; y) = \log L(\theta; y).$$
Some derived quantities for the likelihood function are the score function

\[ u(\theta; y) = l'(\theta; y) = \partial l(\theta)/\partial \theta, \quad (1.4) \]

and the observed Fisher information

\[ j(\hat{\theta}) = -\partial^2 l(\theta; y)/\partial \theta^2|_{\theta = \hat{\theta}}. \quad (1.5) \]

Let \( y \) represent a sample of \( n \) independent and identically distributed random variables:

\( y = (Y^{(1)}, \ldots, Y^{(n)}) \). The corresponding log-likelihood function is a sum of the \( n \) independently distributed quantities

\[ l(\theta; y) = \sum_{i=1}^{n} l(\theta; Y^{(i)}). \]

The maximum likelihood estimate is

\[ \hat{\theta} = \arg \sup_{\theta} l(\theta; y), \quad (1.6) \]

and under certain regularity condition, \( \hat{\theta} \) is asymptotically normal,

\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, J^{-1}(\theta)I(\theta)J^{-1}(\theta)), \]

where \( I(\theta) = \text{Var}_{\theta}(u(\theta)) = E_{\theta}(u(\theta)u^T(\theta)) \) is the expected Fisher information matrix, and \( J(\theta) = E_{\theta}(-u'(\theta)) \). Under the usual regularity conditions, the second Bartlett identity gives
\[ I(\theta) = J(\theta), \text{ so} \]
\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta)). \tag{1.7} \]

We have the following limiting results for likelihood based test statistics:

\[ s(\theta) = u^T(\theta)\{j(\hat{\theta})\}^{-1}u(\theta) \xrightarrow{d} \chi^2_d, \tag{1.8} \]
\[ q(\theta) = (\hat{\theta} - \theta)^T\{j(\hat{\theta})\}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_d, \tag{1.9} \]
\[ r^2(\theta) = 2\{l(\hat{\theta}; y) - l(\theta; y)\} \xrightarrow{d} \chi^2_d, \tag{1.10} \]

where \( \chi^2_d \) is the Chi-squared distribution with \( d \) degrees of freedom, and \( d \) is the dimension of \( \theta \), assuming the model satisfies regularity conditions and assuming the consistency of \( \hat{\theta} \).

With nuisance parameter involved, \( \theta = (\psi, \lambda) \), where \( \psi \), \( p \)-dimensional, is the parameter of interest, and \( \lambda \) is the nuisance parameter. Let \( \hat{\lambda}_\psi \) be the constrained maximum likelihood estimator of \( \lambda \) for \( \psi \) fixed,

\[ \sup_{\lambda} l(\psi, \lambda; y) = l(\psi, \hat{\lambda}_\psi; y) = l_p(\psi), \tag{1.11} \]

and \( l_p(\psi) \) is called the profile log-likelihood function. The pair \( (\psi, \hat{\lambda}_\psi) \) is also denoted by \( \hat{\theta}_\psi \). Similar results as above can be derived, for example

\[ 2\{l(\hat{\theta}; y) - l(\hat{\theta}_\psi; y)\} \xrightarrow{d} \chi^2_p, \tag{1.12} \]
where $p$ is the dimension of $\psi$.

1.2.1 Likelihood asymptotics

The limiting results (1.7) - (1.10) give first order approximations; (1.8) - (1.10) are the quadratic forms for the multi-dimensional $\theta$. For scalar $\theta$, we have one-sided forms, especially

$$ r(\theta) = \text{sign}(\hat{\theta} - \theta) \sqrt{2\{l(\hat{\theta}) - l(\theta)\}}, $$

which we call the signed likelihood root statistic. The limiting distribution of $r(\theta)$ under the model $f(y; \theta)$ is the standard normal distribution. The error in the normal approximation to the distribution of $r(\theta)$ is typically of order $O(n^{-1/2})$. A major development in modern statistical inference based on the likelihood function is to provide more accurate approximations than first order. Third order approximations have been presented in different ways in the literature of higher order asymptotics, such as Barndorff-Nielsen’s $p^*$ approximation, Lugannani and Rice approximation, and the modified likelihood root in the $r^*$ type version.

In the special case that $\theta$ is a scalar parameter, the following expression is the most useful for practical purposes (Reid, 2003):

$$ r^* = r + \frac{1}{r} \log \frac{q}{r}, $$

$$ q = \{l_{\hat{\theta}}(\theta) - l_{\hat{\theta}}(\hat{\theta})\} \{j(\hat{\theta})\}^{-1/2}, $$

$$ l_{\hat{\theta}}(\theta; \hat{\theta}, a) = \frac{\partial}{\partial \theta} l(\theta; \hat{\theta}, a), $$
where $r$ is the signed likelihood root statistic defined by (1.13). $a$ is assumed ancillary and the transformation from $y$ to $(\hat{\theta}, a)$ is one to one.

### 1.3 Main results and Outline

The present work focuses on methodological development for composite likelihood inference. We illustrate the fundamental properties of composite likelihood estimators, and different test statistics along with their asymptotic distributions. Higher order asymptotic properties of the signed composite likelihood root statistic are explored. Though the success of composite likelihood method has been shown in a number of applications, very limited results on the theoretical side. This is partially due to the complex applications where composite likelihood has been exploited as a surrogate for ordinary likelihood. We give attention to multivariate distributions where both ordinary likelihood and composite likelihood can be employed and then compared. The properties of the composite likelihood approach are showed for the multivariate normal distribution, time series, and correlated binary data. Besides, our findings provide empirical evidence that pairwise marginal likelihood can be regarded as a fine, simplest and very flexible version of composite likelihood, on top of the principles that analytical simplicity, computational convenience, and more importantly the questions and parameters of interest in general determine the form of the composite likelihood function.

The first part of the thesis aims in building an inference structure for composite likelihood
similar in the framework of ordinary likelihood inference. Chapter 2 presents the theoretical
background and develops results of the maximum composite likelihood estimate, derived
quantities, and information matrices. The consistency and asymptotic normality of the
maximum composite likelihood estimator are given. Moreover, composite likelihood test
statistics, that is composite score statistic, composite Wald statistic and composite likelihood
ratio statistic, and their asymptotic distribution are shown.

An unexplored aspect of composite likelihood inference is the higher order asymptotics.
Limiting results of the maximum composite likelihood estimate and test statistics are proved
in the first order asymptotics. We employ techniques of approximation to construct higher-
order asymptotic approximations to the distribution of the signed composite likelihood root
test statistic.

The second part of the thesis investigates the performance of composite likelihood method
on three models: multivariate normal distribution, times series, and correlated binary data.
We focus on evaluate the efficiency, accuracy and robustness of composite likelihood infer-
ence.

Chapter 3 studies the performance of composite likelihood for the multivariate normal
distribution. The distribution is of great theoretical interest as it is the foundation and
connected to many fields of statistical problems. In Mardia et al. (2007) the full efficiency
of the composite conditional likelihood applied on a general class of normal distribution is
outlined. We proposed the pairwise likelihood function, and proved the analytical results
that this provides identical estimators as MLE and reaches full efficiency for one widely used
1.3. MAIN RESULTS AND OUTLINE

class of multivariate normal distribution - intraclass correlation normal, and furthermore we extended the conclusion to unrestricted multivariate normal distributions.

Chapter 4 examines composite likelihood inference applied to autoregressive process. The motivation arises partly from the correlation structure of the process which owes a special pattern of interest. Moreover, the AR model allows us to consider scenario of a few long sequences. We exploit the fact that a real-valued stationary process \( \{X_t\} \), from a second-order point of view, is characterized by the mean \( \mu \) and autocovariance function \( \gamma(\cdot) \). We focus on specifying the first- and second-order moments of the joint distribution. In particular, when all the joint distributions are multivariate normal, the second-order properties of \( \{X_t\} \) completely determine the joint distributions and hence give a complete probabilistic characterization of the sequence.

Some major results of composite likelihood inference in time series analysis include Hjort and Varin (2008) who studied the quasi likelihood of subsequent pairs along with its extensions - subsequent triplets and so on. Numerical illustrations of pairwise likelihood inference for AOP(1) model (autoregressive ordered probit model with order 1) were given in Varin and Vidoni (2006). Further, asymptotic results are derived in an unpublished manuscript of Hjort (2008). Our contribution to the topic is that we inspect the model paying attention to its curved exponential family characteristic, reach analytical results for variance matrices, implement simulation studies with regard to potential impacts, and hence compare the performance and efficiency of composite likelihood estimation. Overall, based on the analytical and numerical results, we find explanations for the good behaviour of
composite likelihood inference and conclude it is a suitable approach for the autoregressive model.

Chapter 5 investigates correlated binary data. Pairwise likelihood appears to be an attractive alternative for modeling correlated binary data where the joint distributions are usually intractable. The pairwise likelihood method is expected to be able to capture essential features in data, and has a high flexibility, while the loss of information is not large.

We assume independent experiment individuals. The response measurements from the same individual can be highly correlated, and they can be continuous, or discrete, and usually associated with a vector of covariates or explanatory variables.

If responses are normally distributed, the theory and implementation of composite likelihood, particularly pairwise likelihood, are analytically derived in Chapter 2, as a result of the attribute that multivariate normal distributions are fully parameterized by mean vector and covariance matrix. An intermediate model, multivariate probit model, which allows latent multivariate normal distribution is anticipated to possess the same good properties. Whereas the responses are counts or discrete, difficulty arises caused by the lack of a discrete analogue to the multivariate normal distribution.

In the light of the multivariate normal distribution, previous related work in this area has illustrated that pairwise likelihood method performs well for a group of various models. Simulation study on the asymptotic efficiency of pairwise likelihood for multivariate probit models with exchangeable and hierarchical dependence structure is examined in detail in
Zhao and Joe (2005). Renard et al. (2004) proposed to use pairwise likelihood in multilevel models with binary responses and probit link. Simulation results show that pairwise likelihood outperforms second-order penalized quasi-likelihood (PQL2) and maximum likelihood. The loss of efficiency is generally moderate, and it tends to show more robustness against convergence problems than PQL2. Kuk and Nott (2000) applied pairwise likelihood on the estimation of association parameter. Simulation results suggest comparable asymptotic efficiency relative to alternating logistic regression (ALR) in Carey et al. (1993). Furthermore, a hybrid pairwise likelihood is proposed in Kuk (2007) which leads to better estimation of the marginal distribution parameters.

We adopt logit link function and apply pairwise likelihood to estimation of both marginal and association parameters. Simulation results show that the pairwise likelihood approach outperforms the full likelihood approach, as it provides more accurate estimates, higher efficiency, and is less computational intensive. Furthermore the pairwise likelihood is more robust to model misspecifications.
Chapter 2

Composite Test Statistics and Higher Order Asymptotics

In Sections 1 and 2, the definition and main properties of composite likelihood inference are stated. In Section 3, we present test statistics in the framework of composite likelihood. Higher order asymptotics for composite likelihood are described in Section 4, followed by an example in Section 5.

2.1 Definitions and derived quantities

Let $Y$ be a $q$-dimensional random variable with density function $f(y; \theta)$ where $\theta$ is an unknown parameter. We assume $\theta \in \Theta \subseteq \mathbb{R}^d$. The term composite likelihood here particularly refers to the composite marginal likelihood proposed in Cox and Reid (2004).
Definition The composite marginal likelihood (CML) for a random variable \( Y \) from a model \( f(y; \theta) \) is defined as

\[
CML(\theta; y) = \prod_{s \in S} f_s(y_s; \theta),
\]

where \( f_s(y_s; \theta) \) is the marginal density function corresponding to the subset \( s \), and \( S \) is a set of indices. The associated composite log-likelihood function is \( cl(\theta; y) = \log CML(\theta; y) \).

This class generalizes the usual ordinary likelihoods and contains many other interesting alternatives. We are particularly interested in lower dimensional marginal densities, often univariate and bivariate. For example, if composite likelihood is specified by the first order marginal density, i.e. \( S = \{1, \ldots, q\} \), the CML is called the independent likelihood

\[
IL(\theta; y) = \prod_{r=1}^{q} f_1(y_r; \theta),
\]

hence if we use bivariate marginal densities, then the CML is known as the pairwise likelihood

\[
PL(\theta; y) = \prod_{r<s}^{q} f_2(y_r, y_s; \theta).
\]

We shall also use \( L_2 \) to denote pairwise likelihood function.

The composite score function is defined by composite log-likelihood derivatives in the usual way

\[
U(\theta; y) = \nabla cl(\theta; y) = \sum_{s \in S} U_s(\theta; y_s)
\]

where \( U_s = \log f'_s(y_s; \theta) \). Since the composite score is a linear combination of valid likelihood
score functions, its unbiasedness follows under the regularity conditions, i.e. \( E_\theta[U(\theta; Y)] = 0 \).

Because composite likelihood is not a valid likelihood function, we lose the information identity. This is due to model misspecification: \( H(\theta) \neq J(\theta) \), unless \( U \) is the derivative of a log-density. We define \( H(\theta) \) and \( J(\theta) \) by

\[
H(\theta) = E\{-\nabla U(\theta; Y)\}, \quad (2.3)
\]
\[
J(\theta) = \text{Var}\{U(\theta; Y)\} = E\{(U(\theta; Y))U^T(\theta; Y)\}. \quad (2.4)
\]

The information in the composite score function assumes the form

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

which is known as Godambe information or sandwich information.

For a random sample \( Y = (Y^{(1)}, \ldots, Y^{(n)})^T \), the overall composite log-likelihood function is then

\[
cl(\theta; y) = \sum_{i=1}^{n} cl(\theta; y^{(i)}). \quad (2.6)
\]

The maximum composite likelihood estimator (MCLE) is defined by

\[
\tilde{\theta}_{CL} = \arg \sup_\theta cl(\theta; y). \quad (2.7)
\]
In particular, the pairwise likelihood for the random sample \( Y \) is given by

\[
\text{PL}(\theta; y) = \prod_{i=1}^{n} \prod_{r<s} f_{2}(y_r^{(i)}, y_s^{(i)}; \theta), \tag{2.8}
\]

and we shall use \( \hat{\theta}_{PL} \) for the corresponding maximum pairwise likelihood estimator.

The definition of CML given by (2.1) can be extended to have weights included. The results presented in this thesis use the formulation with no weights involved, unless otherwise specified.

## 2.2 Maximum composite likelihood estimator

In this section we state the results on the consistency and asymptotic normality of the composite likelihood estimators. The regularity conditions on the density function \( f(y; \theta) \), similarly as stated in Geys et al. (1997) and Molenberghs and Verbeke (2005), are

\begin{itemize}
  \item[(A0)] The densities \( f(y; \theta) \) are distinct for different values of the parameter \( \theta \).
  \item[(A1)] The densities \( f(y; \theta) \) have common support, so that the set \( A = \{ y : f(y, \theta) > 0 \} \) is independent of \( \theta \).
  \item[(A2)] The parameter space \( \Omega \) contains an open region \( \omega \) of which the true parameter value \( \theta_0 \) is an interior point.
  \item[(A3)] The first three derivatives of the composite log-likelihood \( cl(\theta; y) \) with respect to \( \theta \)
exist in the neighbourhood of the true parameter value almost surely. Furthermore,

\((A4)\) The first derivatives satisfy \(E_{\theta}[U(\theta; Y)] = 0\) for all \(\theta\) and therefore \(\text{Var}_{\theta}[U(\theta; Y)] = E[U(\theta; Y)U^T(\theta; Y)] = J(\theta)\).

\((A5)\) The second derivatives satisfy \(E_{\theta}[-\nabla U(\theta; Y)] = H(\theta)\). The matrices \(H(\theta)\) and \(J(\theta)\) are positive definite. Note that by definitions \(H(\theta)\) and \(J(\theta)\) are always positive-semidefinite.

\((A6)\) Let \(cl''''_{jkl}(\theta; y)\) be the mixed partial derivative of \(cl(\theta; y)\) with respect to \(\theta_j, \theta_k, \theta_l\) for \(1 \leq j, k, l \leq d\). There exist functions \(M_{jkl}\) such that

\[ |cl''''_{jkl}(\theta; y)| \leq M_{jkl}(y) \]

for all \(\theta\) in \(\omega\) and \(E_{\theta_0}(M_{jkl}(Y)) < \infty\).

The techniques used in the original proof for consistency for maximum likelihood estimators in Wald (1949) or the classical proofs for consistency and asymptotic normality in Lehmann (1983) can be applied to composite likelihoods straightforwardly. A proof for the asymptotic normality using techniques of approximation is given in Appendix A.

**Theorem 2.1 (Consistency and Asymptotic Normality)** Assume that \(Y^{(1)}, \ldots, Y^{(n)}\) are independently and identically distributed with common density that depends on \(\theta\). Under the regularity conditions \((A1) - (A7)\), the maximum composite likelihood estimator \(\tilde{\theta}_{CL}\)
converges in probability to $\theta$, and

\[ \sqrt{n}(\hat{\theta}_{CL} - \theta) \xrightarrow{d} N_d(0, G^{-1}(\theta)), \] (2.9)

where $G(\theta)$ is defined by (2.5)

Note that in general, by using Cramer-Rao inequality, the difference between the variances of MCLE and MLE, i.e. comparing the inverse of Godambe information $G^{-1}$ and the Fisher information $I^{-1}$, is positive semi-definite. The Cramer-Rao lower bound is attained ($G = I$) if and only if the estimate is a linear function of the score function $U(\theta)$ (defined in next section).

As discussed in Arnold and Strauss (1991), the inequality is strict if the composite estimator fails to be a function of a minimal sufficient statistic. This point is not necessarily true. For example, the pairwise estimator of a symmetric normal distribution can be expressed in terms of the sufficient statistics, however as we know pairwise likelihood estimation for this distribution is generally inefficient.

The large-sample properties, consistency and asymptotic normality, of composite likelihood estimators are satisfied under broad model assumptions for the scenario of many independent replicates.

For the scenario where a small number $n$ of individually large sequences is available, there are no more ensured satisfactory properties of the resulting estimator, composite or ordinary maximum. When $n$ is relatively small compared to $q$, consistency of the estimator depends
on the internal correlation structure. If there is too much internal correlation present, $\tilde{\theta}$ will not be a consistent estimator of $\theta$, nor $\hat{\theta}$, as $q$ increases. If the correlation is that of short-range dependent stationary time series, convergence of the overall sample mean to $\theta$ will be at the usual rate, that is $1/\sqrt{q}$; whereas if the correlation is that of a long-range dependent process, convergence will be slower or fail. See discussions in Cox and Reid (2004).

2.3 Composite Likelihood Test Statistics

2.3.1 Composite likelihood test statistics, with no nuisance parameter

In principle, it is straightforward to extend the standard test statistics under the framework of ordinary likelihoods to the framework of composite likelihoods. We first start with treating the parameter $\theta \in \mathbb{R}^d$ as a whole.

For the convenience in notations in the derivation of this section and so on, we shall in general use $\sim$ to indicate composite likelihood objects. Let $\bar{l}(\theta) = \sum \bar{l}(\theta; Y^{(i)})$ be the total composite log-likelihood function for $n$ observations. The score function is $U(\theta) = \nabla \bar{l}(\theta; Y^{(i)})$, and we let $U(\theta) = \sum U(\theta)$ denote the total score statistic.

It is convenient to consider the estimated matrix functions

\[
\hat{H}(\theta) = -n^{-1} \sum \partial U(\theta)/\partial \theta^T, \quad \hat{J}(\theta) = n^{-1} \sum U(\theta)U^T(\theta).
\]
In general, they are consistent estimates of \( H(\theta) \) and \( J(\theta) \), respectively.

From the results stated in Section 2.2,

\[
\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N_d(0, G^{-1}(\theta)),
\]

and from the Central Limit Theorem, it follows that

\[
\frac{1}{\sqrt{n}} U.(\theta) \xrightarrow{d} Z \sim N(0, J(\theta))
\]

and so we have

\[
n(\tilde{\theta} - \theta)^T G(\tilde{\theta} - \theta) \xrightarrow{d} \chi^2_d, \quad n^{-1} U.(\theta)^T J^{-1}(\theta) U.(\theta) \xrightarrow{d} \chi^2_d
\]

where \( \chi^2_d \) is the Chi-squared distribution with \( d \) degrees of freedom, and \( d \) is the dimension of \( \theta \). We can define composite Wald and score tests as \( n(\tilde{\theta} - \theta)^T G(\tilde{\theta} - \theta) \) and \( n^{-1} U.(\theta)^T J^{-1}(\theta) U.(\theta) \), respectively.

The composite likelihood ratio test can be defined and also model-based composite Wald and score statistics, although their asymptotic distribution is a bit more complicated, as

**Theorem 2.2 (Composite likelihood ratio test)** Under the model assumed, asymptotically as \( n \to \infty \), the composite likelihood ratio statistic \( \tilde{w}(\theta) \) is distributed as

\[
\tilde{w}(\theta) = 2\{\tilde{l}(\tilde{\theta}) - \tilde{l}(\theta)\} \xrightarrow{d} \sum_{j=1}^d \nu_j V_j,
\]
where \( V_1, \ldots, V_d \) denote independent \( \chi^2_1 \) variates, and \( \nu_j \) are the eigenvalues of

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

(2.11)

and \( J \) and \( H \) are defined by (2.3) and (2.4), respectively. Further, \( \tilde{W} \) is asymptotically equivalent to both the composite Wald statistic

\[
\tilde{w}_c(\theta) = n(\tilde{\theta} - \theta)^T H(\theta)(\tilde{\theta} - \theta)
\]

(2.12)

and to the composite score statistic

\[
\tilde{w}_s(\theta) = n^{-1}U(\theta)H^{-1}(\theta)U(\theta)
\]

(2.13)

The proof is completed in Appendix B.

### 2.3.2 Composite test statistics for \( \theta = (\psi, \lambda) \)

Now we assume the parameter \( \theta \in \mathbb{R}^d \) is partitioned as \( \theta = (\psi, \lambda) \), where \( \psi \) is a \( p \times 1 \) parameter of interest and \( \lambda \) is a \( q \times 1 \) nuisance parameter, \( d = p + q \). For convenience we shall largely follow the notations of Cox and Hinkley (1974, Chapter 9) and Kent (1982).

We wish to test the null hypothesis that \( \psi = \psi_0 \), for some specified value of \( \psi_0 \). Let \( \tilde{\theta} = (\tilde{\psi}, \tilde{\lambda}) \) denote the unrestricted maximum composite likelihood estimate and let \( \tilde{\theta}_0 = (\psi_0, \lambda_0) \) denote the maximum composite likelihood estimate under the null hypothesis.
We partition the total score statistic $U.(\theta)$ into

$$ U.(\theta) = \begin{bmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{bmatrix}, $$

and the maximum composite likelihood estimates $\tilde{\theta}$ and $\tilde{\theta}_0$ satisfy

$$ U.(\tilde{\theta}) = 0, \quad U_\lambda(\psi_0, \tilde{\lambda}_0) = 0. $$

Furthermore, we write the partition of matrix $H$ and its inverse as

$$ H = \begin{bmatrix} H_\psi,\psi & H_\psi,\lambda \\ H_\lambda,\psi & H_\lambda,\lambda \end{bmatrix}, \quad H^{-1} = \begin{bmatrix} H^{\psi,\psi} & H^{\psi,\lambda} \\ H^{\lambda,\psi} & H^{\lambda,\lambda} \end{bmatrix}. $$

and use the notation $H_{\psi,\lambda} = (H^{\psi,\psi})^{-1} = H^{\psi,\psi} - H^{\psi,\lambda} H^{\lambda,\lambda}^{-1} H^{\lambda,\psi}$.

**Theorem 2.3 (Composite test statistics with nuisance parameter)** Under the model assumed, in testing the null hypothesis $H_0$: $\psi = \psi_0$, asymptotically as $n \to \infty$, the composite likelihood ratio statistic $\bar{w}(\psi_0)$ is distributed as

$$ \bar{w}(\psi_0) = 2\{\bar{l}(\tilde{\theta}) - \bar{l}(\tilde{\theta}_0)\} \overset{d}{\to} \sum_{j=1}^p \nu_j V_j $$

(2.14)

where $V_1, \ldots, V_p$ denote independent $\chi^2_1$ variates, and $\nu_j$ are the eigenvalues of

$$ (H^{\psi,\psi})^{-1} G^{\psi,\psi}, $$

(2.15)
and \( J \) and \( H \) are defined by (2.3) and (2.4), respectively. Further, \( \tilde{W} \) is asymptotically equivalent to both the composite Wald statistic

\[
\tilde{w}_e(\psi_0) = n(\tilde{\psi} - \psi_0)^T H_{\psi\psi,\lambda}(\tilde{\psi} - \psi_0)
\]  

(2.16)

and to the composite score statistic

\[
\tilde{w}_u(\psi_0) = n^{-1} U_{\psi}(\tilde{\theta}_0) H_{\psi\psi,\lambda}^{-1} U_{\psi}(\tilde{\theta}_0)
\]  

(2.17)

Proof. Composite likelihood is a special case of model misspecification. The proof follows Kent (1982); see Appendix B.

Remark 1. If (2.15) reduces to the identity matrix, then \( \tilde{w} \overset{d}{\to} \chi^2_p \).

Remark 2. If \( H(\theta) \) is block diagonal, that is, \( H_{\psi\lambda} = 0 \) evaluated under \( H_0 \), then (2.15) simplifies to \( J_{\psi\psi} H_{\psi\psi}^{-1} \). In particular, \( J_{\psi\psi} H_{\psi\psi}^{-1} \) is appropriate when \( q = 0 \), that is when no nuisance parameters are present.

Remark 3. If \( H_{\psi\lambda} = 0 \) and \( p = 1 \) then (2.15) simplifies further to the ratio of numbers \( J_{\psi\psi}/H_{\psi\psi} \), which acts as a correction factor to the usual \( \chi^2_1 \) distribution.
2.4 Higher accuracy inference in terms of $\tilde{r}$ under simple null hypothesis

In the following part of this chapter, we consider the problem of constructing higher-order asymptotic approximations to the distribution of the maximum composite likelihood estimator, the composite likelihood ratio test statistic and related quantities.

When we implement composite likelihoods as the likelihood objects of interest, we meanwhile bring in the difficulty caused by the failure of Bartlett identities. When the identities, in particular Bartlett second identity, fail, which equates the variance of the score and the expected Fisher information, no simplification in expressions for asymptotic variance is then allowed.

The section investigates the use of mean and variance corrections to construct an approximate pivot based on the signed composite likelihood root statistic $\tilde{r}$ that has the standard normal distribution to error of order $O(n^{-3/2})$. More specifically, the mean and variance of the signed composite likelihood root are expected to be expanded in the form of

\[ E(\tilde{r}) = m(\theta) + O(n^{-3/2}), \quad \text{Var}(\tilde{r}) = c(\theta) + v(\theta) + O(n^{-2}) \]

where $m(\theta)$ is of order $O(n^{-1/2})$ and $v(\theta)$ is of order $O(n^{-1})$, and $c(\theta)$ is of order $O(1)$ and determined by the ratio of the the variance of the score and the expected score derivative, i.e. $JH^{-1}$. Furthermore, the third- and higher-order cumulants of $\tilde{r}$ are of order $O(n^{-3/2})$. 

or smaller. Then by assuming the validity of an Edgeworth expansion for \( \tilde{r} \), it follows that the standard normal approximation to the distribution of

\[
\frac{\tilde{r} - E(\tilde{r})}{\{\text{Var}(\tilde{r})\}^{1/2}}
\]

has an error of order \( O(n^{-3/2}) \).

### 2.4.1 Maximum composite likelihood estimation

Given a statistical model, for notation simplicity, we write the overall composite log-likelihood function \( \tilde{l} = \tilde{l}(\theta) \) for a random sample of size \( n \).

To facilitate the discussion and calculations that follow, we will employ standard tensor notations used in McCullagh (1987) and Barndorff-Nielsen and Cox (1994). Generic coordinates of the parameter \( \theta \), assumed \( d \)-dimensional, will be denoted by \( \theta^r, \theta^s, \ldots \). Differentiation will be denoted by subscripts, so that \( \tilde{l}_r = \partial \tilde{l}(\theta)/\partial \theta^r \), \( \tilde{l}_{rs} = \partial^2 \tilde{l}(\theta)/\partial \theta^r \partial \theta^s \), and so on. We define \( j \) as minus the matrix of second order derivatives of \( \tilde{l} \), i.e.

\[
j_{rs} = -\tilde{l}_{rs},
\]

and \( j^{rs} \) as the observed information, i.e. the inverse matrix of the observed information matrix \( j_{rs} \).

We shall use \( \nu \) to indicate joint moments of the composite log-likelihood derivatives. More
specifically, if $R = r_1 \cdots r_m$ are a set of coordinate indices we let

$$\nu_R = E(\tilde{l}_R)$$

with appropriately partitioned subscripts, so that $\nu_{rs} = E(\tilde{l}_{rs}), \nu_{r,s} = E(\tilde{l}_{r}\tilde{l}_{s}), \nu_{r,st} = E(\tilde{l}_{r}\tilde{l}_{st}),$ and so on.

Furthermore, we will use $u_R = -\nu_R$. Finally, we introduce the general notation

$$H_R = \tilde{l}_R - \nu_R,$$

which is the deviance between the composite log-likelihood and its expectation.

For the sake of brevity in the notation, we shall henceforth generally suppress the dependence on $\theta$ for the derivatives of $\tilde{l}$, at least under clear circumstances. Besides, while $\tilde{l}$ is the composite log-likelihood function we adhere, however, to our remaining notations such as $j, \nu$, etc. by dropping $\sim$ symbol, we write, for instance, $j$ instead of $\tilde{j}$, and so on.

Taking Taylor expansion of $\tilde{l}_r(\tilde{\theta}) = 0$ around $\theta$ gives

$$0 = \tilde{l}_r(\tilde{\theta}) = \tilde{l}_r + \tilde{l}_{rs}(\tilde{\theta} - \theta)^s + \frac{1}{2} \tilde{l}_{rst}(\tilde{\theta} - \theta)^s(\tilde{\theta} - \theta)^t + \frac{1}{6} \tilde{l}_{rstu}(\tilde{\theta} - \theta)^s(\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)^u + O_p(n^{-1}).$$
With using notations \( j_{rs} \) and \( j^{rs} \), we can have an asymptotic expansion for \((\tilde{\theta} - \theta)^r = \tilde{\theta}^r - \theta^r\)

\[
(\tilde{\theta} - \theta)^r = j^{rs} \tilde{l}_s + \frac{1}{2} j^{rs} \tilde{l}_{stu}(\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)^u \\
+ \frac{1}{6} j^{rs} \tilde{l}_{stuv}(\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)^u(\tilde{\theta} - \theta)^w + O_p(n^{-2}) \\
= j^{rs} \tilde{l}_s + \frac{1}{2} j^{rs} j^{tu} j^{vw} \tilde{l}_{stu} \tilde{l}_u \tilde{l}_w \\
+ \frac{1}{6} j^{rs} j^{tu} j^{vw} j^{xy} (\tilde{l}_{suwy} + 3 \tilde{l}_{sup} j^{pq} \tilde{l}_{qwy}) \tilde{l}_t \tilde{l}_u \tilde{l}_v + O_p(n^{-2}).
\] (2.18)

Furthermore, an expansion of \( j^{-1} \) with asymptotic order of \( O_p(n^{-5/2}) \) is desired. To avoid the Bartlett identities, rewrite

\[
j_{rs} = u_{rs} \{I - u^{rs}(u_{rs} - j_{rs})\}, \quad \text{where} \quad u_{rs} = -E(\tilde{l}_{rs}) = E(j_{rs}) \\
\hat{j}^{rs} = \{I - u^{rs}(u_{rs} - j_{rs})\}^{-1} u^{rs} \\
= u^{rs} + u^{rt} u^{su}(u_{tu} - j_{tu}) + u^{rt} u^{su} u^{vw}(u_{tv} - j_{tv})(u_{uw} - j_{uw}) + O_p(n^{-5/2}).
\]

Then by inserting the notation \( H \), we have

\[
j^{rs} = u^{rs} + u^{rt} u^{su} H_{tu} + u^{rt} u^{su} u^{vw} H_{tv} H_{uw} + O_p(n^{-5/2}).
\] (2.19)

Therefore, an expansion for \( \tilde{\theta} - \theta \) with asymptotic magnitude of \( O(n^{-2}) \) can be derived by
plugging (2.19) in (2.18) and collecting terms of the same asymptotic order. Finally,

\[
(\tilde{\theta} - \theta)^r = u^{rs} \tilde{l}_s + u^{rs} u^{tu} H_{st} \tilde{l}_u + \frac{1}{2} u^{rs} u^{tu} u^{vw} \tilde{l}_{stu} \tilde{l}_v \tilde{l}_w \\
+ u^{rs} u^{tu} u^{vw} H_{st} H_{uw} \tilde{l}_w + \frac{1}{2} u^{rs} u^{tu} u^{vw} u^{xy} H_{st} \tilde{l}_{uvw} \tilde{l}_x \tilde{l}_y \\
+ u^{rs} u^{tu} u^{vw} u^{xy} H_{vw} \tilde{l}_{stu} \tilde{l}_x \tilde{l}_y \\
+ \frac{1}{6} u^{rs} u^{tu} u^{vw} u^{xy} (\nu_{stu} + 3 \nu_{stp} u^{pq} \nu_{quv}) \tilde{l}_{wz} \tilde{l}_y + O_p(n^{-2})
\]  

(2.20)

\[
(\tilde{\theta} - \theta)^r = u^{rs} \tilde{l}_s + u^{rs} u^{tu} H_{st} \tilde{l}_u + \frac{1}{2} u^{rs} u^{tu} u^{vw} \nu_{stu} \tilde{l}_w \\
+ u^{rs} u^{tu} u^{vw} H_{st} H_{uw} \tilde{l}_w + \frac{1}{2} u^{rs} u^{tu} u^{vw} u^{xy} H_{st} \nu_{uvw} \tilde{l}_x \tilde{l}_y \\
+ u^{rs} u^{tu} u^{vw} u^{xy} H_{vw} \nu_{stu} \tilde{l}_x \tilde{l}_y \\
+ \frac{1}{6} u^{rs} u^{tu} u^{vw} u^{xy} (\nu_{stu} + 3 \nu_{stp} u^{pq} \nu_{quv}) \tilde{l}_{wz} \tilde{l}_y + O_p(n^{-2}).
\]  

(2.21)

An expansion for the mean of the maximum composite likelihood estimator can be obtained readily by taking termwise expectations in (2.21). Note that under the regularity conditions the mean of \( \tilde{l}_r \) is zero for all \( \theta \), that is

\[
\nu_r = E(\tilde{l}_r) = 0.
\]

The first correction to the leading term \( \tilde{\theta}^r \) is of order \( n^{-1} \) and is given by

\[
E(\tilde{\theta} - \theta)^r = u^{rs} u^{tu} (\nu_{stu} + \frac{1}{2} \nu_{stu} u^{vw} \nu_{v,w}) + O(n^{-2}).
\]  

(2.22)

Note that with the second Bartlett identity, i.e. \( \nu_{r,s} = u_{rs} \) the equation (2.22) deduces to

\[
E(\tilde{\theta} - \theta)^r = \frac{1}{2} u^{rs} u^{tu} (\nu_{stu} + 2 \nu_{st,u}) + O(n^{-2}).
\]
which is consistent with the result for the MLE in Barndorff-Nielsen and Cox (1994, p.150) or McCullagh (1987, p.209).

Using expansions (2.21) and (2.22), the asymptotic normality of the maximum composite likelihood estimator can be proved, with error of order $O(n^{-1/2})$: see Appendix A.

It is straightforward to simplify these results to the univariate case. To make the notations more specific for single parameter, we write

\[
\begin{align*}
\nu_{11} &= E \left( \frac{\partial^3 \tilde{l}}{\partial \theta^3} \right) \equiv \nu_3, \\
\nu_{11,1} &= E \left( \frac{\partial^2 \tilde{l}}{\partial \theta^2} \frac{\partial \tilde{l}}{\partial \theta} \right), \\
\nu_{1,1} &= E \left( \frac{\partial \tilde{l}}{\partial \theta} \frac{\partial \tilde{l}}{\partial \theta} \right), \\
\nu_{1} &= E \left( \frac{\partial^2 \tilde{l}}{\partial \theta^2} \right) \equiv \nu_2 = -u_{11}
\end{align*}
\]

Therefore, in the univariate case we have

\[
\begin{align*}
E(\tilde{\theta} - \theta) &= (u_{11}^{-1})^2(\nu_{11,1} + \frac{1}{2}\nu_3 u_{11}^{-1} \nu_{1,1}) + O(n^{-3/2}), \quad (2.23) \\
Var(\tilde{\theta} - \theta) &= (u_{11}^{-1})^2 \nu_{1,1} + O(n^{-1}). \quad (2.24)
\end{align*}
\]

With ordinary likelihoods, we have $\nu_{1,1} = -\nu_{11}$, so $Var(\hat{\theta}) = 1/\nu_{1,1} + O(n^{-1})$ for the MLE $\hat{\theta}$ where $\nu_{1,1} = E(l_\theta^2)$ is the expected Fisher information, as expected.
2.4.2 Composite likelihood ratio statistic

In this subsection we work with the composite likelihood ratio statistic defined by (2.10)

\[ \tilde{w}(\theta) = 2\{\tilde{l}(\tilde{\theta}) - \tilde{l}(\theta)\} \]

where \( \tilde{l}(\theta) \) is the composite log-likelihood function for the sample of \( n \) independent observations. We find by Taylor expansion around \( \theta \) that

\[
\tilde{l}(\tilde{\theta}) - \tilde{l}(\theta) = \tilde{l}_r(\tilde{\theta} - \theta)^r + \frac{1}{2} \tilde{l}_{rs}(\tilde{\theta} - \theta)^r(\tilde{\theta} - \theta)^s + \frac{1}{6} \tilde{l}_{rst}(\tilde{\theta} - \theta)^r(\tilde{\theta} - \theta)^s(\tilde{\theta} - \theta)^t \\
+ \frac{1}{24} \tilde{l}_{rstu}(\tilde{\theta} - \theta)^r(\tilde{\theta} - \theta)^s(\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)^u + O_p(n^{-3/2}).
\]

We insert the expansion for \( (\tilde{\theta} - \theta)^r \) of (2.21) and rearrange terms according to their asymptotic order. Simplification renders an expansion for \( \tilde{l}(\tilde{\theta}) - \tilde{l}(\theta) \) with asymptotic order of \( O_p(n^{-3/2}) \) as follows

\[
\tilde{l}(\tilde{\theta}) - \tilde{l}(\theta) = \frac{1}{2} u^{rs} \tilde{l}_r \tilde{l}_s + \frac{1}{2} u^{rs} u^{tu} H_{rt} \tilde{l}_s \tilde{l}_u + \frac{1}{6} u^{rs} u^{tu} u^{vw} \nu_{rtu} \tilde{l}_s \tilde{l}_u \tilde{l}_w \\
+ \frac{1}{2} u^{rs} u^{tu} u^{vw} H_{rt} H_{sv} \tilde{l}_u \tilde{l}_w + \frac{1}{8} u^{rs} u^{tu} u^{vw} u^{xy} u^{z_p} \nu_{rtu} \nu_{szw} \tilde{l}_u \tilde{l}_w \tilde{l}_y \tilde{l}_p + \frac{1}{2} u^{rs} u^{tu} u^{vw} u^{xy} \nu_{rtu} H_{sx} \tilde{l}_u \tilde{l}_w \tilde{l}_y \\
+ \frac{1}{6} u^{rs} u^{tu} u^{vw} H_{rtu} \tilde{l}_s \tilde{l}_u \tilde{l}_w + \frac{1}{24} u^{rs} u^{tu} u^{vw} u^{xy} \nu_{rtu} \nu_{rtw} \tilde{l}_s \tilde{l}_u \tilde{l}_w \tilde{l}_y + O_p(n^{-3/2}).
\]

With Bartlett identities, (2.25) is consistent with the expansion given in Barndorff-Nielsen and Cox (1994) for the log-likelihood difference.
An immediate result for $\tilde{w}$ is obtained

$$
\tilde{w} = u^{rs}l_r l_s + u^{rs}u^{tu}H_{rt}l_l l_u + \frac{1}{3} u^{rs}u^{tu}u^{vw}v_{rtv}l_s l_u l_w
$$

$$
+ u^{rs}u^{tu}u^{vw}H_{rt}H_{sv}l_u l_w + \frac{1}{4} u^{rs}u^{tu}u^{xy}u^{zp}v_{rtv}v_{sxz}l_u l_w l_y l_p + u^{rs}u^{tu}u^{vw}u^{xy}v_{rtv}l_{sz} l_u l_w + \frac{1}{12} u^{rs}u^{tu}u^{vw}u^{xy}v_{rtv}v_{sxz}l_u l_w l_y l_p + O_p(n^{-3/2}).
$$

(2.26)

Note that the leading term in the above expansion for the composite likelihood ratio statistic $\tilde{w}$ is $u^{rs}l_r l_s$, which is the composite score statistic. Therefore, we have

$$
\tilde{w} = \tilde{w}_u + O_p(n^{-1/2}),
$$

which is consistent with the conclusions we made in the previous section that composite test statistics are asymptotically equivalent, and moreover the error term in the approximation appears to be $O_p(n^{-1/2})$.

An expansion for the more general composite log likelihood ratio with nuisance parameter involved

$$
\tilde{w}(\psi_0) = 2\{\tilde{l}(\tilde{\theta}) - \tilde{l}(\tilde{\theta}_0)\}
$$

where $\tilde{\theta} = (\tilde{\psi}, \tilde{\lambda})$ and $\tilde{\theta}_0 = (\psi_0, \lambda_0)$ is readily obtained by writing $\tilde{l}(\tilde{\theta}) - \tilde{l}(\tilde{\theta}_0)$ as the difference between $\tilde{l}(\tilde{\theta}) - l(\tilde{\theta})$ and $\tilde{l}(\tilde{\theta}_0) - l(\tilde{\theta})$ and applying (2.25) to both these terms.
2.4.3 Signed composite likelihood root statistic and its approximation

We go on to consider higher-order expansion for the signed composite likelihood root statistic such as

\[
\tilde{r}(\psi_0) = \text{sign}(\tilde{\psi} - \psi_0)\sqrt{\tilde{w}(\psi_0)}.
\]

To avoid further derivation of the expansion for the composite likelihood ratio statistic with nuisance parameter, we concentrate on the signed composite likelihood root statistic

\[
\tilde{r}(\theta) = \text{sign}(\tilde{\theta} - \theta)\sqrt{\tilde{w}(\theta)}.
\]

Specifically, when \( \theta \) is a single parameter or it contains a scalar parameter of interest with a nuisance parameter, under the simple null hypothesis, \( \tilde{r} \) is asymptotically normally distributed. The error in the normal approximation is typically of order \( O(n^{-1/2}) \). We aim to show that besides the above fact, a small adjustment by means of mean and variance corrections can make it much closer to normally distributed with the error of order \( O(n^{-3/2}) \) and hence improve the accuracy of this approximation.

To be more specific in terms of tensor notation, rather than applying the Taylor’s series expansion for the square-root function to attain an expansion for the signed square root of \( \tilde{r}^2(= \tilde{w}) \), we define the vector similar as in McCullagh (1987, Section 7.4) as part of the
decomposition of the composite likelihood ratio statistic

\[
\tilde{r}_t = \tilde{l}_t + \frac{1}{6} u^{rs} (3 H_{rt} \tilde{l}_s + u^{tw} \nu_{rtv} \tilde{l}_s \tilde{l}_w) \\
+ \frac{1}{72} u^{rs} u^{vw} (27 H_{rt} H_{svw} \tilde{l}_w + 8 u^{xy} u^{wp} \nu_{rtu} \nu_{sxz} \tilde{l}_w \tilde{l}_y \tilde{l}_p + 30 u^{xy} \nu_{rtv} H_{sxz} \tilde{l}_w \tilde{l}_y \\
+ 12 H_{rtuv} \tilde{l}_s \tilde{l}_w + 3 u^{xyz} \nu_{rtuv} \tilde{l}_s \tilde{l}_w \tilde{l}_y) + O_p(n^{-1})
\] (2.28)

such that

\[
\tilde{r}^2 = \tilde{r}_t \tilde{r}_u + O_p(n^{-3/2}).
\] (2.29)

Note that \( \tilde{r}_t \) is a tensor. Hereafter we attempt to derive the expressions for the cumulants of \( \tilde{r}_t \) and so those of \( \tilde{r} \).

From (2.28) it follows that the expectation of \( \tilde{r}_t \) with error \( O(n^{-1}) \) is given by

\[
E(\tilde{r}_t) = \frac{1}{6} u^{rs} \{ 3 \nu_{rt,s} + u^{vw} \nu_{rtv} \nu_{s,w} \} + O(n^{-1}).
\] (2.30)

Notice that if the second Bartlett identity holds, then equation (2.30) deduces to

\[
E(\tilde{r}_t) = \frac{1}{6} u^{rs} \{ 3 \nu_{rt,s} + \nu_{rst} \} + O(n^{-1}).
\]

With the further third order identity, \( \nu_{r,s,t} + 3 \nu_{rs,t} + \nu_{rst} = 0 \), a consistent result as given in Barndorff-Nielsen and Cox (1994, p.154) or (McCullagh, 1987, p.214) is rendered, i.e.

\[
E(\tilde{r}_t) = -\frac{1}{6} u^{rs} \nu_{r,s,t} + O(n^{-1}).
\]
2.4. HIGHER ACCURACY INFERENCE IN TERMS OF $\tilde{R}$ UNDER SIMPLE NULL HYPOTHESIS

We continue to further derive the joint cumulants of $\tilde{r}_t$. Returning to (2.26), we find the mean value of the composite likelihood ratio statistic $\tilde{r}^2$ has an asymptotic expansion of the form

$$E(\tilde{r}^2) = u^r s \nu_{r,s} + \tilde{R} + O(n^{-2}) \quad (2.31)$$

where $\tilde{R}$, which can be interpreted as the expected Bartlett adjustment with no Bartlett identities, is of order $n^{-1}$ and is given by

$$\tilde{R} = \frac{1}{12} \{ \tilde{K}_{rstu} u^r s u^t u^w + \tilde{K}_{rstuvw} u^r s u^t u^w \} \quad (2.32)$$

with

$$\tilde{K}_{rstu} = 3 \nu_{rstu} u^w u^x y \nu_{v,w} \nu_{x,y} + 12(\nu_{rt,su} u^w v_{v,w} + \nu_{r,t,su} + \nu_{r,stu} u^w \nu_{v,w}), \quad (2.33)$$

$$\tilde{K}_{rstuvw} = (3 \nu_{rst} u^v w + 6 \nu_{rtv} u^w v_{v,w} \nu_{x,y} \nu_{z,p} + 12 \{ \nu_{r,stv} u^u \nu_{w,sw} + \nu_{r,tv} \nu_{su,w} \})$$

$$+ 12 \nu_{rstw} u^x y \nu_{v,w} + 24 \nu_{rtv} u^w v_{v,w} \nu_{x,y} + 4 \nu_{r,t,v} \nu_{su}. \quad (2.34)$$

Using these expansions we are able to obtain the expression for Cov($\tilde{r}_t, \tilde{r}_s$). Recall that $\tilde{r}^2 = \tilde{r}_t \tilde{r}_u u^t u + O_p(n^{-3/2})$. It follows that

$$\text{Cov}(\tilde{r}_t, \tilde{r}_s) = E(\tilde{r}_t \tilde{r}_s) - E(\tilde{r}_t) E(\tilde{r}_s) = E(\tilde{r}^2) u_{t,s} - E(\tilde{r}_t) E(\tilde{r}_s)$$

$$= \nu_{t,s} + \frac{1}{12} \{ \tilde{K}_{rstu} u^u + \tilde{K}_{rstuvw} u^u u^v w \} - E(\tilde{r}_t) E(\tilde{r}_s) + O(n^{-1}).$$
Inserting the result of (2.30), at the end we have

\[
\text{Cov}(\tilde{r}_t, \tilde{r}_s) = \nu_{t,s} + \frac{1}{12} \{ \tilde{K}_{rstu} u^{ru} + (\tilde{K}_{rstuw} - 3\nu_{r,st}\nu_{uv,w} - 2\nu_{r,st}\nu_{uvu} u^{xy} \nu_{x,y} \\
- \frac{1}{3}\nu_{rst}\nu_{uvu} u^{xy} u^{zp} \nu_{x,y} \nu_{z,p}) u^{ru} u^{vw} \} + O(n^{-1})
\] (2.35)

with \(\tilde{K}_{rstu}\) and \(\tilde{K}_{rstuw}\) given by (2.33) and (2.34), respectively.

Loosely speaking, if we define the signed composite likelihood ratio statistic \(\tilde{r}\) as \(\tilde{r} = \tilde{r}_t \tilde{i}_t^{-1/2}\), where \(\tilde{i}_t^{-1/2}\) denotes an arbitrary matrix square root of the expected formation \(u^{rs}\), then

\[
\begin{align*}
\mathbb{E}(\tilde{r}) & = \frac{1}{6} u^{rs}/\sqrt{\tilde{i}_t} \{3v_{rt,s} + u^{vw} v_{rt,v,sw} \} + O(n^{-3/2}) = m(\theta) + O(n^{-3/2}) \\
\text{Var}(\tilde{r}) & = \text{Cov}(\tilde{r}_t, \tilde{r}_s) u^{ts} + \frac{1}{12} \{ \tilde{K}_{rstu} u^{ts} u^{tu} + \tilde{K}_{rstuw} u^{rs} u^{tu} u^{vw} \\
& - \frac{1}{12} \{3\nu_{r,st}\nu_{uv,w} + 2\nu_{r,st}\nu_{uvu} u^{xy} \nu_{x,y} + \frac{1}{3}\nu_{rst}\nu_{uvu} u^{xy} u^{zp} \nu_{x,y} \nu_{z,p} \} u^{rs} u^{tu} u^{vw} + O(n^{-2}) \\
& = u^{ts} \nu_{t,s} + v(\theta) + O(n^{-2})
\end{align*}
\]

where \(m(\theta)\) and \(v(\theta)\) are of order \(O(n^{-1/2})\) and \(O(n^{-1})\) respectively as desired. The first term \(u^{ts} \nu_{t,s}\) in \(\text{Var}(\tilde{r})\) indicates that it is not standard normal, as \(\tilde{r}^2\) has a scaled non-central Chi-square distribution with non-centrality parameter, in the simplest case, given by the eigenvalues of \(JH^{-1}\).

Since in our derivations we did not specify the form for the composite log-likelihood function \(\tilde{l}, \tilde{l}\) can be understood as a generic class of likelihood functions which includes the usual
likelihoods as well as many other interesting alternatives. Thus the results derived here for the signed composite likelihood ratio statistic contain and further generalize the ordinary signed likelihood root statistic $r$ ((1.13), Chapter 1).

For the scalar parameter case, $d = 1$, we will give out the rigorous expressions of the expectation and variance together with the explicit formulae for $m(\theta)$ and $v(\theta)$ for the mean and variance corrections.

Notations are similar but simpler for the single parameter case. Let $\tilde{l}_\theta$, $\tilde{l}_{\theta\theta}$, and so on, denote the derivatives of $\tilde{l}$ with respect to $\theta$. Again use $\nu_{111} = E(\tilde{l}_{\theta\theta\theta}) \equiv v_3$ and $\nu_1 = E(\tilde{l}_{\theta\theta\theta\theta})$. In particular, $\nu_2 = E(\tilde{l}_{\theta\theta}) = -u_{11}$ and $H_2 = \tilde{l}_{\theta\theta} - \nu_2$ and so on. Then the composite likelihood ratio statistic can be written as

$$\tilde{r}^2 = \frac{1}{u_{11}} \tilde{l}_\theta^2 + \frac{1}{3} \{(u_{11})^3 v_3 \tilde{l}_\theta^3 + 3(u_{11})^2 H_2 \tilde{l}_\theta^2 \}
+ \frac{1}{12} \{(u_{11})^4 \nu_4 + 3(u_{11})^5 \nu_3^2 \tilde{l}_\theta^4 + (u_{11})^3 H_2^2 \tilde{l}_\theta^2 + (u_{11})^4 v_3 H_2 \tilde{l}_\theta^2 + \frac{1}{3} (u_{11})^3 H_3 \tilde{l}_\theta^3 \}
+ O_p(n^{-3/2}). \quad (2.36)$$

Returning to (2.31), the expectation of $\tilde{r}^2$ is given by

$$E(\tilde{r}^2) = u_{11}^{-1} \nu_{1,1} + \tilde{R} + O(n^{-2}) \quad (2.37)$$
where $\tilde{R}$ is of order $n^{-1}$ and is given by

$$\tilde{R} = \frac{1}{12} \{\tilde{K}_4(u_{11}^{-1})^2 + \tilde{K}_6(u_{11}^{-1})^3\} \quad (2.38)$$

with

$$\tilde{K}_4 = 3\nu_4(u_{11}^{-1})^2\nu_{1,1}^2 + 12(\nu_{2,2}u_{11}^{-1}\nu_{1,1} + \nu_{1,11}u_{11}^{-1}\nu_{1,1}), \quad (2.39)$$

$$\tilde{K}_6 = 9\nu_3^2(u_{11}^{-1})^2\nu_{1,1}^2 + 24\nu_{1,11}^2 + 36\nu_{1,11}\nu_{3}u_{11}^{-1}\nu_{1,1} + 4\nu_{1,11}\nu_{3}. \quad (2.40)$$

Furthermore, the signed composite likelihood ratio statistic has a form of

$$\tilde{r} = u_{11}^{-1}\bar{l}_{\theta} \left[ 1 + \frac{1}{6}u_{11}^{-1}(3H_2 + u_{11}^{-1}\nu_3\bar{l}_{\theta}) \right. \right.
\left. \left. + \frac{1}{72}(u_{11}^{-1})^2 \{ 27H_2^2 + 30u_{11}^{-1}\nu_3H_2\bar{l}_{\theta} + 12H_3\bar{l}_{\theta} + 3u_{11}^{-1}\nu_4\bar{l}_{\theta}^2 + 8(u_{11}^{-1})^2\nu_3^2\bar{l}_{\theta}^2 \} \right] + O_p(n^{-3/2}). \quad (2.41)$$

Hence, the first and second cumulants of $\tilde{r}$ can be obtained as

$$\kappa_1(\tilde{r}) = \frac{1}{6}u_{11}^{-3/2}(3\nu_{11,1} + \nu_3u_{11}^{-1}\nu_{1,1}) + O(n^{-3/2}), \quad (2.42)$$

$$\kappa_2(\tilde{r}) = \kappa_1(\tilde{r})^2 - \kappa_1^2(\tilde{r})$$

$$= u_{11}^{-1}\nu_{1,1} + \tilde{R} - \frac{1}{36}u_{11}^{-3}(9\nu_{11,1}^2 + 6\nu_3\nu_{11,1}u_{11}^{-1}\nu_{1,1} + \nu_3^2(u_{11}^{-1})^2\nu_{1,1}^2) + O(n^{-2}). \quad (2.43)$$
If we write

\begin{align*}
m(\theta) &= \frac{1}{6} u_{11}^{-3/2} (3 \nu_{11,1} + u_{11}^{-1} \nu_{3,1,1}) \\
v(\theta) &= \tilde{R} - \frac{1}{36} u_{11}^{-3} (9 \nu_{11,1}^2 + 6 \nu_3 \nu_{11,1} u_{11}^{-1} \nu_{1,1} + \nu_3^2 (u_{11}^{-1})^2 \nu_{1,1}^{2})
\end{align*}

(2.44)  (2.45)

where \( m(\theta) \) and \( v(\theta) \) are of orders \( n^{-1/2} \) and \( n^{-1} \), respectively, then

\begin{align*}
E(\tilde{r}) &= m(\theta) + O(n^{-3/2}), \\
Var(\tilde{r}) &= u_{11}^{-1} \nu_{1,1} + v(\theta) + O(n^{-2})
\end{align*}

(2.46)

as desired, except that the first term in variance is not 1, which can be easily modified by scaling \( \tilde{r} \).

Another key step in deriving higher order inference based on \( \tilde{r} \) is to verify whether its third- and higher-order cumulants are of order \( O(n^{-3/2}) \) or smaller. The derivation for the skewness of \( \tilde{r} \) is given in Appendix C, and the result is

\begin{align*}
\kappa_3(\tilde{r}) &= u_{11}^{-3/2} \{ \nu_{1,1,1} + u_{11}^{-1} \nu_{1,1} (\nu_3 u_{11}^{-1} \nu_{1,1} + 3 \nu_{11,1}) \} + O(n^{-3/2}).
\end{align*}

(2.47)

Equivalently, the standardized third cumulant is

\begin{align*}
\rho_3(\tilde{r}) &= \nu_{1,1}^{-3/2} \{ \nu_{1,1,1} + u_{11}^{-1} \nu_{1,1} (\nu_3 u_{11}^{-1} \nu_{1,1} + 3 \nu_{11,1}) \} + O(n^{-3/2}).
\end{align*}

(2.48)

Let \( s(\theta) = u_{11}^{-3/2} \{ \nu_{1,1,1} + u_{11}^{-1} \nu_{1,1} (\nu_3 u_{11}^{-1} \nu_{1,1} + 3 \nu_{11,1}) \} \). Therefore, if \( s(\theta) = O(n^{-1}) \) or
smaller, then the normal approximation to the distribution of the modified signed composite likelihood ratio has an error of order \(O(n^{-3/2})\) as desired. The Bartlett identities indicate that this condition is satisfied if \(\tilde{l}\) is the ordinary log likelihood. As pointed out in Stern (2006), in the discussion for \(M\)-estimators, this condition can be satisfied more widely, in particular for the location family with a symmetric density about the origin, which leads to all third-order or odd-order cumulants zero, in particular, \(\nu_{1,1,1} = \nu_{11,1} = \nu_3 = 0\). However, such symmetry is not typical for composite likelihood, and \(s(\theta)\) is only of typical order \(O(n^{-1/2})\).

### 2.5 Example

We shall illustrate the results above using the symmetric normal distribution. The loss of efficiency of the composite likelihood with increasing \(q\) is illustrated in Cox and Reid (2004). We aim to use this single parameter case to check the third order cumulant for the signed composite likelihood ratio statistic.

Consider \(n\) independent symmetric normal random variables with correlation \(\rho\), each \(q\)-dimensional. The pairwise likelihood, which is defined through the bivariate normal density function for distinct pairs, is given by

\[
l_2(\rho) = -\frac{\text{nq}(q-1)}{4} \log(1 - \rho^2) - \frac{q - 1 + \rho}{2(1 - \rho^2)} \text{SSW} - \frac{(q - 1)(1 - \rho)}{2(1 - \rho^2)} \text{SSB} \cdot q, \tag{2.49}
\]
where

\[ SS_W = \sum_{i=1}^{n} \sum_{r=1}^{q} (Y_r^{(i)} - \bar{Y}_i)^2, \quad SS_B = \sum_{i=1}^{n} Y_i^{(i)^2}, \quad Y_i^{(i)^2} = \left( \sum_{r=1}^{q} Y_r^{(i)} \right)^2. \]

The corresponding pairwise score function is

\[ U_2(\rho) = l'_2(\rho) = \frac{nq(q - 1)\rho}{2(1 - \rho^2)} - \frac{1 + \rho^2 + 2(q - 1)\rho}{2(1 - \rho^2)^2} SS_W + \frac{(q - 1)(1 - \rho^2)^2 SS_B}{2(1 - \rho^2)^2}. \] (2.50)

Furthermore, the second derivative of the pairwise likelihood is

\[ l''_2(\rho) = \frac{nq(q - 1)(1 + \rho^2)}{2(1 - \rho^2)^2} - \frac{(1 - \rho^2)^2}{(1 - \rho^2)^4} \{(q - 1) + 3\rho + 3(q - 1)\rho^2 + \rho^3\} SS_W \\
- \frac{(q - 1)(1 - \rho^2)(1 - \rho^3 SS_B)}{(1 - \rho^2)^4}. \]

In addition, for future use, the explicit forms for the third- and forth-order derivatives are

\[ l'''_2(\rho) = \frac{nq(q - 1)(3\rho + \rho^3)}{(1 - \rho^2)^3} - \frac{3}{(1 - \rho^2)^4} \{1 + 4(q - 1)\rho + 6\rho^2 + 4(q - 1)\rho^3 + \rho^4\} SS_W \\
+ \frac{3(q - 1)(1 - \rho^4 SS_B)}{(1 - \rho^2)^4}. \]

\[ l^{(4)}_2(\rho) = \frac{3nq(q - 1)(1 + 6\rho^2 + \rho^4)}{(1 - \rho^2)^4} - \frac{12}{(1 - \rho^2)^5} \{(q - 1) + 5\rho + 10(q - 1)\rho^2 + 10\rho^3 + 5(q - 1)\rho^4 + \rho^5\} SS_W \\
- \frac{12(q - 1)}{(1 - \rho^2)^5} \{1 + 2\rho + 7\rho^2 + 13\rho^3 + 8\rho^4 - \rho^5\} SS_B. \]

It can be shown that \( SS_B \) and \( SS_W \) are independent, and \( SS_W \sim \sum n \chi_{q-1}^2(1 - \rho) \) and \( SS_B/q \sim \sum n \chi^2(1 + (q - 1)\rho) \). Hence, it is feasible to find the cumulants of various orders of the pairwise likelihood together with its derivatives. We are particularly interested in the
second- and third-order cumulants, since we know Bartlett identities do not hold in general for composite likelihood case, except for the first-order.

In the notation of Section 4, we have

\[ u_{11} = \mathbb{E} [-l''_2(\rho)] = \frac{nq(q - 1)}{2} \frac{\rho^2 + 1}{(1 - \rho^2)^2} \]  

(2.51)

and

\[ \nu_{1,1} = \mathbb{E} [\nu^2_2(\rho)] = \frac{nq(q - 1)}{2} \frac{(1 - \rho)^2 c(q, \rho)}{(1 - \rho^2)^4} \]  

(2.52)

where

\[ c(q, \rho) = (1 - \rho)^2 (3 \rho^2 + 1) + q \rho (-3 \rho^3 + 8 \rho^2 - 3 \rho + 2) + q^2 \rho^2 (1 - \rho)^2, \]

and \( \nu_{1,1} \neq u_{11} \) unless \( q = 2 \), when the pairwise likelihood function equates the full likelihood function.

The details of the derivation for the third-order cumulants are shown in Appendix D. We only list the results here

\[ \nu_3 = -2nq(q - 1) \frac{\rho (3 + \rho^2)}{(1 - \rho^2)^3} \]  

(2.53)

\[ \nu_{1,1,1} = nq(q - 1) \frac{(1 - \rho)^3 d(p, \rho)}{(1 - \rho^2)^6} \]  

(2.54)

\[ \nu_{11,1} = nq(q - 1) \frac{(1 - \rho)^2 p e(q, \rho)}{(1 - \rho^2)^5} \]  

(2.55)
where

\[ d(p, \rho) = -\rho^3(p - 1)^3q^4 + \rho^2(p - 1)^3(5\rho - 3)q^3 - \rho(-47\rho^2 + 69\rho^3 - 42\rho^4 + 10\rho^5 + 21p - 3)q^2 \]
\[ + (p - 1)^2(10\rho^4 - 28\rho^3 + 15\rho^2 - 10\rho + 1)q - (p - 1)^4(5\rho^2 - \rho + 2), \]
\[ e(q, \rho) = \rho(p - 1)^3q^2 - \rho(p - 3)(3\rho^2 - 2\rho + 3)q + (p - 1)^2(3\rho^2 - 2\rho + 3). \]

We can see that the ratio \( u_{11}^{-1} \nu_{1,1} \) is \( O(q^2) \), and the Bartlett second identity doesn’t hold. Furthermore, the skewness of \( \tilde{r} \) is \( O(n^{-1/2}) \) for fixed \( q \) as \( n \to \infty \) and is \( O(q^3) \) for fixed \( n \) as \( q \to \infty \). The standardization of the skewness results in \( \rho_3(\tilde{r}) \) of order \( O(n^{-1/2}) \). However, what we need is \( O(n^{-1}) \) or smaller. Accordingly the goal to construct higher accuracy inference in terms of \( \tilde{r} \), by means of the mean and variance correction, can not be achieved as we had hoped.

The attainment of higher accuracy inference primarily depends on the Bartlett identities. The required condition on the third- and fourth-order cumulants are in direct analogy with the Bartlett third and fourth identities. Given that, in general, unless the composite likelihood objects are not too far away from the true likelihood, higher accuracy inference based on the signed composite likelihood ratio statistic is not achievable.
Chapter 3

Multivariate normal distribution

This chapter is devoted to studying the performance of the composite likelihood approach applied to the multivariate normal distribution, by illustrating explicit computation of the composite likelihood estimates together with the associated information quantities. The model allows us to explore a number of interesting aspects of composite likelihood method. It is shown that, for a general class of multivariate normal distribution, the composite likelihood is fully efficient and provides identical estimates compared to the ordinary full likelihood.

3.1 Intraclass Correlation

We start with a particular simple example of multivariate normal distribution. Consider a random sample $Y^{(1)}, \ldots, Y^{(n)}$ from the equicorrelation multivariate normal distribution,
3.1. INTRACLASS CORRELATION

that is

\[ Y^{(i)} \sim N_q(\mathbf{\mu}, \sigma^2 V), \]

where the correlation matrix is specified as \( V = (1 - \rho)I + \rho J \) with \( J = \mathbf{1}_q \mathbf{1}_q^T \), \( \mathbf{1}_q \) being a \( q \)-dimensional vector of 1's. For simplicity, we assume equal mean, i.e. \( \mathbf{\mu} = \mu \mathbf{1}_q \).

### 3.1.1 Maximum estimate of full likelihood

The joint probability density function is given by

\[
f(Y; \theta) = (2\pi)^{-\frac{nq}{2}} \cdot |\Sigma|^{-\frac{n}{2}} \cdot \exp\{\text{tr}(-\frac{1}{2} \Sigma^{-1} \sum_{i=1}^{n} (Y^{(i)} - \mathbf{\mu})(Y^{(i)} - \mathbf{\mu})^T)\}, \tag{3.1}
\]

where \( \theta = (\mu, \sigma^2, \rho)^T \). We have \( \sum_{i=1}^{n} (Y^{(i)} - \mathbf{\mu})(Y^{(i)} - \mathbf{\mu})^T = A + n(\bar{Y}^{(\cdot)} - \mathbf{\mu})(\bar{Y}^{(\cdot)} - \mathbf{\mu})^T \),

where the matrix \( A \) and the average \( \bar{Y}^{(\cdot)} \) are given by

\[
A = \sum_{i=1}^{n} (Y^{(i)} - \bar{Y})(Y^{(i)} - \bar{Y})^T, \quad \bar{Y}^{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)} \tag{3.2}
\]

respectively. The general results for MLE of a multivariate normal distribution are

\[
\hat{\mathbf{\mu}} = \bar{Y}^{(\cdot)}, \quad \hat{\Sigma} = A/n.
\]
By imposing the model assumptions \( \mu = \mu_1 q \) and \( \Sigma = \sigma^2 V \), we can deduce \( \hat{\mu}, \hat{\sigma}^2 \) and \( \hat{\rho} \) from \( \hat{\mu} \) and \( \hat{\Sigma} \) as

\[
\hat{\mu} = \bar{Y}, \quad (3.3)
\]

\[
\hat{\sigma}^2 = \frac{S_1}{nq}, \quad (3.4)
\]

\[
\hat{\rho} = \frac{2S_2}{(q-1)S_1}, \quad (3.5)
\]

by using the notations

\[
\bar{Y} = \frac{1}{nq} \sum_{i=1}^{n} \sum_{r=1}^{q} Y_r^{(i)},
\]

\[
S_1 = \sum_{i=1}^{n} \sum_{r=1}^{q} (Y_r^{(i)} - \bar{Y})^2,
\]

\[
S_2 = \sum_{i=1}^{n} \sum_{s>r}^{q} (Y_r^{(i)} - \bar{Y})(Y_s^{(i)} - \bar{Y}).
\]

### 3.1.2 Expected Fisher information

To derive the expected Fisher information matrix, we use the results (3.7) - (3.8) to obtain the log-likelihood function from (3.1) as

\[
l(\theta) = -\frac{nq}{2} \log \sigma^2 - \frac{n(q-1)}{2} \log (1 - \rho) - \frac{n}{2} \log (1 + (q-1)\rho)
\]

\[
- \frac{1}{2\sigma^2(1-\rho)} \sum_{i=1}^{n} \sum_{r=1}^{q} (Y_r^{(i)} - \mu)^2
\]

\[
+ \frac{\rho}{2\sigma^2(1-\rho)(1 + (q-1)\rho)} \sum_{i=1}^{n} \sum_{r,s=1}^{q} (Y_r^{(i)} - \mu)(Y_s^{(i)} - \mu), \quad (3.6)
\]
where

\[ |\Sigma| = |\sigma^2 V| = \sigma^{2q} (1 - \rho)^{q-1} (1 + (q - 1)\rho), \tag{3.7} \]

\[ \Sigma^{-1} = \sigma^{-2} V^{-1} = \sigma^{-2} \left( \frac{1}{1 - \rho} I - \frac{\rho}{(1 - \rho)(1 + (q - 1)\rho)} J \right), \tag{3.8} \]

since

\[ \text{tr}\left\{ \sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T \right\} = \sum_{i=1}^{n} \sum_{r=1}^{q} (Y^{(i)}_r - \mu)^2, \]

\[ \text{tr}\left\{ J \sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T \right\} = \sum_{i=1}^{n} \sum_{r,s=1}^{q} (Y^{(i)}_r - \mu)(Y^{(i)}_s - \mu)^T. \]

We note that the MLEs (3.3) - (3.5) can be obtained by differentiating (3.6) directly.

The notations \( \hat{\Sigma}_{\sigma^2}, \tilde{\Sigma}_{\sigma^2}, \) and \( \hat{\Sigma}_{\rho}, \tilde{\Sigma}_{\rho} \) are used to denote the elementwise derivatives of \( \Sigma \) with respect to \( \sigma^2 \) and \( \rho \), and, similarly, the notations \( \hat{\Sigma} \) and \( \tilde{\Sigma} \). We have

\[ \frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} 1_q^T \Sigma^{-1}(Y^{(i)} - \mu), \]

\[ \frac{\partial l}{\partial \sigma^2} = -\frac{nq}{2\sigma^2} + \frac{1}{2\sigma^4} \text{tr}(\sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T V^{-1}), \]

\[ \frac{\partial l}{\partial \rho} = -\frac{n}{2} \text{tr}(V^{-1} \dot{V}_\rho) + \frac{1}{2\sigma^4} \text{tr}(\sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T V^{-1} \dot{V}_\rho V^{-1}). \]
and

\[
\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} 1_q^T V^{-1} 1_q,
\]

\[
\frac{\partial^2 l}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^{n} 1_q^T V^{-1} (Y^{(i)} - \mu),
\]

\[
\frac{\partial^2 l}{\partial \mu \partial \rho} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} 1_q^T V^{-1} \dot{V}_\rho V^{-1} (Y^{(i)} - \mu),
\]

\[
\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{nq}{2\sigma^4} - \frac{1}{\sigma^6} \text{tr} \left( \sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T V^{-1} \right),
\]

\[
\frac{\partial^2 l}{\partial \rho^2} = \frac{n}{2} \text{tr}(V^{-1} \dot{V}_\rho V^{-1} \dot{V}_\rho) - \frac{n}{2} \text{tr}(V^{-1} \ddot{V}_\rho),
\]

\[
\frac{\partial^2 l}{\partial \rho \partial \sigma^2} = -\frac{1}{2\sigma^4} \text{tr} \left( \sum_{i=1}^{n} (Y^{(i)} - \mu)(Y^{(i)} - \mu)^T V^{-1} \dot{V}_\rho V^{-1} \dot{V}_\rho \right)
\]

The expected Fisher information matrix is then obtained as

\[
I(\theta) = \mathbb{E} \left( -\frac{\partial^2 l}{\partial \theta \partial \theta^T} \right) = \begin{pmatrix}
\frac{1}{\sigma^2} \sum_{i=1}^{n} 1_q^T V^{-1} 1_q & 0 & 0 \\
0 & \frac{nq}{2\sigma^4} & \frac{n}{2\sigma^2} \text{tr}(V^{-1} \dot{V}_\rho) \\
0 & \frac{n}{2\sigma^2} \text{tr}(V^{-1} \dot{V}_\rho) & \frac{n}{2} \text{tr}(V^{-1} \dot{V}_\rho V^{-1} \dot{V}_\rho)
\end{pmatrix}
\]
where

\[ \sum_{i=1}^{n} 1_i^T V^{-1} 1_i = \frac{nq}{1 + (q-1)\rho}, \]

\[ \text{tr}(V^{-1} \dot{V}_\rho) = -\frac{q(q-1)\rho}{(1-\rho)(1 + (q-1)\rho)}, \]

\[ \text{tr}(V^{-1} \dot{V}_\rho V^{-1} \dot{V}_\rho) = \frac{q(q-1)}{(1-\rho)^2} - \frac{2q(q-1)^2\rho}{(1-\rho)^2(1 + (q-1)\rho)} + \left(\frac{q(q-1)\rho}{(1-\rho)(1 + (q-1)\rho)}\right)^2. \]

3.1.3 Maximum estimate of pairwise likelihood

We now apply the composite likelihood approach to this equicorrelation multivariate normal distribution. Since no parameters more than means, variances and correlations are included in the distribution, pairwise likelihood would be a natural choice, defined via the bivariate normal density functions for distinct pairs, and hence the pairwise log-likelihood is

\[ l_2(\theta; Y) = \sum_{i} \sum_{s > r} \log f(Y^{(i)}_r, Y^{(i)}_s; \theta), \]

where

\[ f(Y^{(i)}_r, Y^{(i)}_s; \theta) = \frac{1}{2\pi\sigma^2 \sqrt{1 - \rho^2}} \exp\left[-\frac{1}{2\sigma^2(1 - \rho^2)}[(Y^{(i)}_r - \mu)^2 - 2\rho(Y^{(i)}_r - \mu)(Y^{(i)}_s - \mu) + (Y^{(i)}_s - \mu)^2]\right] \]
Simplification gives the pairwise log-likelihood function as

\[
l_2(\theta; Y) = -\frac{nq(q-1)}{2} \log \sigma^2 - \frac{nq(q-1)}{4} \log(1 - \rho^2) - \frac{1}{2\sigma^2(1 - \rho^2)} \sum_i \sum_{s>r} [(Y_r^{(i)} - \mu)^2 - 2\rho(Y_r^{(i)} - \mu)(Y_s^{(i)} - \mu) + (Y_s^{(i)} - \mu)^2]
\]

(3.10)

Solving the first derivatives of the pairwise likelihood with respect to each component of \( \theta \) yields the maximum pairwise likelihood estimators, denoted by \( \tilde{\theta} \), which are

\[
\tilde{\mu} = \bar{Y}_- \quad \tilde{\sigma}^2 = \frac{S_1}{nq} \quad \tilde{\rho} = \frac{2S_2}{(q-1)S_1}.
\]

(3.11)

The calculation is straightforward but tedious, and details would be skipped here. We can see that the estimators obtained from the pairwise likelihood and those of the full likelihood are identical.

Removing the assumption \( \mu = \mu_1 \), the same conclusion will still hold, but \( S_1 \) and \( S_2 \) will have different forms, because \( \hat{\mu} \) will be \( \bar{Y}^- \) not \( \bar{Y}_- \).

### 3.1.4 Godambe information

Since \( \hat{\theta} = \tilde{\theta} \), they have the same variance, it must be the case that \( \tilde{\theta} \) is fully efficient, i.e. that

\[
I(\theta) = H(\theta)J^{-1}(\theta)H(\theta)
\]
even though it is easy to check that $H(\theta) \neq J(\theta)$. Note that the left hand side of the equation is the Fisher information matrix derived from the full likelihood, while the RHS is the Godambe matrix consisting of pairwise objects. To verify this, we give the explicit form of the Godambe information matrix.

Of the two matrices on the RHS, it is easier to calculate the expected score derivative matrix $H(\theta)$

$$ H(\theta) = E\left(-\frac{\partial^2 l_2}{\partial \theta \partial \theta^T}\right) = nq(q - 1) \begin{pmatrix} \frac{1}{\sigma^2(1 + \rho)} & 0 & 0 \\ 0 & \frac{1}{2\sigma^4} & \frac{-\rho}{2\sigma^2(1 - \rho^2)} \\ 0 & \frac{-\rho}{2\sigma^2(1 - \rho^2)} & \frac{1 + \rho^2}{2(1 - \rho^2)^2} \end{pmatrix}. \hspace{1cm} (3.12) $$

A first result concerning the expected squared score matrix $J(\theta)$ is

$$ J(\theta) = E\left(\frac{\partial l_2}{\partial \theta} \frac{\partial l_2}{\partial \theta^T}\right) = \begin{pmatrix} \frac{nq(q - 1)^2(1 + (q - 1)\rho)}{\sigma^2(1 + \rho)^2} & 0 & 0 \\ 0 & j_{\sigma^2\sigma^2} & j_{\sigma^2\rho} \\ 0 & j_{\rho\sigma^2} & j_{\rho\rho} \end{pmatrix}. \hspace{1cm} (3.13) $$

where $j_{\sigma^2\sigma^2} = E\left\{\left(\frac{\partial l_2}{\partial \sigma^2}\right)^2\right\}$, $j_{\sigma^2\rho} = E\left\{\frac{\partial l_2}{\partial \sigma^2} \frac{\partial l_2}{\partial \rho}\right\}$, and $j_{\rho\rho} = E\left\{\left(\frac{\partial l_2}{\partial \rho}\right)^2\right\}$. Our ultimate goal is to compare the two variance matrices of the identical estimators. It is computationally easier to verify, instead of solving $J(\theta)$ completely,

$$ I^{-1}(\theta) = H^{-1}(\theta)J(\theta)H^{-1}(\theta). $$
and meanwhile we avoid to specify the forms of $j_{\sigma^2\sigma^2}$, $j_{\sigma^2\rho}$, and $j_{\rho\rho}$.

It is straightforward to verify $(I^{-1}(\theta))_{11} = (H^{-1}(\theta)J(\theta)H^{-1}(\theta))_{11}$, i.e. $I^{\mu\mu} = G^{\mu\mu}$. Since $\mu$ is orthogonal to $(\sigma^2, \rho)$, we have all zeros in the information matrices except the lower diagonal $2 \times 2$ submatrices for $(\sigma^2, \rho)$, denoted by $I_{\sigma^2\rho}$, $H_{\sigma^2\rho}$, and $J_{\sigma^2\rho}$, respectively. Tedious yet straightforward calculation yields

$$I_{\sigma^2\rho}^{-1} = \frac{2}{nq(q-1)} \begin{pmatrix} (q-1)(1 + (q-1)\rho^2)\sigma^4 & (q-1)\rho(1-\rho)(1 + (q-1)\rho)\sigma^2 \\ (1-\rho)(1 + (q-1)\rho)^2 & (1-\rho^2)^2 \end{pmatrix}$$

$$H_{\sigma^2\rho}^{-1} = \frac{2}{nq(q-1)} \begin{pmatrix} (1+\rho^2)\sigma^4 & \rho(1-\rho^2)\sigma^2 \\ (1-\rho^2)^2 & (1-\rho^2)^2 \end{pmatrix}$$

To further simplify the calculation, it is reasonable to verify the equivalence for one of the components in the $2 \times 2$ submatrices, rather than verifying all of them. The reason is that if $j_{\sigma^2\sigma^2}$, $j_{\sigma^2\rho}$, and $j_{\rho\rho}$ satisfy one equivalence, they must work for all of them, otherwise it will cause contradictions. For simplicity we choose to verify the following one

$$(I_{\sigma^2\rho}^{-1})_{\rho\rho} = ((H^{-1}JH^{-1})_{\sigma^2\rho})_{\rho\rho}$$
3.1. INTRACLASS CORRELATION

which stands for the variances for $\rho$. More explicitly,

$$(I_{\sigma^2 \rho})_{\rho \rho} = \frac{2}{nq(q-1)}((1-\rho)(1+(q-1)\rho))^2$$

$$((H^{-1}JH^{-1})_{\sigma^2 \rho})_{\rho \rho} = \left(\frac{2}{nq(q-1)}\right)^2 [\rho^2(1-\rho^2)^2 \sigma^4 j_{\sigma^2 \rho} + 2\rho(1-\rho^2)^3 \sigma^4 j_{\sigma^2 \rho} + (1-\rho^2)^4 j_{\rho \rho}]$$

$$= \left(\frac{2}{nq(q-1)}\right)^2 E \{[\rho(1-\rho^2)\sigma^2(\frac{\partial l_2}{\partial \sigma^2}) + (1-\rho^2)^2(\frac{\partial l_2}{\partial \rho})]^2\}$$

Denote the target expectation as

$$E\{\ast\} = E\{[\rho(1-\rho^2)\sigma^2(\frac{\partial l_2}{\partial \sigma^2}) + (1-\rho^2)^2(\frac{\partial l_2}{\partial \rho})]^2\}$$

Substituting the pairwise score functions, we have

$$E\{\ast\} = E \left\{ \left( -\frac{\rho}{2\sigma^2} \sum_i \sum_{s>r} [(Y_r^{(i)} - \mu)^2 + (Y_s^{(i)} - \mu)^2] + \frac{1}{\sigma^2} \sum_i \sum_{s>r} (Y_r^{(i)} - \mu)(Y_s^{(i)} - \mu) \right)^2 \right\}$$

$$= E \left\{ \left( -\frac{\rho}{2} \sum_i \sum_{s>r} (Z_r^{(i)} + Z_s^{(i)}) + \sum_i \sum_{s>r} Z_r^{(i)} Z_s^{(i)} \right)^2 \right\}$$

$$= \frac{\rho^2}{4} E \{[\sum_i \sum_{s>r} (Z_r^{(i)} + Z_s^{(i)} - 2\frac{Z_r^{(i)} Z_s^{(i)}}{\rho})]^2\}$$

where $Z^{(i)} \sim N(0, V)$, and recall that we denote $V = (1-\rho)I + \rho J$.

Let $W_i = \sum_{s>r}(Z_r^{(i)} + Z_s^{(i)} - 2\frac{Z_r^{(i)} Z_s^{(i)}}{\rho})$. The independence and identity of $Z^{(i)}$’s secure

$$E \{ \sum_i W_i^2 \} = E \{ \sum_i W_i^2 \} = nE \{ W_1^2 \}$$
3 Multivariate normal distribution

Hence, we can further simplify the target to \( \mathbb{E}\{*\} = \frac{n\rho^2}{4} \mathbb{E}\{(\sum_{s>r} V_{rs})^2\} \) by introducing

\[
V_{rs} = Z_r^2 + Z_s^2 - \frac{2}{\rho} Z_r Z_s.
\]

In order to find an expression for \((\sum_{s>r} V_{rs})^2\), for the sake of brevity in the notation, we shall henceforth suppress \( V \) for \( V_{rs} \), and only write down the subscripts which display all information we are interested. To be more specific,

\[
(\sum_{s>r} V_{rs})^2 = \frac{1}{4} \sum_r \sum_{s \neq r} \sum_{r' \neq s'} V_{rs} V_{r's'}
\]

\[
= \frac{1}{4} \left[ \sum_r \sum_{s \neq r} r s s r + \sum_r \sum_{s \neq r, s' \neq r} r s s s' + \sum_r \sum_{s \neq r} r s r s + \sum_r \sum_{s \neq r} r s r s'ight]
\]

As \( r, s, r' \) and \( s' \) are equally exchangeable, we have

\[
\mathbb{E}\{\sum_r \sum_{s \neq r} r s s r\} = \mathbb{E}\{\sum_r \sum_{s \neq r} r s r s\},
\]

\[
\mathbb{E}\{\sum_r \sum_{s \neq r} \sum_{s' \neq r,s} r s s s'\} = \mathbb{E}\{\sum_r \sum_{s \neq r} \sum_{s' \neq s,r} r s r s'\} = \mathbb{E}\{\sum_r \sum_{s \neq r} \sum_{s' \neq s,r} r s r s'\} = \mathbb{E}\{\sum_r \sum_{s \neq r} \sum_{s' \neq s,r} r s r s'\}.
\]
Finally, after thorough calculations, we obtain the values of the following expectations

\[
E\left\{ \sum_r \sum_{s \neq r} rssr \right\} = 4q(q-1)(\rho - \frac{1}{\rho})^2
\]

\[
E\left\{ \sum_r \sum_{s \neq r} \sum_{s' \neq r} rs's'r's' \right\} = 2q(q-1)(q-2)(3\rho^2 - 4\rho + \frac{2}{\rho} - 1)
\]

\[
E\left\{ \sum_r \sum_{s \neq r} \sum_{s' \neq r,s} rsr's'r's' \right\} = 8q(q-1)(q-2)(q-3)(\rho - 1)^2
\]

Therefore, combining all the results together gives us

\[
((H^{-1}JH^{-1})_{\sigma^2 \rho})_{\mu \mu} = \left( \frac{2}{nq(q-1)} \right)^2 E\{*\}
\]

\[
= \left( \frac{2}{nq(q-1)} \right)^2 \frac{n\rho^2}{4} E\left\{ \left( \sum_{s>r} V_{rs} \right)^2 \right\}
\]

\[
= \frac{2}{nq(q-1)} \left[ 1 + 2(q-2)\rho + (q^2 - 6q + 6)\rho^2 - 2(q-1)(q-2)\rho^3 + (q-1)^2 \rho^4 \right]
\]

\[
= \frac{2}{nq(q-1)} ((1 - \rho)(1 + (q-1)\rho))^2
\]

\[
= (I^{-1}_{\sigma^2 \rho})_{\mu \mu}
\]

Furthermore, the equivalence with respect to the variance of \( \sigma^2 \) or the covariance of \( \sigma^2 \) and \( \rho \) follows consequently to ensure consistency of the values of \( j_{\sigma^2 \sigma^2} \), \( j_{\sigma^2 \rho} \), and \( j_{\rho \rho} \). That is, \( I^{-1} = H^{-1}JH^{-1} \), or equivalently, \( I = HJ^{-1}H \). Therefore, we can conclude that not only the pairwise likelihood estimator and the full likelihood estimator are identical, but also they yield the same information matrices. The pairwise likelihood estimator is fully efficient for Intraclass Correlation Normal.
3.1.5 Maximum estimate of pseudolikelihood

We now consider a pseudo log-likelihood proposed in Cox and Reid (2004),

\[ l^* = l_2 - aql_1 \]

\[
= \sum_i \sum_{s>r}^q \log f(Y_r^{(i)}, Y_s^{(i)}; \theta) - aq \sum_i \sum_{r=1}^q \log f(Y_r^{(i)}; \theta)
\]

\[
= -\frac{nq(q-1)}{2} \log \sigma^2 + \frac{aq}{2} nq \log \sigma^2 - \frac{nq(q-1)}{4} \log(1 - \rho^2)
\]

\[
-\frac{1}{2\sigma^2(1-\rho^2)} \sum_i \sum_{s>r}^q [(Y_r^{(i)} - \mu)^2 - 2\rho(Y_r^{(i)} - \mu)(Y_s^{(i)} - \mu) + (Y_r^{(i)} - \mu)^2]
\]

\[
+ \frac{aq}{2\sigma^2} \sum_i \sum_{r=1}^q (Y_r^{(i)} - \mu)^2
\]

where \( l_1 \) is the independence likelihood, and \( a \) is a constant to be chosen, in principle as a solution to an optimality problem.

Let \( \theta^* \) denote the maximum pseudolikelihood estimator. Differentiating the pseudo log-likelihood with respect to each component of \( \theta \) provides the following results:

\[
\mu^* = \bar{Y}_r \quad (3.15)
\]

\[
\sigma^{2*} = \frac{aqS_1 - (q-1)S_1/2}{anq^2 - nq(q-1)/2} \quad (3.16)
\]

\[
\rho^* = \frac{2S_2}{(q-1)S_1} \quad (3.17)
\]
If \( a = 0 \), the pseudo mle \( \theta^* = \bar{\theta} \), and \( \bar{\theta} \) is identical to the MLE \( \hat{\theta} \). If \( a \neq 0 \), we have

\[
\hat{\mu} = \tilde{\mu} = \mu^*,
\]

\[
\hat{\rho} = \tilde{\rho} = \rho^*,
\]

but \( \sigma^2^* \neq \hat{\sigma}^2 \).

However, if \( a = r/q \), we have

\[
\sigma^2^* = \frac{rS_1 - (q - 1)S_1/2}{nrq - nq(q - 1)/2} = \frac{(r - q^{-1})S_1}{(r - q^{-1})nq} = \frac{S_1}{nq} = \hat{\sigma}^2 = \tilde{\sigma}^2.
\]

The special case, \( a = 0 \), is the usual pairwise likelihood, and \( a = 1/2q \) is the pairwise pseudolikelihood defined in Mardia et al. (2007). This latter is a composite conditional likelihood.

### 3.2 General Unrestricted Multivariate Normal

The discussion in this section follows arguments developed in Mardia et al. (2007). They showed the efficiency of composite likelihood based on conditional densities for the multivariate normal distribution. In their paper the composite conditional likelihood is called pairwise pseudolikelihood, and is defined as

\[
PPL(\Sigma) = \prod_{i=1}^{n} \prod_{r \neq s} f(Y_{s}^{(i)}|Y_{r}^{(i)}; \Sigma)
\]
where $\Sigma$ is the joint covariance matrix, and mean vector is taken to be zero without loss of
generality.

We consider the composite marginal likelihood functions, and compare the pairwise like-
lihood estimation with the ordinary likelihood estimation.

### 3.2.1 MLE and Fisher information

Consider a random sample $Y^{(1)}, \ldots, Y^{(n)}$, and $Y^{(i)} \sim N_q(\mu, \Sigma)$ with an unrestricted covari-
ance matrix $\Sigma$. Well known results are that the maximum likelihood estimator $\hat{\mu} = \bar{Y}^{(i)}$
with $\bar{Y}^{(i)}$ given in (3.2), and

$$
E \left( -\frac{\partial^2 l}{\partial \mu \partial \mu^T} \right) = \Sigma^{-1}, \quad E \left( -\frac{\partial^2 l}{\partial \mu \partial \Sigma} \right) = 0. \quad (3.18)
$$

Without loss of generality from now on we assume $\mu = 0$, and the parameters of interest
are all contained in $\Sigma$. In the following, we adopt the notation $\sigma_{ij}$ for the $(i, j)$th element of
$\Sigma$ with $i = j, \ldots, q$ and $j = 1, \ldots, q$, and write $\mathbf{\sigma}^T = (\sigma_{11}, \sigma_{21}, \ldots, \sigma_{q1}; \sigma_{22}, \ldots, \sigma_{q2}; \ldots; \sigma_{qq})$, for the vector of $q(q + 1)/2$ elements of $\Sigma$.

The log-likelihood function for $\Sigma$ is

$$
l(\Sigma) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \left( \sum_{i=1}^{n} \Sigma^{-1} Y^{(i)} Y^{(i)^T} \right).
$$
Differentiating \( l(\Sigma) \) with respect to \( \Sigma \) gives

\[
\frac{\partial l}{\partial \Sigma} = -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \sum \Sigma^{-1} Y^{(i)} Y^{(i)T} \Sigma^{-1},
\]

\[
\frac{\partial^2 l}{\partial \Sigma \partial \Sigma^T} = \frac{n}{2} \Sigma^{-1} \otimes \Sigma^{-1} - \frac{1}{2} \sum_{i=1}^{n} \Sigma^{-1} \otimes \Sigma^{-1} Y^{(i)} Y^{(i)T} \Sigma^{-1} - \frac{1}{2} \sum_{i=1}^{n} \Sigma^{-1} Y^{(i)} Y^{(i)T} \Sigma^{-1} \otimes \Sigma^{-1},
\]

and the second equation yields the maximum likelihood estimate and the Fisher information matrix for \( \Sigma \)

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)} Y^{(i)T}, \quad I(\Sigma) = \frac{n}{2} \Sigma^{-1} \otimes \Sigma^{-1}
\]

(3.19)

The above results can be extended straightforwardly to \( \mu \neq 0 \): the maximum likelihood estimators are

\[
\hat{\mu} = \bar{Y}(\cdot), \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)} Y^{(i)T},
\]

(3.20)

and the corresponding Fisher information matrix is given by

\[
I(\mu, \Sigma) = n \begin{pmatrix}
    \Sigma^{-1} & 0 \\
    0 & \frac{1}{2} \Sigma^{-1} \otimes \Sigma^{-1}
\end{pmatrix}
\]

(3.21)

### 3.2.2 MPLE and Godambe information

Now we turn into the computation of \( \theta \) based on pairwise likelihood. Let \( \Sigma_{rs} \) be the submatrix of \( \Sigma \) which contains variances and covariance \( \sigma_{rr}, \sigma_{ss}, \) and \( \sigma_{rs} \). The pairwise log-likelihood
function is

\[
  l_2(\sigma) \quad = \quad \sum_{i=1}^{n} \sum_{s>r} f_r^{(i)}(\Sigma) = \sum_{i=1}^{n} \sum_{s>r} \log f(Y_r^{(i)}, Y_s^{(i)}; \Sigma_{rs})
  = \quad -\frac{n}{2} \sum_{s>r} \log |\Sigma_{rs}| - \frac{1}{2} \sum_{i=1}^{n} \sum_{s>r} \text{tr}(\Sigma_{rs}^{-1} Y_r^{(i)} Y_s^{(i)} T_{rs} \Sigma_{rs}^{-1}).
\]

(3.22)

In order to proceed with the calculations, we write the covariance matrix as a parameter containing \(q(q - 1)/2\) \(2 \times 2\) submatrices \(\Sigma_{rs}\), and denote all of them by \(\Sigma_{..} = (\Sigma_{12}, \Sigma_{13}, \ldots, \Sigma_{1q}; \Sigma_{23}, \ldots, \Sigma_{2q}; \ldots; \Sigma_{q-1q})\). Taking derivatives of \(l_2\) with respect to \(\Sigma\) is equivalent to taking derivatives with respect to each \(\Sigma_{rs}\) and rearrange the result according to components. That is

\[
  \frac{\partial l_2}{\partial \Sigma} = \begin{pmatrix}
    \vdots \\
    \frac{\partial l_2}{\partial \Sigma_{rs}} \\
    \vdots
  \end{pmatrix},
\]

and

\[
  \frac{\partial l_2}{\partial \Sigma_{rs}} = -\frac{n}{2} \Sigma_{rs}^{-1} + \frac{1}{2} \sum_{i=1}^{n} \Sigma_{rs}^{-1} Y_r^{(i)} Y_s^{(i)} T_{rs} \Sigma_{rs}^{-1},
\]

(3.23)

which yields the estimate for each submatrix \(\Sigma_{rs}\)

\[
  \hat{\Sigma}_{rs} = \frac{1}{n} \sum_{i=1}^{n} Y_r^{(i)} Y_s^{(i)} T_{rs} = (\hat{\Sigma})_{rs},
\]

(3.24)

and it implies the estimates for \(\hat{\sigma}_{rr}\), \(\hat{\sigma}_{ss}\) and \(\hat{\sigma}_{rs}\) are fully determined, such as \(\sum (Y_r^{(i)})^2/n\), \(\sum (Y_s^{(i)})^2/n\) and \(\sum Y_r^{(i)} Y_s^{(i)}/n\). So compatibility or overlap in \(\hat{\Sigma}_{rs}\) is not a problem. The last
3.2. GENERAL UNRESTRICTED MULTIVARIATE NORMAL

Equation shows that the maximum pairwise likelihood estimator (MPLE) of $\Sigma_{rs}$ is equal to the $2 \times 2$ submatrix of the maximum likelihood estimator (MLE) of $\Sigma$. In addition, we can include the mean vector in the parameter, and it is straightforward to obtain $\hat{\mu}_{rs} = \bar{Y}_{rs}$, where $\bar{Y}_{rs} = (\bar{Y}^{(r)}, \bar{Y}^{(s)})^T$, and so $\hat{\mu} = \bar{Y}^{(c)}$.

In summary, we can conclude that MPLE and MLE for the parameter $\theta^T = (\mu^T, \sigma^T)$ are identical for general multivariate normal distribution with unrestricted covariance matrix.

For each bivariate density $l^{(i)}_{rs}$ we can compute the Fisher information for $\Sigma_{rs}$ as

$$i_2(\Sigma_{rs}) = \frac{1}{2} \Sigma_{rs}^{-1} \otimes \Sigma_{rs}^{-1},$$  

which, to avoid overlapping, can be also written in terms of $\sigma$’s by incorporating matrix multiplications,

$$i_2(\sigma_{rr}, \sigma_{rs}, \sigma_{ss}) = \frac{1}{2} C_2(\Sigma_{rs}^{-1} \otimes \Sigma_{rs}^{-1}) C_2^T,$$  

where $C_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

Since $l^{(i)}_{rs}$ is a valid bivariate log likelihood depending only on $\Sigma_{rs}$, following the Bartlett second identity, we have

$$H(\theta) = E \left( - \frac{\partial^2 l_2}{\partial \Sigma \partial \Sigma^T} \right) = E \left( - \sum_{i=1}^{n} \sum_{s>r}^{q} \frac{\partial^2 l^{(i)}_{rs}}{\partial \Sigma_i \partial \Sigma^T} \right) = \sum_{i=1}^{n} \sum_{s>r}^{q} E \left[ (\frac{\partial l^{(i)}_{rs}}{\partial \Sigma_i}) (\frac{\partial l^{(i)}_{rs}}{\partial \Sigma_i})^T \right]$$  

(3.27)
Moreover, due to independence over $i$ we have

$$J(\theta) = E\left[\left(\frac{\partial l_2}{\partial \Sigma}\right)\left(\frac{\partial l_2}{\partial \Sigma}\right)^T\right] = E\left[\left(\sum_{i=1}^{n} \sum_{s>r}^{q} \frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{rs}}\right)\left(\sum_{i=1}^{n} \sum_{s>r}^{q} \frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{rs}}\right)^T\right]$$

(3.28)

where $\frac{\partial l_{rs}^{(i)}}{\partial \Sigma} = \left(\frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{12}}, \cdots, \frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{rs}}, \cdots, \frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{q-1q}}\right)^T$.

Therefore, for submatrix $\Sigma_{rs}$, we have

$$H(\Sigma_{rs}) = \sum_{i=1}^{n} E\left[\left(\frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{rs}}\right)\left(\frac{\partial l_{rs}^{(i)}}{\partial \Sigma_{rs}}\right)^T\right] = J(\Sigma_{rs}) = n\Sigma_{rs}. \quad (3.29)$$

We can extend the result to $\Sigma$ without worrying about overlap or compatibility, since transformations, such as (3.26), equating the parameter $\Sigma$ to the parameter $\sigma$, are always right and there is no overlap issue in the parameter set-up of $\sigma$. Hence, first we can conclude that we have information identity for the multivariate normal distribution, i.e. $H(\theta) = J(\theta)$.

Furthermore, comparing $ni_2(\Sigma_{rs})$ to $I(\Sigma)$ in equation (3.20), we have

$$H(\Sigma_{rs}) = J(\Sigma_{rs}) = (I(\Sigma))_{rs} \quad (3.30)$$

It follows that pairwise likelihood estimation is as efficient as full likelihood estimation.

We now extend to non-zero mean vector $\mu$. It is straightforward to show

$$H(\mu_{rs}) = J(\mu_{rs}) = n\Sigma_{rs}^{-1} = (I(\mu))_{rs}. \quad (3.31)$$
where $\mathbf{\mu} = (\mu_r, \mu_s)^T$, and that the $(\mathbf{\mu}, \Sigma)$ elements in the information matrices are zero. Therefore, we have $G(\theta) = I(\theta)$ where $\theta = (\mathbf{\mu}, \Sigma)$. To summarize,

**Theorem 3.1** For a general multivariate normal distribution with unrestricted covariance matrix $\Sigma$ and mean vector $\mathbf{\mu}$, the pairwise likelihood method provides identical estimators as MLE, and achieves full efficiency.

### 3.3 Discussions

A counter example to the full efficiency of the pairwise likelihood is the symmetric normal distribution, with a restricted variance matrix $V = (1 - \rho)I + \rho J$, such that all variances are fixed to be 1. The loss of information with increasing $q$ is discussed in Cox and Reid (2004).

The joint likelihood for such a random sample $Y^{(1)}, \ldots, Y^{(n)}$ is

$$L(\rho; Y) = (2\pi)^{-n/2} \cdot |V|^{-n/2} \cdot \exp(-\frac{1}{2} \text{tr}(V^{-1} \sum_{i=1}^{n} Y^{(i)}Y^{(i)^T})), $$

where $V = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$. 
The associated log-likelihood is

\[ l(\rho; Y) = -\frac{n(q-1)}{2} \log(1 - \rho) - \frac{n}{2} \log\{1 + (q-1)\rho\} - \frac{1}{2(1 - \rho)} SS_W - \frac{1}{2(1 + (q-1)\rho)} SS_B. \]  \hspace{1cm} (3.32)

where \( SS_W = \sum_{i=1}^{n} \sum_{r=1}^{q} (Y_r^{(i)} - \bar{Y}^{(i)})^2 \), \( SS_B = \sum_{i=1}^{n} Y^{(i)}^2 \), and \( Y^{(i)}^2 = (\sum_{r=1}^{q} Y_r^{(i)})^2 \). Consequently, \( \hat{\rho} \) is estimated by the cross products

\[ \hat{\rho} = \frac{1^T Y^T Y 1_q - \text{tr}(Y^T Y)}{nq(q-1)} = \frac{2 \sum_{i=1}^{n} \sum_{s>r}^{q} Y_r^{(i)} Y_s^{(i)}}{nq(q-1)}. \]  \hspace{1cm} (3.33)

Furthermore, the pairwise likelihood is

\[ l_2(\rho; Y) = \sum_{i} \sum_{s>r}^{q} \log f(Y_r^{(i)}, Y_s^{(i)}; \theta) \]

\[ = -\frac{nq(q-1)}{4} \log(1 - \rho^2) - \frac{q - 1 + \rho^2}{2(1 - \rho^2)} SS_W - \frac{(q - 1)(1 - \rho)^2 SS_B}{q}. \]  \hspace{1cm} (3.34)

The MPLE \( \tilde{\rho} \), which is a solution to the pairwise score function, needs to be solved numerically. To be more specific,

\[ U_2(\rho) = \frac{nq(q-1)\rho}{2(1 - \rho^2)} - \frac{1 + \rho^2 + 2(q-1)\rho}{2(1 - \rho^2)^2} SS_W + \frac{(q - 1)(1 - \rho)^2 SS_B}{2(1 - \rho^2)^2}. \]  \hspace{1cm} (3.35)

is the pairwise score function.

For different choices of \( q \), we plot in Figure 3.1 and 3.2 the full and pairwise likelihoods together with the corresponding estimators for \( \rho \).
3.3. DISCUSSIONS

Figure 3.1: Likelihood functions of $\rho$ with $n = 30$ and $q = 3$. The graphs compare the standardized likelihood curves and the estimators of $\rho$; the full likelihood and MLE (in red), and the pairwise likelihood and MPLE (in green). Note that the scale of $\rho$ values is restricted on the same range suitable for both likelihoods.

Figure 3.2: Likelihood functions of $\rho$ with $n = 30$ and $q = 8$. The graphs compare the standardized likelihood curves and the estimators of $\rho$; the full likelihood and MLE (in red), and the pairwise likelihood and MPLE (in green). Note that the scale of $\rho$ values is restricted on the same range suitable for both likelihoods.
To end the chapter, we would mention that the good performance of pairwise likelihood applied to the multivariate normal distribution is believed due to the fact that the multivariate normal distribution is fully parameterized by the means and the covariance matrix. In other words, the first- and second-order structures and parameters fully determine a multivariate normal distribution, and so pairwise likelihood contains all aspects of information as the full likelihood does.
Chapter 4

AR model

This chapter studies the composite likelihood method for first-order autoregressive process.

First, we give a brief review of ordinary likelihood inference. Next, the composite likelihood is presented, the maximum composite likelihood estimator is reviewed and connections with the maximum likelihood estimator are exhibited. Details of the computation for a closed form of the Godambe information are given, and based on these, the efficiency of the composite likelihood relative to the ordinary likelihood is evaluated. The advantage of the analytical form of the Godambe information enables us to examine the performance of the composite likelihood when few observations are available and the asymptotic quantities could be poor approximations to the true values. Finally, the numerical performance of ordinary and composite inference is assessed in a simulation study. We find that composite likelihood method loses little to the ordinary likelihood, appears to be large sample equivalent and is always highly efficient.
4.1 The full likelihood

Let \( \{X_t\} \) be a zero mean Gaussian autoregressive process of order one with correlation coefficient \( a \). It is defined by the difference equation

\[
X_t = aX_{t-1} + \varepsilon_t,
\]

where \( \varepsilon_t \) is a sequence of Gaussian white noise with mean zero and constant variance \( \sigma^2 \). The distribution of a sample \( X = (X_1, \cdots, X_T) \) is not uniquely determined by (4.1) but also depends on the specification of an initial condition. Usually, the initial value \( X_0 \) is assumed to be one of the following:

i) a given constant;

ii) a random variable from \( N(0, \sigma^2/(1 - a^2)) \), and independent of \( \varepsilon_1, \cdots, \varepsilon_T \);

iii) equal to \( X_T \).

It can be shown that under (iii) the minimal sufficient statistic for the parameter \( (a, \sigma^2) \) is \( (\sum X_t^2, \sum X_{t-1}X_t) \) summed over \( t = 1, \ldots, T \), but under (i) and (ii) it has to be supplemented by end corrections. Hence, under (i) and (ii) the proposed AR(1) model belongs to a curved exponential family, while the case (iii) is of less practical interest. For a long sequence there is no need to distinguish the first two cases. We choose the distribution under the case (ii) as our working model with the time series starting at its equilibrium.
4.1. THE FULL LIKELIHOOD

The likelihood function for a sample $X_1, \ldots, X_T$ is given by

$$L(a, \sigma^2) = f(x_1) \prod_{t=2}^{T} f(x_t|x_{t-1})$$

$$= \frac{1}{\sqrt{2\pi \frac{\sigma^2}{1-a^2}}} \exp\left\{-\frac{x_1^2}{2\frac{\sigma^2}{1-a^2}}\right\} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi \sigma}} \exp\left\{-\frac{(x_t - ax_{t-1})^2}{2\sigma^2}\right\}.$$ 

The log-likelihood is

$$l(a, \sigma^2) = \frac{1}{2} \log(1 - a^2) - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2}[S_1 + a^2 S_2 - 2a S_{12}], \quad (4.2)$$

where $S_1 = \sum_{t=1}^{T} x_t^2$, $S_2 = \sum_{t=2}^{T-1} x_t^2$, and $S_{12} = \sum_{t=1}^{T-1} x_t x_{t+1}$.

The maximum likelihood estimators, $\hat{a}$ and $\hat{\sigma}^2$, are the solutions to

$$\begin{cases} 
0 &= (1 - \frac{1}{T})S_2 \hat{a}^3 - (1 - \frac{2}{T})S_{12} \hat{a}^2 - (S_2 + \frac{S_1}{T})\hat{a} + S_{12}, \\
\hat{\sigma}^2 &= (S_1 + \hat{a}^2 S_2 - 2\hat{a} S_{12})/T. 
\end{cases} \quad (4.3)$$

Solving the first equation in (4.3) amounts to finding the roots of the cubic function of $a$,

$$g(a) = a^3 - \frac{T - 2 S_{12}}{T - 1} a^2 - \frac{S_1 + TS_2}{(T - 1)S_2} a + \frac{T}{T - 1} \frac{S_{12}}{S_2},$$

and for large $T$, we have

$$g(a) = a^3 - \frac{S_{12}}{S_2} a^2 - a + \frac{S_{12}}{S_2} + O\left(\frac{1}{T}\right)$$

$$= (a - \frac{S_{12}}{S_2})(a^2 - 1) + O\left(\frac{1}{T}\right).$$
As the root of interest is in \((-1, 1)\), we see that \(\hat{a}\) is asymptotically equal to \(S_{12}/S_2\), and it is readily verified that \(S_{12}/S_2 \xrightarrow{p} a\).

Furthermore, the expected Fisher information can be obtained as

\[
I(\theta) = \mathbb{E} \left( -\frac{\partial^2 l}{\partial \theta \partial \theta^T} \right) = \begin{pmatrix}
\frac{1 + a^2}{(1 - a^2)^2} + \frac{T - 2}{1 - a^2} & \frac{a}{\sigma^2(1 - a^2)} \\
\frac{a}{\sigma^2(1 - a^2)} & \frac{T}{2\sigma^4}
\end{pmatrix}.
\]

\[\text{(4.4)}\]

### 4.2 The composite likelihood

For the AR(1) model, we propose a composite likelihood function formed by adjacent pairs, denoted by \(L_2\)

\[
L_2(a, \sigma^2) = \prod_{t=2}^{T} f(x_t, x_{t-1}),
\]

\[\text{(4.5)}\]

where \((X_t, X_{t-1})\) follows a bivariate normal distribution with mean zero, variance \(\sigma^2/(1-a^2)\), and correlation \(a\). The proposed pairwise likelihood function is

\[
L_2(a, \sigma^2) = f(x_1) \prod_{t=2}^{T} f(x_t | x_{t-1}) \prod_{t=2}^{T-1} f(x_t) = L(a, \sigma^2) \prod_{t=2}^{T-1} f(x_t),
\]

where \(L(a, \sigma^2)\) is the full likelihood function. The pairwise log-likelihood function is

\[
l_2(a, \sigma^2) = l(a, \sigma^2) + \sum_{t=2}^{T-1} \log f(x_t).
\]

\[\text{(4.6)}\]
We can see that the pairwise log-likelihood is composed of the full log-likelihood and an additional term of the sum of the first order margins.

It is straightforward to write the pairwise log-likelihood in terms of \( S_1, S_2, \) and \( S_{12}, \)

\[
l_2(a, \sigma^2) = \frac{T}{2} - \frac{1}{2} \log(1 - a^2) - (T - 1) \log \sigma^2 - \frac{1}{2\sigma^2} (S_1 + S_2 - 2aS_{12}) \tag{4.7}
\]

which yields explicit forms for the maximum pairwise likelihood estimators, MPLE \((\tilde{a}, \tilde{\sigma}^2)\)

\[
\begin{align*}
\tilde{a} &= \frac{2S_{12}}{S_1 + S_2}, \\
\tilde{\sigma}^2 &= \frac{(S_1 + S_2)^2 - 4S_{12}^2}{2(T - 1)(S_1 + S_2)}.
\end{align*}
\tag{4.8}
\]

Simplify the equation of \( \tilde{\sigma}^2 \) to

\[
\tilde{\sigma}^2 = \frac{S_1 + S_2 - \frac{4S_{12}^2}{S_1 + S_2}}{2(T - 1)}
\]

\[
= \frac{S_1}{T} - \frac{S_{12}^2}{TS_1} + O\left(\frac{1}{T^2}\right),
\]

and since

\[
\frac{S_1}{T} \xrightarrow{p} \mathbb{E}[X_t^2] \quad \text{and} \quad \frac{S_{12}^2}{TS_1} = \frac{S_{12}}{T} \cdot \frac{S_{12}}{S_1} \xrightarrow{p} a \mathbb{E}[X_t^2] \cdot a = a^2 \mathbb{E}[X_t^2].
\]

Hence, \( \tilde{\sigma}^2 \xrightarrow{p} \mathbb{E}[X_t^2] - a^2 \mathbb{E}[X_t^2] = \sigma^2. \) That is, the composite likelihood estimator \( \tilde{\sigma}^2 \) is
a consistent estimator of $\sigma^2$ as $T$ increases, which is not always one property of composite likelihood estimators. As noted in Cox and Reid (2004), when we consider the problem where a small number $n$ of individually large sequences is available, i.e. we let the length of the sequence, say $T$ here, increase for fixed number of sequences $n$, the consistency of the resulting estimator relies on the correlation structure.

4.3 Godambe matrix

The main purpose of the present section is to give details for the computation of the Godambe information, which has a general form, as given in Chapter 2

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

where $H(\theta)$ is the sensitivity matrix and $J(\theta)$ is the variability matrix.

It’s straightforward to obtain $H(\theta)$ as

$$H(\theta) = E\left( -\frac{\partial^2 l_2}{\partial \theta \partial \theta^T} \right) = \frac{T - 1}{(1 - a^2)^2 \sigma^4} \begin{pmatrix} (1 + a^2)\sigma^4 & a(1 - a^2)\sigma^2 \\ a(1 - a^2)\sigma^2 & (1 - a^2)^2 \end{pmatrix}. \quad (4.9)$$
4.3. GODAMBE MATRIX

4.3.1 Steps to derive $J(\theta)$

The remaining is devoted to providing details for attaining an explicit form of $J(\theta)$

$$J(\theta) = E\left(\frac{\partial l_2}{\partial \theta} \frac{\partial l_2}{\partial \theta^T}\right) = \begin{pmatrix} j_{aa} & j_{a\sigma^2} \\ j_{\sigma^2a} & j_{\sigma^2\sigma^2} \end{pmatrix}.$$ 

Differentiating the pairwise log-likelihood (4.7) with respect to $a$ and $\sigma^2$ yields

$$\frac{\partial l_2}{\partial a} = -\frac{(T-1)a}{1-a^2} + \frac{S_{12}}{\sigma^2},$$

$$\frac{\partial l_2}{\partial \sigma^2} = -\frac{T-1}{\sigma^2} + \frac{1}{2\sigma^4}(S_1 + S_2 - 2aS_{12}).$$

To compute $j_{aa} = E\left\{\left(\frac{\partial l_2}{\partial a}\right)^2\right\}$, we need $E\{S_{12}^2\}$. An explicit form for $S_{12}^2$ is available in matrix form as

$$S_{12}^2 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1^2x_2^2 & x_1x_2^2x_3 & x_1x_2x_3x_4 & \cdots & x_1x_2x_{T-1}x_T \\
x_1x_2^2x_3 & x_2^2x_3^2 & x_2x_3^2x_4 & \cdots & x_2x_3x_{T-1}x_T \\
\vdots & \vdots & \ddots & \vdots \\ x_{T-1}^2x_Tx_1x_2 & \cdots & x_{T-1}^2x_T^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

For convenience, we suppress $x$ for $x_i$ and focus on the superscripts and subscripts, and suggest so from here to next page. More specifically, $S_{12}^2$ is re-written and further simplified.
as

\[
S_{12}^2 = \begin{pmatrix}
1^2 & 12^2 & 12^3 & 12^4 & \cdots & 12(T-1)T \\
2^2 & 23^2 & 23^3 & 23^4 & \cdots & 23(T-1)T \\
& \ddots & \ddots & \ddots & \ddots & \ddots \\
(T-2)^2 & (T-1)^2 & (T-2)(T-1)^2 & \ddots & \ddots & \ddots \\
& & & (T-1)^2T^2 & \ddots & \ddots \\
& & & & & (T-1)^2T^2 \\
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
\vdots \\
1 \\
1 \\
\end{pmatrix}
\]

\[
= (T-1) \cdot 1^2 + 2 \{(T-2) \cdot 12^2 + (T-3) \cdot 12^3 + (T-4) \cdot 12^4 \}
+ \cdots + 2 \cdot 12(T-2)(T-1) + 12(T-1)T
\]

The last equation is obtained by collecting terms sharing the same 4th-order moment patterns which are determined by subscripts. Now, we can write \( S_{12} \) explicitly in terms of \( x_i \)'s

\[
S_{12}^2 = \sum_{i=1}^{T-1} x_i^2 x_{i+1}^2 + 2 \sum_{i=2}^{T-1} (T-i)x_1x_2x_{i+1}.
\]

As

\[
E \{(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l)\} = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},
\] (4.10)
we have

$$E \{ x_1 x_2 x_i x_{i+1} \} = \sigma_{12} \sigma_{i,i+1} + \sigma_{1i} \sigma_{2,i+1} + \sigma_{1,i+1} \sigma_{2i}$$

$$= (a \frac{\sigma^2}{1 - a^2})^2 + (a^{i-1} \frac{\sigma^2}{1 - a^2})^2 + a^i \frac{\sigma^2}{1 - a^2} a^{i-1} \frac{\sigma^2}{1 - a^2}, \text{ for } i = 2, \cdots, T - 1,$$

as $\sigma_{kl} = a^{\vert l-k \vert} \sigma^2/(1 - a^2)$.

Eventually, the above results lead to

$$E \{ S_{12}^2 \} = (T - 1)E \{ X_1^2 X_2^2 \} + 2 \sum_{i=2}^{T-1} (T - i)(\frac{\sigma^2}{1 - a^2})^2(a^2 + 2a^2(i-1))$$

$$= (\frac{\sigma^2}{1 - a^2})^2 \left[ (T - 1)(1 + 2a^2) + 2 \sum_{i=2}^{T-1} (T - i)(a^2 + 2a^2(i-1)) \right].$$

To find the remaining quantities required in $J(\theta)$, we repeatedly use (4.10) and $\sigma_{ij} = a^{\vert i-j \vert} \sigma^2/(1 - a^2)$. Another key point in the calculation is that we simplify all necessary quantities by combining terms together if they share the same 4th-order moment patterns. The reason is that we only need the values of the expectations, and those values are completely determined by (4.10).
We have the squares or cross product of pairwise score functions as follows

\[
\left( \frac{\partial l_2}{\partial a} \right)^2 = \left( \frac{(T-1)a}{1-a^2} \right)^2 - 2 \frac{(T-1)a}{\sigma^2(1-a^2)} S_{12} + \frac{S_{12}^2}{\sigma^4},
\]

\[
\frac{\partial l_2}{\partial a} \frac{\partial l_2}{\partial \sigma^2} = \left( - \frac{(T-1)a}{1-a^2} + \frac{S_{12}}{\sigma^2} \right) \left( - \frac{T-1}{\sigma^2} + \frac{1}{2\sigma^4}(S_1 + S_2 - 2aS_{12}) \right)
\]

\[
= \left( \frac{\sigma^2}{1-a^2} \right) - \frac{T-1}{2\sigma^4(1-a^2)}(S_1 + S_2 - 2aS_{12}) - \frac{T-1}{\sigma^4} S_{12} + \frac{1}{2\sigma^6} S_{12}(S_1 + S_2 - 2aS_{12}),
\]

\[
\left( \frac{\partial l_2}{\partial \sigma^2} \right)^2 = \left( \frac{T-1}{\sigma^2} + \frac{1}{2\sigma^4}(S_1 + S_2 - 2aS_{12}) \right)^2
\]

\[
= \left( \frac{T-1}{\sigma^2} \right)^2 - \frac{T-1}{\sigma^6}(S_1 + S_2 - 2aS_{12}) + \frac{1}{4\sigma^8}(S_1 + S_2 - 2aS_{12})^2,
\]

and after certain simplification their expectations are

\[
E\{\left( \frac{\partial l_2}{\partial a} \right)^2\} = - \left( \frac{(T-1)a}{1-a^2} \right)^2 + \frac{1}{\sigma^4} E\{S_{12}^2\},
\]

\[
E\left\{ \frac{\partial l_2}{\partial a} \frac{\partial l_2}{\partial \sigma^2} \right\} = - \frac{(T-1)^2a}{\sigma^2(1-a^2)} + \frac{1}{2\sigma^6} E\{S_1S_{12} + S_2S_{12} - 2aS_{12}^2\},
\]

\[
E\left\{ \left( \frac{\partial l_2}{\partial \sigma^2} \right)^2 \right\} = - \left( \frac{T-1}{\sigma^2} \right)^2 + \frac{1}{4\sigma^8} E\{S_1^2 + S_2^2 + 4a^2S_{12}^2 + 2S_1S_2 - 4aS_1S_{12} - 4aS_2S_{12}\}.
\]

Using the same techniques in deriving $S_{12}^2$, after diligent calculations, we have all necessary
quantities as follows

\[
S_{12}^2 = \sum_{i=1}^{T-1} x_i^2 x_{i+1}^2 + 2 \sum_{i=2}^{T-1} (T - i)x_1 x_2 x_i x_{i+1}, \quad (4.11)
\]

\[
S_{1} S_{12} = 2(T - 1)X_1^3 X_2 + 2 \sum_{i=2}^{T-1} (T - i)X_1^2 X_i X_{i+1}, \quad (4.12)
\]

\[
S_{2} S_{12} = 2(T - 2)X_1^3 X_2 + 2 \sum_{i=2}^{T-1} (T - i)X_1 X_2 X_i^2, \quad (4.13)
\]

\[
S_1^2 = T X_1^4 + 2 \sum_{i=2}^{T} (T - i + 1)X_1^2 X_i^2, \quad (4.14)
\]

\[
S_2^2 = (T - 2)X_1^4 + 2 \sum_{i=2}^{T-2} (T - i)X_1^2 X_i^2, \quad (4.15)
\]

\[
S_1 S_2 = (T - 2)X_1^4 + 2 \sum_{i=2}^{T-1} X_1^2 X_i^2, \quad (4.16)
\]

with the corresponding expectations

\[
E\{S_{12}^2\} = \frac{\sigma^2}{1 - a^2}[(T - 1)(1 + 2a^2) + 2 \sum_{i=2}^{T-1} (T - i)(a^2 + 2a^{2(i-1)})], \quad (4.17)
\]

\[
E\{S_1 S_{12}\} = \frac{\sigma^2}{1 - a^2}[6(T - 1)a + 2 \sum_{i=2}^{T-1} (T - i)(a + 2a^{2i-1})], \quad (4.18)
\]

\[
E\{S_2 S_{12}\} = \frac{\sigma^2}{1 - a^2}[6(T - 2)a + 2 \sum_{i=2}^{T-1} (T - i)(a + 2a^{2i-3})], \quad (4.19)
\]

\[
E\{S_1^2\} = \frac{\sigma^2}{1 - a^2}[3T + 2 \sum_{i=1}^{T-1} (T - i)(1 + 2a^{2i})], \quad (4.20)
\]

\[
E\{S_2^2\} = \frac{\sigma^2}{1 - a^2}[3(T - 2) + 2 \sum_{i=3}^{T-1} (T - i)(1 + 2a^{2(i-2)})], \quad (4.21)
\]

\[
E\{S_1 S_2\} = \frac{\sigma^2}{1 - a^2}[3(T - 2) + 2 \sum_{i=2}^{T-1} (T - i)(1 + 2a^{2(i-1)})]. \quad (4.22)
\]
Assembling these facts together forms each item of $J(\theta)$. That is,

\[
\begin{align*}
    j_{aa} &= -\left(\frac{(T-1)a}{1-a^2}\right)^2 + \frac{1}{\sigma^4} E\{S_{12}^2\}, \\
    j_{a\sigma^2} &= -\frac{(T-1)^2 a}{\sigma^2(1-a^2)} + \frac{1}{2\sigma^6} E\{S_1 S_{12} + S_2 S_{12} - 2a S_{12}^2\}, \\
    j_{\sigma^2 \sigma^2} &= -\left(\frac{T-1}{\sigma^2}\right)^2 + \frac{1}{4\sigma^8} E\{S_1^2 + S_2^2 + 4a^2 S_{12}^2 + 2S_1 S_2 - 4a S_1 S_{12} - 4a S_2 S_{12}\},
\end{align*}
\]

(4.23) - (4.25)

with the expectation quantities given by (4.17) - (4.22).

### 4.3.2 Relative efficiency

Based on the forms of $H(\theta)$ and $J(\theta)$ obtained from the previous subsections, we are able to determine the Godambe matrix $G(\theta)$ and further evaluate the efficiency of pairwise likelihood method. The asymptotic efficiency of one estimation method relative to another is defined in terms of the ratio of information matrices for the two methods. Therefore, the efficiency of pairwise likelihood compared to full likelihood for AR(1) model can be obtained by

\[
\left\{ \frac{|H(\theta)J^{-1}(\theta)H(\theta)|}{|I(\theta)|} \right\}^{1/2}
\]

(4.26)

where $\theta = (a, \sigma^2)^T$, and notation $|\cdot|$ stands for matrix determinant.
4.4 Analytical efficiency results

Table 4.1 lists the results of the relative efficiency according to different lengths of AR(1) model with parameter \((a, \sigma) = (0.55, 0.1)\).

<table>
<thead>
<tr>
<th>T</th>
<th>100</th>
<th>20</th>
<th>8</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency</td>
<td>0.995</td>
<td>0.979</td>
<td>0.954</td>
<td>0.940</td>
</tr>
</tbody>
</table>

Table 4.1: Relative efficiency of pairwise likelihood relative to full likelihood based on AR(1) model \(X_t = 0.55X_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, 0.1^2)\) with different lengths \(T\).

In addition, we plot the efficiency as a function of the length \(T\). The table and plot illustrate high efficiency of pairwise likelihood method. With increasing \(T\) the efficiency is getting exceptionally close to 1, and for small \(T = 4\) the relative efficiency is still close to 1.

![Efficiency curve](image)

Figure 4.1: Efficiency of pairwise likelihood relative to full likelihood for AR(1) model \(X_t = 0.55X_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, 0.1^2)\), as a function of \(T\).

The two 3-dimensional plots show the relative efficiency as a function of \(a\) and \(\sigma^2\) for
$T = 4, 20$, respectively. We reach the same conclusions as above. These two plots also show that the relative efficiency drops a little bit, when the coefficient values approach the stationary boundaries.

Figure 4.2: Given $T = 4$, efficiency of pairwise likelihood relative to full likelihood for AR(1) model $X_t = aX_{t-1} + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2)$, as a function of $(a, \sigma)$.
4.5 Simulation study

The simulation study includes mainly, for comparison, the estimation and efficiency of the two methods, full likelihood and composite likelihood, which particularly refers to pairwise likelihood here, applied to AR(1) model.

4.5.1 MLE and MPLE

To get an instant idea of the pairwise likelihood estimation, we draw scatter plots and histograms of MPLEs and compare them to those of MLEs. All plots are made based on 100 simulations of AR(1) model $X_t = 0.55X_{t-1} + \varepsilon_t$, with $\varepsilon \sim N(0, 0.1^2)$ and length $T = 100$. 
Figure 4.4: Scatter plot of MLE and MPLE for $a$ (left panel) and $\sigma^2$ (right panel), from 100 simulations for AR(1) model with $(a, \sigma^2) = (0.55, 0.12)$ and length $T = 100$.

Figure 4.4 displays the scatter plots of the MLE ($\hat{a}, \hat{\sigma}^2$) and MPLE ($\tilde{a}, \tilde{\sigma}^2$) for $a$ (left panel) and $\sigma^2$ (right panel). We can see that the MLEs and MPLEs for $a$ and $\sigma^2$ are nearly identical. Figure 4.5 and 4.6 are the histograms of the MLEs (left panel) and the MPLEs (right panel) for $a$ and $\sigma^2$ respectively from the same 100 simulations. The histograms compare the distribution of estimates together with the corresponding density curves and normal fits. The figures show similar and good estimation for both $a$ and $\sigma^2$ of the pairwise likelihood approach compared to the full likelihood approach.
4.5. SIMULATION STUDY

Figure 4.5: Histograms of the estimators of \( a \), MLE (left panel) and MPLE (right panel), from the same 100 simulations, together with corresponding density curves and normal fits. Note that the scale of \( a \) values is the same for both panels.

Figure 4.6: Histograms of the estimators of \( \sigma^2 \), MLE (left panel) and MPLE (right panel), from the same 100 simulations, together with corresponding density curves and normal fits. Note that the scale of \( \sigma^2 \) values is the same for both panels.
4.5.2 Simulation results

To further evaluate the numerical performance of pairwise likelihood compared with full likelihood, we apply both approaches to the AR(1) model with various choices of \(a(=0.1, 0.55, 0.9)\) and \(\sigma^2(=0.1^2, 1^2, 3^2)\), which permits us to investigate whether different values of \(a\) or \(\sigma^2\), or the length of AR(1) influence the results. For instance, when \(a\) approaches the stationarity bounds, the estimation can be very non-stationary; and when \(a\) goes to 0, the two parameters are asymptotically orthogonal. Simulation results of only positive autocorrelation are presented, as seen from Figures 4.2 and 4.3 and further the formulae of the information matrices that the relative efficiency is symmetric around \(a\). For each pair of the values, we simulate \(N = 1000\) series with lengths \(T = 100, 20\), respectively. For convenience, the tables exhibiting the numerical results are presented at the end.

The estimates (MPLE or MLE) are the averages of the 1000 parameter estimates \(\tilde{\theta}_n\) or \(\hat{\theta}_n\), that is

\[
\tilde{\theta}_{PL} = \frac{1}{1000} \sum_{n=1}^{1000} \tilde{\theta}_n \quad \text{or} \quad \hat{\theta}_F = \frac{1}{1000} \sum_{n=1}^{1000} \hat{\theta}_n.
\]

Pairwise likelihood estimators have derived closed forms, see (4.8). Optimization of the full likelihood function is performed through the quasi-Newton BFGS algorithm with supplied bounds.

Since both \(G(\theta)\) and \(I(\theta)\) have a closed-form, we can substitute the estimators for \(\theta\). The
asymptotic standard errors for MPLE and MLE are

$$\hat{se}(\hat{\theta}_{PL}) = \left(\text{Diag}\{G^{-1}\}\right)^{1/2} \quad \text{and} \quad \hat{se}(\hat{\theta}_F) = \left(\text{Diag}\{I^{-1}\}\right)^{1/2}.$$ 

The general ARE in terms of the expected information matrices is given by

$$\text{ARE}_{\theta}(\hat{\theta}_{PL}, \hat{\theta}_F) = \left\{ \frac{|I^{-1}|}{|G^{-1}|} \right\}^{1/2}.$$ (4.27)

Additionally, it is also reasonable to evaluate AREs for each component of the parameter separately as

$$\text{ARE}_a(\hat{a}_{PL}, \hat{a}_F) = \frac{(I^{-1})_{aa}}{(G^{-1})_{aa}} \quad \text{and} \quad \text{ARE}_{\sigma^2}(\hat{\sigma}^2_{PL}, \hat{\sigma}^2_F) = \frac{(I^{-1})_{\sigma^2\sigma^2}}{(G^{-1})_{\sigma^2\sigma^2}}.$$ (4.28)

Before we comment on the simulation results, an explanation concerning the standard errors is addressed first. Except the results we extracted, as shown in Brockwell and Davis (1991), an asymptotic result on sample autocorrelation functions $\hat{\rho}(h)$ is, for large $T$

$$\hat{\rho}(h) \sim N(\rho(h), T^{-1}W)$$

where $W$ is the covariance matrix whose $(i, j)$-element is given by Bartlett’s formula. In particular, we are interested in

$$\hat{\rho}(1) \sim N(\rho(1), \frac{1-a^2}{T})$$ (4.29)
and $\rho(1) = a$. That is, $\text{Var}(\hat{a}) = (1 - a^2)/T$.

In addition, if we set $a = 0$, then we get a series of i.i.d $N(0, \sigma^2)$'s. The estimator of $\sigma^2$ is given by $\hat{\sigma}^2 = \sum X_i^2 / T$ with variance $\text{Var}(\hat{\sigma}^2) = 2\sigma^4 / T$. So roughly speaking, for the estimator $\hat{\sigma}^2$ of AR(1) we have $\text{Var}(\hat{\sigma}^2) = 2\sigma^4 / T$. As a result, we see rather large standard errors for $\sigma = 3$ scenario. In fact, if $T = 100$, as in Table 4.2, when $\sigma > 2.66$ we have larger than 1 standard errors for the estimators of $\sigma^2$.

In summary, the asymptotic variances for $a$ and $\sigma^2$ are determined by their true values together with the length of the time series. For example, the larger $\sigma^2$, the larger its variance; whereas the closer the absolute values of $a$ to 1, the smaller its variance. Furthermore, since the estimates from pairwise likelihood or full likelihood are fairly close to the true values, we find that the standard errors which are derived from the squared roots of Monte Carlo variances are particularly connected to the parameter values.

From the simulation results we can see that pairwise likelihood and full likelihood synchronize to produce good or bad estimation. Mainly, when the length of AR(1) model is long enough, e.g. $T = 100$, both methods generate good estimators; whereas when $T = 20$, even full likelihood can not yield highly accurate estimates, but, however, pairwise likelihood doesn’t do worse. What’s more, pairwise likelihood approach provides slightly larger estimators for $\sigma^2$, which tends to correct the underestimate problem to a certain extent. Aside from the effect of the length $T$, the selection of different values of coefficients and variances has no particular influence on the performance here. No matter what value the correlation $a$ is chosen, close to boundary $a = 0.9$ or weak correlation $a = 0.1$, and no matter the noise
level is high $\sigma = 3$ or reduced to $\sigma = 0.1$, there is no significant evidence of difference in results.

To complete the simulation study, at the end we provide a table of simulations with long length but few individuals. We choose $N = 3$, $T = 1000$ to verify that under the AR(1) scenario, the longer the time series, the better the performance, for both pairwise likelihood and full likelihood, From Table 4.4 we can see that the results provided by the pairwise likelihood and the full likelihood are nearly identical with efficiency over 99%.

Now, we conclude that the two methods are large-sample, in the sense of long sequence, equivalent, and pairwise likelihood method loses very little to the full likelihood and is always highly efficient.
Table 4.2: AR(1) model, simulation results with $N = 1000$, $T = 100$

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$a$</th>
<th>Full</th>
<th>Pairwise</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{a}$</td>
<td>$\hat{\sigma}^2$</td>
<td>$\hat{a}$</td>
<td>$\hat{\sigma}^2$</td>
</tr>
<tr>
<td>0.1$^2$</td>
<td>0.55</td>
<td>0.539</td>
<td>0.010</td>
<td>0.539</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.084)</td>
<td>(0.001)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.099</td>
<td>0.010</td>
<td>0.099</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.099)</td>
<td>(0.001)</td>
<td>(0.099)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.884</td>
<td>0.010</td>
<td>0.883</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.044)</td>
<td>(0.001)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>1$^2$</td>
<td>0.55</td>
<td>0.542</td>
<td>0.977</td>
<td>0.541</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.084)</td>
<td>(0.139)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.101</td>
<td>0.974</td>
<td>0.101</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.099)</td>
<td>(0.138)</td>
<td>(0.099)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.885</td>
<td>0.980</td>
<td>0.884</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.044)</td>
<td>(0.139)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>3$^2$</td>
<td>0.55</td>
<td>0.537</td>
<td>8.684</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.084)</td>
<td>(1.234)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.099</td>
<td>8.808</td>
<td>0.099</td>
<td>8.899</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.100)</td>
<td>(1.252)</td>
<td>(0.100)</td>
</tr>
<tr>
<td>0.9</td>
<td>0.884</td>
<td>8.799</td>
<td>0.883</td>
<td>8.877</td>
</tr>
<tr>
<td></td>
<td>$se$</td>
<td>(0.044)</td>
<td>(1.251)</td>
<td>(0.045)</td>
</tr>
</tbody>
</table>

$^1$Estimates (Ests) of MLE and MPLE are listed for $a$ and $\sigma^2$, under Full and Pairwise respectively, so are the corresponding standard errors ($se$) in parentheses. $^2$Relative efficiency is given by the general efficiency $ARE_\theta$, and the efficiency in terms of each parameter $ARE_a$ and $ARE_{\sigma^2}$. $^3$Same layout applies to the next two tables.
### 4.5 Simulation Study

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$\alpha$</th>
<th></th>
<th></th>
<th></th>
<th>$\theta$</th>
<th>$\alpha$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha}$ ($\hat{\sigma}^2$)</td>
<td>$\hat{\sigma}^2$ ($\hat{\sigma}^2$)</td>
<td>$\hat{\sigma}^2$ ($\hat{\sigma}^2$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1²</td>
<td>0.55</td>
<td>0.515 (0.186)</td>
<td>0.009 (0.003)</td>
<td>0.909 (0.003)</td>
<td>0.938</td>
<td>0.972</td>
<td>0.907</td>
</tr>
<tr>
<td>0.1</td>
<td>0.55</td>
<td>0.091 (0.222)</td>
<td>0.009 (0.003)</td>
<td>0.091 (0.003)</td>
<td>0.944</td>
<td>0.994</td>
<td>0.897</td>
</tr>
<tr>
<td>0.9</td>
<td>0.55</td>
<td>0.838 (0.101)</td>
<td>0.009 (0.003)</td>
<td>0.838 (0.003)</td>
<td>0.891</td>
<td>0.885</td>
<td>0.901</td>
</tr>
<tr>
<td>1²</td>
<td>0.55</td>
<td>0.524 (0.184)</td>
<td>0.894 (0.291)</td>
<td>0.932 (0.307)</td>
<td>0.937</td>
<td>0.970</td>
<td>0.906</td>
</tr>
<tr>
<td>0.1</td>
<td>0.55</td>
<td>0.101 (0.221)</td>
<td>0.905 (0.294)</td>
<td>0.941 (0.311)</td>
<td>0.943</td>
<td>0.994</td>
<td>0.895</td>
</tr>
<tr>
<td>0.9</td>
<td>0.55</td>
<td>0.844 (0.098)</td>
<td>0.889 (0.290)</td>
<td>0.835 (0.306)</td>
<td>0.893</td>
<td>0.889</td>
<td>0.901</td>
</tr>
<tr>
<td>3²</td>
<td>0.55</td>
<td>0.510 (0.186)</td>
<td>7.899 (2.572)</td>
<td>8.229 (2.706)</td>
<td>0.938</td>
<td>0.971</td>
<td>0.907</td>
</tr>
<tr>
<td>0.1</td>
<td>0.55</td>
<td>0.084 (0.222)</td>
<td>7.872 (2.558)</td>
<td>8.260 (2.712)</td>
<td>0.942</td>
<td>0.993</td>
<td>0.894</td>
</tr>
<tr>
<td>0.9</td>
<td>0.55</td>
<td>0.839 (0.101)</td>
<td>8.130 (2.659)</td>
<td>8.519 (2.802)</td>
<td>0.894</td>
<td>0.889</td>
<td>0.902</td>
</tr>
</tbody>
</table>

Table 4.3: AR(1) model, simulation results with $N = 1000$, $T = 20$.  
Estimates (Ests) of MLE and MPLE are listed for $\alpha$ and $\sigma^2$, under Full and Pairwise respectively, so are the corresponding standard errors (se) in parentheses. Relative efficiency is given by the general efficiency $ARE_{\theta}$, and the efficiency in terms of each parameter $ARE_{\alpha}$ and $ARE_{\sigma^2}$. 

---

4Estimates (Ests) of MLE and MPLE are listed for $\alpha$ and $\sigma^2$, under Full and Pairwise respectively, so are the corresponding standard errors (se) in parentheses. 5Relative efficiency is given by the general efficiency $ARE_{\theta}$, and the efficiency in terms of each parameter $ARE_{\alpha}$ and $ARE_{\sigma^2}$. 

---
4.6 Discussion

The symmetric normal and the AR(1) models are curved exponential models. However, the AR(1) has asymptotically no loss of information, whereas this is not the case for the symmetric normal. The loss of efficiency of pairwise likelihood applied to the symmetric normal is shown in Cox and Reid (2004) with increasing dimension $q$; when $q = 10$ there is about 15% loss of efficiency. The good behavior of pairwise likelihood here can be explained...
by the facts that, first of all, as $T$ increases, the dependence of the time series decays exponentially to zero. Secondly, for large $T$, $S_1 \approx S_2$, the curved exponential model becomes asymptotically a full exponential model with minimum sufficient statistic $(S_1, S_{12})$, 
Chapter 5

Correlated binary data

This chapter investigates the properties of composite likelihood inference for correlated data with binary responses. Due to the complex structure of the dependence, the ordinary likelihood function is usually intractable. Moreover, since the likelihood method relies on the specification of the full joint distribution, it is sensitive to model misspecification.

The model used here for correlated binary data is the Bahadur representation. This is introduced in Section 1, and model set-up and notation are described in Section 2. In Section 3, we outline estimation based on both the pairwise likelihood function and the full likelihood function. Sections 4 and 5 describe two simulations and Section 6 provides results of the simulation study, reporting accuracy, asymptotic efficiency, and robustness. The pairwise likelihood approach is shown to provide accurate estimates, demonstrate high efficiency, and be less computationally intensive. Moreover, it is more robust to model misspecification.
5.1 Bahadur representation

The joint distribution of multivariate binary (MVB) variables can be specified using a representation first suggested by Bahadur (1961), and later by Cox (1972a). In the Bahadur representation, the joint probabilities of MVB are functions of marginal probabilities and second- and higher-order correlations.

Consider \( N \) independent individuals or experiment units and \( n_i \) observations for each individual. For \( i = 1, \cdots, N \), let \( Y_i = (Y_{i1}, \cdots, Y_{in_i})^T \) be the binary response measurements from the \( i \)th individual. For \( j = 1, \cdots, n_i \) and \( i \) in \( 1, \cdots, N \), let \( \mu_{ij} = P(y_{ij} = 1|x_i) \), or equivalently

\[
\mu_{ij} = \mathbb{E}(y_{ij}|x_i), \tag{5.1}
\]

given a covariate vector \( x_i = (x_{i1}, \cdots, x_{in_i})^T \). Next, setting

\[
z_{ij} = (y_{ij} - \mu_{ij})/\sqrt{\mu_{ij}(1 - \mu_{ij})}, \tag{5.2}
\]

define

\[
\rho_{ij;jk} = \mathbb{E}(z_{ij}z_{ik}), \quad j < k,
\]

\[
\rho_{ij;kl} = \mathbb{E}(z_{ij}z_{ik}z_{il}), \quad j < k < l,
\]

\[\vdots\]

\[
\rho_{i;12\cdots n_i} = \mathbb{E}(z_{i1}z_{i2}\cdots z_{in_i}),
\]
as the pairwise correlations, third-order correlations, ..., and the \( n_i \)th-order correlation. The correlation parameters, totally \( 2^n_i - n_i - 1 \) in number, together with the \( n_i \) marginal probabilities \( \mu_{ij} \)'s, determine the probability distribution of the multivariate binary distribution.

The Bahadur representation of the joint distribution of the \( Y_i \)'s is

\[
f(y_i) = \prod_{j=1}^{n_i} P(Y_{ij} = y_{ij}) \\
\times \left\{ 1 + \sum_{j<k} \rho_{i,jk} z_{ij} z_{ik} + \sum_{j<k<l} \rho_{i,jkl} z_{ij} z_{ik} z_{il} + \cdots + \rho_{i,12\ldots n_i} z_{i1} \cdots z_{in_i} \right\}. \tag{5.4}
\]

This second factor describes the effect of departures from independence. It is obtained by an expansion in terms of an orthonormal basis \( S \) relative to the joint probability distribution of the \( Y_i \) when all \( Y_i \)'s are independent.

The set

\[
S = \{1; z_{i1}, \cdots, z_{in_i}; z_{i1} z_{i2}, \cdots, z_{in_i-1} z_{in_i}; \cdots; z_{i1} z_{i2} \cdots z_{in_i}\}
\]

is a basis in the space of real-valued functions on \( Y_i \), and each function on \( Y_i \) admits one and only one form as a linear combinations of functions in \( S \). Therefore, the joint probability distribution of any subset of \( Y_i \) takes the same form as (5.4), i.e.

\[
f(y_{ir1}, \cdots, y_{irn}) = \prod_{j=r_1}^{r_n} P(y_{ij}) \\
\times \left\{ 1 + \sum_{j<k} \rho_{ir,jk} z_{irj} z_{irk} + \sum_{j<k<l} \rho_{ir,jkl} z_{irj} z_{irk} z_{irl} + \cdots + \rho_{ir,12\ldots n_i} z_{ir1} \cdots z_{irn} \right\}.
\]
5.2 Model set-up and marginal distributions

The Bahadur representation is particularly useful for our analysis, primarily because the representation specifies the joint distribution of MVB as a function of distributions of different orders. For example, \( f \) is of first order if all \( 2^n - n - 1 \) correlations are zero, i.e. \( y_i \)'s are independent and \( f \) is simplified to the first factor in (5.4),

\[
f(y_{ij}, \cdots, y_{in_i}) = \prod_{j=1}^{n_i} P(y_{ij}) = \prod_{j} \mu_{ij} (1 - \mu_{ij})^{(1-y_{ij})};
\]

and \( f \) is of second order if all correlations of order exceeding 2 are zeros,

\[
f(y_{ij}, \cdots, y_{in_i}) = \prod_{j=1}^{n_i} P(y_{ij}) \times \{1 + \sum_{j<k} \rho_{ijk} z_{ij} z_{ik}\};
\]

and so on.

5.2 Model set-up and marginal distributions

We assume independent clusters. Within clusters, the structure of dependence of the multivariate binary distribution is of great complexity. Because it is difficult to interpret measures of association beyond the order of two and pairwise associations are usually of most interest, we assume only second order dependence within clusters.

For simplicity, we assume equal cluster sizes, \( n_i = m \), and we assume that \( x_i \) is a \( m \times 1 \) vector of covariates \( x_{ij} \), where \( x_{ij} \) is scalar. The binary response of the \( j \)th observation in \( i \)th cluster is \( y_{ij}, i = 1, \cdots, N; j = 1, \cdots, m \). The dependence of univariate marginal
probabilities $\mu_{ij} = E(y_{ij}|x_{ij})$ on $x_{ij}$ is modeled through logit link function

$$\text{logit} \mu_{ij} = \beta_0 + \beta_1 x_{ij}. \quad (5.5)$$

The pairwise dependence can be measured by correlation coefficient, the odds ratio or
the relative risk. We choose odds ratio over correlations as odds ratios are considered less
constrained and more flexible than correlations. The compatibility of the second order asso-
ciation parameters and the first order mean vector is discussed in detail in Chaganty and Joe
(2004). The odds ratio for a pair of binary random variables $y_{ij}$ and $y_{ik}$ is defined as

$$\psi_{i,jk} = \frac{P(y_{ij} = 1, y_{ik} = 1|x_i) \cdot P(y_{ij} = 0, y_{ik} = 0|x_i)}{P(y_{ij} = 1, y_{ik} = 0|x_i) \cdot P(y_{ij} = 0, y_{ik} = 1|x_i)},$$

and its range is $[0, \infty)$. We use the log odds ratio $\phi$

$$\log \psi_{i,jk} = \phi,$$

and $\phi$ is defined over the entire real line.

The parameter of interest is $\theta = (\beta_1, \phi)^T$, where $\beta_1$ is the first order regression parameter
and $\phi$ is the second order association parameter characterizing the pairwise association within
clusters. With no loss of generality, $\beta_0$ is taken as known for the sake of calculation savings.
From the definition of the odds ratio, we have

$$\psi = \frac{\mu_{ijk}(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})},$$

where we have omitted the subscripts of $\psi$ since we assume a common (log) odds ratio, and using quadratic formula and taking only the positive root value, the bivariate probability, $\mu_{i,j,k} = P(y_{ij} = 1, y_{ik} = 1|x_i)$, is given in terms of $\psi$ by

$$\mu_{i,j,k} = \begin{cases} 
\frac{a_{i,j,k} - \sqrt{b_{i,j,k}}}{2(\psi - 1)} & \text{if } \psi \neq 1 \\
\mu_{ij}\mu_{ik} & \text{if } \psi = 1 
\end{cases} \tag{5.6}$$

where

$$a_{i,j,k} = 1 - (1 - \psi)(\mu_{ij} + \mu_{ik}),$$
$$b_{i,j,k} = a_{i,j,k}^2 - 4\psi(\psi - 1)\mu_{ij}\mu_{ik}.$$ 

Note that we assume common log odds ratio, but pairwise correlations $\rho_{i,j,k}$ between $y_{ij}$ and $y_{ik}$, i.e.

$$\rho_{i,j,k} = \frac{\mu_{ijk} - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}} \tag{5.7}$$

are still different: see Appendix E.
5.3 Estimation procedures

5.3.1 Full likelihood

Using the Bahadur representation (5.4) and omitting correlations of order higher than two, the joint probability density for $Y_i$ under the model assumptions is given by

$$f(y_{i1}, \cdots, y_{im}) = \prod_{j=1}^{m} P(y_i) \times \left\{1 + \sum_{j<k} \rho_{ij}z_{ij}z_{ik}\right\}.$$  (5.8)

Substituting the correlation and standardized variables, the likelihood function for $Y_i$ is

$$L(\theta; y_i) = \left(\prod_{j=1}^{m} \mu_{ij}(1 - \mu_{ij})^{1-y_{ij}}\right) \times \left\{1 + \sum_{j<k} \frac{(\mu_{ijk} - \mu_{ij}\mu_{ik})(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})}{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}\right\}.$$  (5.9)

The overall full log-likelihood function for $Y = (Y_1, \cdots, Y_N)$ is then

$$l(\theta) = \sum_{i=1}^{N} l_i(\theta) = \sum_{i=1}^{N} \log f(y_{i1}, \cdots, y_{im}).$$  (5.10)

The maximum likelihood estimator $\hat{\theta} = (\hat{\beta}_1, \hat{\phi})^T$ is the solution to the score function

$$l'(\theta) = \sum_{i=1}^{N} l_i'(\theta) = 0.$$  (5.11)

Optimization details are given in Appendix F.
The full likelihood function, as we assume, under the regularity conditions carries $I = J$.

Furthermore, $\hat{\theta}$ is asymptotically normal for large $N$

$$
(\hat{\theta} - \theta) \sim N(0, I^{-1}),
$$

and the Fisher information matrix $I$, to avoid deriving second derivatives, is estimated by

$$
\hat{I} = \sum_{i=1}^{N} l'_{i}l'^{T}_{i},
$$

where $l'_{i}$ is evaluated at $\hat{\theta}$.

### 5.3.2 Pairwise likelihood

The pairwise log-likelihood function for the random sample $Y$ is

$$
l_p = \sum_{i=1}^{N} l_{ip} = \sum_{i=1}^{N} \sum_{j<k} \log f(y_{ij}, y_{ik}),
$$

where $f(y_{ij}, y_{ik})$ is the bivariate density functions for within cluster pairs given by

$$
f(y_{ij}, y_{ik}) = \mu_{ij}^{y_{ij}y_{ik}} \cdot (\mu_{ij} - \mu_{i,j,k})^{y_{ij}(1-y_{ik})} \cdot (\mu_{ik} - \mu_{i,j,k})^{(1-y_{ij})y_{ik}}
$$

$$
\cdot (1 - \mu_{ij} - \mu_{ik} + \mu_{i,j,k})^{(1-y_{ij})(1-y_{ik})},
$$

(5.15)
and $l_{ip}$ is the pairwise log-likelihood function based on cluster $i$. We let $l_{ijk} = \log f(y_{ij}, y_{ik})$, representing the bivariate log-likelihood function based on observations of $j \neq k$ in cluster $i$.

Maximizing $l_p$ yields the maximum pairwise likelihood estimator $\tilde{\theta}_{PL} = (\tilde{\beta}_{1PL}, \tilde{\phi}_{PL})^T$, a solution to the pairwise score equations

$$
\begin{align*}
\sum_{i=1}^{N} \sum_{j<k}^{m} \frac{\partial l_{ij}}{\partial \beta_1} &= 0 \\
\sum_{i=1}^{N} \sum_{j<k}^{m} \frac{\partial l_{ij}}{\partial \phi} &= 0
\end{align*}
$$

The formulae of these derivatives are given in Appendix F.

As described in Chapter 2, the resulting estimator is asymptotically normal with mean $\theta$ and variance matrix $G^{-1}(\theta)$, i.e.

$$(\tilde{\theta}_{PL} - \theta) \sim N(0, G^{-1}(\theta)) \quad (5.17)$$

where $G.(\theta) = NG(\theta)$ and $G(\theta)$ is the Godambe information or sandwich information matrix given in Chapter 2,

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

where

$$
\begin{align*}
H.(\theta) &= E\left(-\frac{\partial^2 l_p(\theta)}{\partial \theta \partial \theta^T}\right), \\
J.(\theta) &= E\left(\frac{\partial l_p(\theta)}{\partial \theta} \cdot \frac{\partial l_p(\theta)}{\partial \theta^T}\right).
\end{align*}
$$
5.3. ESTIMATION PROCEDURES

It’s difficult to achieve a closed form for the Godambe information for the multivariate binary distribution, because the expressions for \( l''_{i,jk} \) are long and complicated. However, since \( l_{i,jk} \) is a valid log-likelihood function for pairs \((y_{ij}, y_{ik})\), we have \( E\{-l''_{i,jk}\} = E\{l'_{i,jk}l'^T_{i,jk}\} \), and the clusters are independent, so we have

\[
H = E(-l''_p) = E\left(-\sum_{i=1}^{N} \sum_{j<k}^m l''_{i,jk}\right) = E\left(\sum_{i=1}^{N} \sum_{j<k}^m l'_{i,jk}l'^T_{i,jk}\right), \quad (5.18)
\]

\[
J = E(l'_p l'^T_p) = E\left(\sum_{i=1}^{N} \sum_{j<k}^m l'_{i,jk}\right) = E\left(\sum_{i=1}^{N} l'_{i,jk}l'^T_{i,jk}\right), \quad (5.19)
\]

The expressions show that we need first derivatives only.

Hence, \( G(\tilde{\theta}_{PL}) \) can be estimated using

\[
\hat{H} = \sum_{i=1}^{N} \sum_{j<k}^m (l'_{i,jk}l'^T_{i,jk}), \quad (5.20)
\]

\[
\hat{J} = \sum_{i=1}^{N} l'_{i,p}l'^T_{i,p} = \sum_{i=1}^{N} \left(\sum_{j<k}^m l'_{i,jk}\right) \left(\sum_{j<k}^m l'_{i,jk}\right)^T, \quad (5.21)
\]

where \( l'_{i,jk} \) are evaluated at \( \tilde{\theta}_{PL} \).

5.3.3 Independence case

The simplest MVB distribution assumes that \( y_{ij} \)'s are independent. Some analytical results can be derived straightforwardly. First, the joint pdf of \( Y_i \) is simplified to its first-order
distribution

\[ f(y_{ij}, \cdots, y_{in_i}) = \prod_{j=1}^{n_i} P(y_{ij}) = \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}}(1 - \mu_{ij})^{1-y_{ij}}, \]

and the full log-likelihood function for cluster \(i\) is given by

\[ l_i(\theta) = \sum_{j=1}^{n_i} \{y_{ij} \log \mu_{ij} + (1 - y_{ij}) \log(1 - \mu_{ij})\}. \]

Secondly, the pairwise log-likelihood function for cluster \(i\) is

\[ l_{ip}(\theta) = \sum_{j<k}^{n_i} \log \{P(y_{ij})P(y_{ik})\} = \sum_{j<k}^{n_i} \log \{\mu_{ij}^{y_{ij}}(1 - \mu_{ij})^{1-y_{ij}}\mu_{ik}^{y_{ik}}(1 - \mu_{ik})^{1-y_{ik}}\} = (n_i - 1) \sum_{j=1}^{n_i} \{y_{ij} \log \mu_{ij} + (1 - y_{ij}) \log(1 - \mu_{ij})\} \]

Hence, under independence within clusters we have

\[ l_{ip}(\theta) = (n_i - 1)l_i(\theta) \]

(5.22)

As a result, we have immediately that pairwise likelihood inference is fully efficient relative to full likelihood inference. In fact, when all \(y_{ij}\)'s are independent, the MVB distribution is fully determined by its marginal means \(\mu_{ij}\)'s, and reduced to a first order distribution such that \(l_i(\theta) \equiv l_1(\theta)\), and \(l_1(\theta)\) is called the independence log-likelihood.
5.4 Simulated data I – True model

To carry out the simulation study, we aim at generating $N$ clusters of multivariate binary variables of cluster size $m = 4$ from the multivariate logit model described in Section 2. We provide two simulation methods.

The first is called the inversion method in Devroye (1986), when an explicit specification of the joint distribution of $Y$ is feasible. Inspired by Preisser et al. (2002), the task is accomplished by a multinomial random number generator which produces the number of times that each $y_{ij}$ occurs, denoted by $n_j$. The simulation process is implemented in 3 steps:

i) Define $Y_i[.]$, an array of $2^4$ 4-dimensional binary vectors with the later components changing fastest. Use (5.8) to calculate the $2^4$ probabilities.

ii) Generate 1000 groups of multinomial random numbers $(n_1, n_2, \cdots, n_{16})$ with the total number of objects in each group which are assigned to 16 probabilities equal to $N(=100, 30)$, i.e. $n_1 + \cdots + n_{16} = N$.

iii) Transform $n_i$ to $Y_i[.]$. Determine the times of each $Y_i[.]$ using the generated multinomial numbers $(n_1, n_2, \cdots, n_{16})$. Return $Y$, 1000 random samples with each $N \times 4$.

Table 5.1 shows one example of the simulated 4-vector binary data with $N = 100$.

When the $2^m$ probabilities $\mu_{ij}$’s can be computed explicitly, this inversion method can generate random values from arbitrary distributions, such as the distribution of the true
Distribution of four binary variates

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$n_i$</th>
<th>$f(y_{i1}, \ldots, y_{i4})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1</td>
<td>0</td>
<td>0.006897319</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>1</td>
<td>0.010241185</td>
</tr>
<tr>
<td>1 1 0 1</td>
<td>1</td>
<td>0.010241185</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>2</td>
<td>0.016491342</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>5</td>
<td>0.021001699</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>4</td>
<td>0.034022752</td>
</tr>
<tr>
<td>1 0 0 1</td>
<td>3</td>
<td>0.034022752</td>
</tr>
<tr>
<td>1 0 0 0</td>
<td>7</td>
<td>0.064897878</td>
</tr>
<tr>
<td>0 1 1 1</td>
<td>2</td>
<td>0.021001699</td>
</tr>
<tr>
<td>0 1 1 0</td>
<td>2</td>
<td>0.034022752</td>
</tr>
<tr>
<td>0 1 0 1</td>
<td>1</td>
<td>0.034022752</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>10</td>
<td>0.064897878</td>
</tr>
<tr>
<td>0 0 1 1</td>
<td>9</td>
<td>0.070244122</td>
</tr>
<tr>
<td>0 0 1 0</td>
<td>11</td>
<td>0.134380701</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>18</td>
<td>0.134380701</td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>24</td>
<td>0.309233285</td>
</tr>
</tbody>
</table>

| Total | 100 | 1 |

Table 5.1: An example of simulated binary random sample with $x_i = (0, 0, 1, 1)$, $\beta_0 = -1.4$, and $\theta = (\beta_1, \phi) = (0.7, \log(1.2))$

model specified by (5.8). We call the model simulated by the inversion method the true model.

### 5.5 Simulated data II – Working model

With increasing cluster size, the expression for the joint density of the MVB distribution becomes more intractable. The problem of simulating binary data with specified lower dimension structure is of more interest. Various authors have put effort on the topic. Appealing methods would allow for unrestricted marginal probabilities and pairwise correlations, avoid
computing the joint distribution or solving nonlinear equations, and would be able to handle large \( m \).

Emrich and Piedmonte (1991) described a method based on the existence of a multivariate normal distribution. This discretised multivariate normal method is easy to interpret and implement. A minor drawback of the method is that it requires numerical integration. Lee (1993) presented two alternative methods, based on linear programming and Archimedean copulas, for generating binary data with arbitrary marginal distributions and consistent odds ratios. Lee’s approaches involve solving a large number of nonlinear equations. Gange (1995) proposed an iterative proportional fitting algorithm that allows unequal means and general correlations, but the algorithm may be discouraging in the sense that it requires a second iterative fitting procedures. Park et al. (1996) developed a method for generating non-negatively correlated binary variables by using correlated Poisson variables. But the method cannot be extended to negative correlations. Lunn and Davies (1998) introduced a simple and efficient method based on a finite mixture of Bernoulli variables. However, this method is only suitable for generating binary variables with exchangeable correlations, or correlation structures which are autoregressive or stationary \( M \)-dependent. Recently, Kang and Jung (2001) provided a method based on specifying the joint distribution. Oman and Zucker (2001) proposed a method similar to Lunn and Davies (1998) with constraints to only non-negative correlations. So as we see, the proposed approaches in Park et al. (1996) and Lunn and Davies (1998), as well as Oman and Zucker (2001), are limited to certain correlation ranges. More recently, Qaqish (2003) presented a method of generating a conditional linear family
of multivariate binary distributions with specified marginal means and correlations. The methods of Emrich and Piedmonte (1991) and Qaqish (2003) seem to be the most practically applicable methods for generating correlated binary data of large cluster size with unconstrained marginal probabilities and correlation structures.

For ease of implementation, we use the discretised normal method in Emrich and Piedmonte (1991) as the second method to generate data, using specified lower-dimensional distributions.

5.5.1 The algorithm

Let \( Z_i = (Z_{i1}, \ldots, Z_{im})^T \) be a \( m \)-dimensional multivariate normal random variable with mean zero, variance one, and correlation matrix \( \{ \delta_{i,jk} \} \). Set \( y_{ij} = 1 \) if \( Z_{ij} > z(\mu_{ij}) \), and \( y_{ij} = 0 \) otherwise, i.e. \( y_{ij} = 1 \Leftrightarrow Z_{ij} > z(\mu_{ij}) \). Therefore, we have

\[
E(y_{ij}) = P(y_{ij} = 1) = P(Z_{ij} > z(\mu_{ij})) = \mu_{ij},
\]

where \( \mu_{ij} \) is calculated by (5.5), and \( z(\mu_{ij}) \) is the \( \mu_{ij} \)th quantile of the standard normal distribution.

The joint probability of \( y_{ij} \) and \( y_{ik} \) is given by

\[
\mu_{ijk} = P(y_{ij} = 1, y_{ik} = 1) = \mu_{ij}\mu_{ik} + \rho_{i,jk}\sqrt{\nu_{ij}\nu_{jk}}, \tag{5.23}
\]
where \( \nu_{ij} = \mu_{ij}(1 - \mu_{ij}) \) and \( \nu_{ik} = \mu_{ik}(1 - \mu_{ik}) \), and \( \rho_{i;jk} \) is given by (5.7). Since \( y_{ij} \) and \( y_{ik} \) are converted from the normal random variables \( Z_{ij} \) and \( Z_{ik} \), it yields

\[
\mu_{i;jk} = P(Z_{ij} > z(\mu_{ij}), Z_{ik} > z(\mu_{ik})) = \Phi^*(z(\mu_{ij}), z(\mu_{ik}); \delta_{i;jk})
\] (5.24)

and

\[
\Phi^*(x, y, \rho) = \int_x^\infty \int_y^\infty \varphi(s, t; \delta) dt ds
\]

where \( \varphi \) is the probability density function of bivariate normal distribution mean zero, variance one, and correlation coefficient \( \delta \). A simple case is median dichotomy; i.e. \( z(\mu_{ij}) = 0 \), an example given in Cox and Reid (2004). Under such circumstance, an analytical result for \( \Phi^* \) is given by

\[
\Phi^*(0, 0, \rho) = \frac{\sin^{-1}(\rho)}{2\pi} + \frac{1}{4}.
\]

Otherwise, \( \Phi^* \) needs to be determined numerically.

Another step is to check the compatibility of the calculated marginal probabilities \( \mu_{ij} \) and second order probabilities \( \mu_{i;jk} \), or correlations \( \rho_{i;jk} \), i.e.

\[
\begin{aligned}
0 \leq \mu_{ij} &\leq 1 & j = 1, \ldots, 4 \\
\max(0, \mu_{ij} + \mu_{ik} - 1) \leq \mu_{i;jk} &\leq \min(\mu_{ij}, \mu_{ik}) & j \neq k \\
\mu_{ij} + \mu_{ik} + \mu_{il} - \mu_{i;jk} - \mu_{i;jl} - \mu_{i;lk} &\leq 1 & j \neq k, k \neq l, l \neq j.
\end{aligned}
\] (5.25)

Chaganty and Joe (2004) provides detailed compatibility conditions for lower dimension multivariate binary distributions. The simulation process of the discretised normal method is
summarized as below

i) Calculate the 1st- and 2nd-order distributions of the multivariate binary distribution by (5.5), (5.6), or (5.23). Check the compatibility with (5.25).

ii) Set \[ z(\mu_{ij}) = \Phi^{-1}(\mu_{ij}) \] where \( \Phi \) is the standard normal distribution function.

iii) Solve (5.24) numerically to obtain the correlation matrix \( \{(\delta_{i,jk})\} \) for MVN.

iv) Generate a sample of 4-vector MVN \( Z_{ij} \) with mean \( z(\mu_{ij}) \) variance one and correlation matrix \( \{(\delta_{i,jk})\} \).

v) Discretize MVN to MVB, i.e. \( y_{ij} = 1 \Leftrightarrow Z_{ij} > z(\mu_{ij}) \).

The discretised normal method is easy to implement and places no constraints on cluster size. It can generate binary data with large cluster size fast and efficiently. However, there are some concerns in the algorithm.

First, Emrich and Piedmonte (1991) and Chaganty and Joe (2004) noted that the conditions satisfied by the algorithm for compatibility are only sufficient, and it’s possible that the simulation method may fail to produce a binary distribution even if one exists.

We are more concerned that when data is generated by dichotomizing the MVN distribution in this way, it is not guaranteed to follow the true model (5.8) given by the 2nd order Bahadur representation. The specification of the MVN distribution does not ensure that correlations of the order three and higher are zero. In other words, the discretised MVN method generates MVB of satisfactory first and second order marginal distributions, but not
from the assumed model (5.8), but rather from a model with implicit dependence of higher order, i.e.

\[ f(y_1, \cdots, y_m) = \prod_{j=1}^{m} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}} \]

\[ \times \left\{ 1 + \sum_{j<k} \rho_{i,jk} z_{ij} z_{ik} + \sum_{j<k<l} r_{i,jkl} z_{ij} z_{ik} z_{il} + \cdots + r_{i,12\cdots m} z_{ij} \cdots z_{im} \right\}. \]

We call the simulation model the working model.

In contrast, the inversion method gives simulated data from the true model (5.8), although it can not handle large cluster sizes. However, the working model provides us a chance to consider robustness when higher order distributions are misspecified.

5.6 Simulation study

5.6.1 Simulation design

The simulation study is conducted under different conditions pertaining to the correlation structure and data simulating mechanism. Correlated binary data are generated by the inversion method and the discretised normal method. Each of the two methods was used to simulate data with three different correlation structures: representing almost no correlation, weak correlation and strong correlation, respectively.

We follow closely the simulations presented in Kuk and Nott (2000), and throughout we
fix $\beta_0 = -1.4$, $\beta_1 = 0.7$ and covariates $x_i = (0, 0, 1, 1)$, giving the marginal probabilities $
abla_{i1} = \nabla_{i2} = 0.1978$ and $
abla_{i3} = \nabla_{i4} = 0.3318$. Three correlation structures selected are determined using a common log odds ratio $\phi$: almost uncorrelated $\phi = \log 1.2$, weak negative $\phi = \log 0.8$, and medium positive $\phi = \log 6$. As shown in Appendix E, correlation is an increasing function of the odds ratio $\psi$. Table 5.2 displays the correlations corresponding to each log odds ratio selected.

<table>
<thead>
<tr>
<th>log odds ratio</th>
<th>correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>$\rho_{00}$</td>
</tr>
<tr>
<td>log 1.2</td>
<td>0.0300</td>
</tr>
<tr>
<td>log 0.8</td>
<td>-0.0339</td>
</tr>
<tr>
<td>log 6</td>
<td>0.3427</td>
</tr>
</tbody>
</table>

Table 5.2: Correlations corresponding to different log odds ratios

For each of the 12 ways that data are generated (two data simulation models $\times$ three correlation structures $\times$ two numbers of clusters $N = 100, 30$), we compare pairwise and full likelihood inference using 1000 simulated observations.

The maximum pairwise likelihood estimator $\tilde{\theta}_{PL}$ and the ordinary maximum full likelihood estimator $\hat{\theta}_F$ are compared, and for each of them, we evaluate the accuracy of the variance estimates, observed coverage of a nominal 95 percent confidence interval, and the asymptotic relative efficiency. More specifically, let $\tilde{\theta}_t/\hat{\theta}_t$ represent the estimate of $\theta = (\beta_1, \phi)^T$, for pairwise or ordinary estimators, from the $t$th simulated replicate. The Monte Carlo standard errors ($MC\ se$) for the pairwise estimates are evaluated by the average values over all simulations of the 1000 $\tilde{\theta}_t/\hat{\theta}_t$ using the estimated Godambe matrix, while those for the full likelihood method use the counterparts of the estimated Fisher information matrix. The
empirical standard deviations ($EM \ sd$) are the sample standard deviations of 1000 $\tilde{\theta}_t/\hat{\theta}_t$'s. Coverage rate is calculated as the proportion of normal 95% confidence intervals that contain $\theta = (\beta_1, \phi)$ for $\tilde{\theta}_t/\hat{\theta}_t$ using Monte Carlo standard errors. The asymptotic relative efficiency of $\tilde{\theta}_{PL}$ to $\hat{\theta}_F$ is evaluated in different ways according to correlation structures. We report both Monte Carlo and empirical AREs when applicable. The empirical ARE is more reasonable when the observations are mildly correlated and MLE and Fisher information are not highly accurate estimates. Moreover, when neither of the above AREs reflects the reality appropriately due to relatively poor estimation, alternatively we choose the ratio of the mean squared errors of the estimators as the last choice.

Under the working model, the full likelihood function (5.8) is a second order approximation to the actual distribution, and sensitive to large correlations. Moreover, this approximation may not be a probability distribution function at all, i.e. it could fail to be non-negative for some points $y_{ij}$'s. It occurs in the estimation procedures that the full likelihood method fails occasionally; the Newton-Raphson iterations fall out of the narrow compatible parameter definition region for certain simulated sample $Y$. We introduce $N_{good}$ to count the number of successful runs for full likelihoods. Therefore, for a distribution involving certain orders of structure of dependence complicate and unknown, such as the working model, using the full likelihood method can be risky. The accomplishment of the full likelihood method loses with increasing correlations. On the other hand, the pairwise likelihood is always a valid bivariate likelihood function.

In the following, we summarize what’s reported in the simulation. The estimates are the
averages of the 1000 parameter estimates $\tilde{\theta}_t/\hat{\theta}_t$:

$$\tilde{\theta}_{PL} = \frac{1}{1000} \sum_{t=1}^{1000} \tilde{\theta}_{PLt} \text{ and } \hat{\theta}_F = \frac{1}{N_{\text{good}}} \sum_{t=1}^{N_{\text{good}}} \hat{\theta}_{Ft}.$$ 

The estimated Godambe matrix $\hat{G}_PL^{-1}$ is the average value over all the estimated matrices, and each inverse of the Godambe information matrix from the $t$th replicate is

$$\hat{G}_{PLt}^{-1}(\hat{\theta}_{PL}) = \hat{H}^{-1}(\hat{\theta}_{PL})\hat{J}(\hat{\theta}_{PL})\hat{H}^{-1}(\hat{\theta}_{PL}),$$

where $\hat{J}$ and $\hat{H}$ are given by (5.20) and (5.21). The standard errors for the pairwise estimators are given by

$$\hat{se}(\hat{\theta}_{PL}) = \left( \text{Diag}\{\hat{G}_{PL}^{-1}\} \right)^{1/2}.$$ (5.26)

Concerning the full likelihood, the inverse of the Fisher information matrix $\hat{I}_F^{-1}$ is given by average over the estimated matrices for each $t$th replicate, calculated by (5.13), and the standard errors for full likelihood estimators are

$$\hat{se}(\hat{\theta}_F) = \left( \text{Diag}\{\hat{I}_F^{-1}\} \right)^{1/2}. $$ (5.27)

Sample variances and sample standard deviations are:

$$s^2_{PL} = \frac{1}{N-1} \sum_{t=1}^{1000} (\tilde{\theta}_{PLt} - \tilde{\theta}_{PL})^2 \text{ and } s^2_F = \frac{1}{N_{\text{good}}-1} \sum_{t=1}^{N_{\text{good}}} (\hat{\theta}_{Ft} - \tilde{\theta}_{F})^2,$$

and $\text{sd}(\hat{\theta}_{PL}) = (s^2_{PL})^{1/2}$ and $\text{sd}(\hat{\theta}_F) = (s^2_F)^{1/2}$. Coverage rate is calculated as the per cent of
95\% confidence intervals using the standard error formulae derived above

\[
\tilde{se}(\bar{\theta}_{PLt}) = \left( \text{Diag}\{\hat{G}^{-1}_{PLt}(\bar{\theta}_{PLt})\} \right)^{1/2} \quad \text{and} \quad \tilde{se}(\bar{\theta}_{Ft}) = \left( \text{Diag}\{\hat{I}^{-1}_{Ft}(\bar{\theta}_{Ft})\} \right)^{1/2}.
\]

Asymptotic relative efficiency is illustrated in different ways.

**Monte Carlo ARE in terms of estimated variance matrix:**

\[
\text{ARE}_{\theta}(\bar{\theta}_{PL}, \bar{\theta}_{F}) = \left\{ \frac{\left| \hat{I}^{-1}_{F} \right|}{\left| \hat{G}^{-1}_{PL} \right|} \right\}^{1/2}.
\]

Or, it can be more reasonable to report the ARE for each component of the parameter as

\[
\text{ARE}_{mc}(\bar{\beta}_{PL}, \bar{\beta}_{F}) = \frac{\hat{I}^{-1}_{F_{\beta\beta}}}{\hat{G}^{-1}_{PL_{\beta\beta}}} \quad \text{and} \quad \text{ARE}_{mc}(\bar{\phi}_{PL}, \bar{\phi}_{F}) = \frac{\hat{I}^{-1}_{F_{\phi\phi}}}{\hat{G}^{-1}_{PL_{\phi\phi}}}. \tag{5.28}
\]

**Empirical ARE in terms of sample variance:** Since

\[
\frac{1}{N-1} \sum_{t=1}^{1000} (\bar{\theta}_{PLt} - \bar{\theta}_{PL})^2 \overset{p}{\rightarrow} G^{-1} \quad \text{and} \quad \frac{1}{N_{\text{good}}-1} \sum_{t=1}^{N_{\text{good}}} (\bar{\theta}_{Ft} - \bar{\theta}_{F})^2 \overset{p}{\rightarrow} I^{-1},
\]

we can also use sample variance to evaluate the relative efficiency. We especially rely more on this empirical ARE when the observations are mildly correlated for the reasons that full likelihood estimates are far from the true values and the estimate of the inverse Fisher information matrix is not reliable. The empirical ARE is given by

\[
\text{ARE}_{\theta}(\bar{\theta}_{PL}, \bar{\theta}_{F}) = \left\{ \frac{\text{s.var}(\bar{\theta}_{F})}{\text{s.var}(\bar{\theta}_{PL})} \right\}^{1/2},
\]
where the sample variance-covariance matrix is computed as
\[
s_{\text{var}}(\hat{\theta}) = s_{\text{var}} \left( \begin{pmatrix} \hat{\beta}_1 \\ \hat{\phi} \end{pmatrix} \right) = \frac{1}{N-1} \left( \frac{\sum (\hat{\beta}_{1t} - \bar{\hat{\beta}}_{1t})^2}{\sum (\hat{\beta}_{1t} - \bar{\hat{\beta}}_{1t})(\hat{\phi}_t - \bar{\hat{\phi}}_t)} \right).
\]

We need to distinguish \( N_{\text{good}} \) from \( N \) for the full likelihood method. What we report are the empirical AREs in terms of each component

\[
ARE_{\text{em}}(\tilde{\beta}_{PL}, \hat{\beta}_F) = \frac{s_{\text{var}}(\hat{\beta}_F)}{s_{\text{var}}(\tilde{\beta}_{PL})} \quad \text{and} \quad ARE_{\text{em}}(\tilde{\phi}_{PL}, \hat{\phi}_F) = \frac{s_{\text{var}}(\hat{\phi}_F)}{s_{\text{var}}(\tilde{\phi}_{PL})}. \tag{5.29}
\]

**ARE in terms of MSE:** Among all the scenarios, the full likelihood function is the least accurate approximation for the working model with mild correlations. Under the circumstances, the ordinary MLE of \( \phi \) is rather biased, which yields incomparable Monte Carlo or empirical variance estimates. Another choice to examine the relative efficiency is using the mean squared error:

\[
ARE_{\text{mse}}(\tilde{\beta}_{PL}, \hat{\beta}_F) = \frac{\text{MSE}(\hat{\beta}_F)}{\text{MSE}(\tilde{\beta}_{PL})} \quad \text{and} \quad ARE_{\text{mse}}(\tilde{\phi}_{PL}, \hat{\phi}_F) = \frac{\text{MSE}(\hat{\phi}_F)}{\text{MSE}(\tilde{\phi}_{PL})}. \tag{5.30}
\]

### 5.6.2 Simulation results

The simulation results are given in Tables 5.3 - 5.6. Tables 5.3 and 5.5 report results for \( \beta_1 \) based on different models and likelihoods, for \( N \) equal to 100 and 30, respectively, and likewise Tables 5.4 and 5.6 for \( \phi \). Each table includes estimators with Monte Carlo standard
errors and empirical standard deviations, coverage rates, and asymptotic relative efficiency calculated by Monte Carlo variances, empirical variances, or MSE as applicable. Results for the bias of the estimators and significance test for the regression parameter $\beta_1$ are not shown in the tables since the results are consistent with the estimators or coverage rates.

Some observations can be made: (i) more clusters gives better estimates; (ii) pairwise estimators are generally good; (iii) the pairwise likelihood approach is comparable to the full likelihood approach, and in fact, it tends to outperform the full likelihood approach.

Tables 5.3 and 5.5 for $\beta_1$ show that the pairwise likelihood method provides highly accurate estimates for large $N = 100$, and sufficiently good estimates for small $N = 30$, regardless of the type of models or correlation structures. The full likelihood method also provides generally good results, except when the data are strongly correlated. The asymptotic relative efficiencies of the pairwise likelihood method to the full likelihood method are mostly 1. One exception occurs for true model with $\phi = \log 6$ where the full likelihood method fails to be a valid probability function more often and gives inaccurate variance estimates and small relative efficiency. The coverage rates of the pairwise likelihood estimation are generally higher than those of full likelihood estimation, and are close to nominal 95 per cent level. For the estimation of the first order parameter $\beta_1$, the pairwise likelihood approach is accurate and efficient.

The results for $\phi$ are contained in Tables 5.4 and 5.6, and are in general agreement with the results for $\beta_1$. Further, full likelihood estimation for the association parameter is observed to be more sensitive to model misspecification. For example, for the working model with
medium correlations, the MLEs are not as good as the maximum pairwise estimators in Table 5.4 for $N = 100$, and worse for $N = 30$. The efficiency of the pairwise likelihood approach can be more than twice as large as the full likelihood approach, underscoring the poor properties of the latter method in the situations of model misspecification. Full likelihood estimation results in undercoverage, while pairwise likelihood estimation has coverages near the nominal 92 per cent level. These facts lead us to conclude that for the second order parameter $\phi$ pairwise likelihood approach outperforms the full likelihood approach.

The simulation results reveal that for multivariate binary data the pairwise likelihood method provides more accurate estimators, higher efficiency, and is more robust to model misspecification. Besides, the pairwise likelihood function is simple and neat, and less computationally intensive. It is a method that is especially preferable to the full likelihood method when model misspecification is inherent.
## 5.6. Simulation Study

| $\phi$ | **True Model** | | **Working Model** | |
| --- | --- | | --- | |
|  | pairwise | full | $N_{good}$ | pairwise | full | $N_{good}$ |
| log 1.2 | Estimates | 0.698 | 0.698 | 999 | 0.701 | 0.703 | 1000 |
| |  | (0.182) | | | | | |
| | MC se | 0.153 | 0.155 | | 0.153 | 0.155 | |
| | EM sd | 0.155 | 0.155 | | 0.154 | 0.154 | |
| | Coverage rate | 956 | 952 | | 944 | 948 | |
| | AREmc | 1.024 | 1.025 | |
| | AREem | 0.995 | 0.991 | |
| log 0.8 | Estimates | 0.699 | 0.694 | 982 | 0.699 | 0.695 | 985 |
| | (−0.223) | | | | | | |
| | MC se | 0.146 | 0.146 | | 0.146 | 0.147 | |
| | EM sd | 0.148 | 0.149 | | 0.149 | 0.150 | |
| | Coverage rate | 948 | 928 | | 953 | 937 | |
| | AREmc | 1.003 | 1.006 | |
| | AREem | 1.012 | 1.018 | |
| log 6 | Estimates | 0.701 | 0.722 | 870 | 0.696 | 0.756 | 998 |
| (1.792) | | | | | | | |
| | MC se | 0.159 | 0.134 | | 0.160 | 0.148 | |
| | EM sd | 0.156 | 0.131 | | 0.156 | 0.153 | |
| | Coverage rate | 968 | 818 | | 956 | 918 | |
| | AREmc | n/a | n/a | |
| | AREem | 0.711 | n/a | |
| | AREmse | - | 1.082 | |

Table 5.3: Simulation results for $\beta_1$, for $N = 100$, $m = 4$, $\beta_1 = 0.7$.

1. n/a indicates that MC se or EM sd is not applicable nor is their ratio.
2. Efficiency of working model with medium correlations is calculated using mse.
3. AREmc or AREem is preferred than AREmse in general.
<table>
<thead>
<tr>
<th>$\phi$</th>
<th>True Model</th>
<th></th>
<th>Working Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pairwise</td>
<td>full</td>
<td>$N_{\text{good}}$</td>
</tr>
<tr>
<td>log 1.2</td>
<td>$\text{Estimates}$</td>
<td>0.180</td>
<td>0.180</td>
<td>999</td>
</tr>
<tr>
<td>(0.182)</td>
<td>$MC$ se</td>
<td>0.223</td>
<td>0.228</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>$EM$ sd</td>
<td>0.235</td>
<td>0.242</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>929</td>
<td>934</td>
<td>932</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td></td>
<td>1.045</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREem</td>
<td></td>
<td>1.060</td>
<td></td>
</tr>
<tr>
<td>log 0.8</td>
<td>$\text{Estimates}$</td>
<td>-0.225</td>
<td>-0.279</td>
<td>982</td>
</tr>
<tr>
<td>(-0.223)</td>
<td>$MC$ se</td>
<td>0.199</td>
<td>0.242</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>$EM$ sd</td>
<td>0.210</td>
<td>0.280</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>937</td>
<td>885</td>
<td>939</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td></td>
<td>1.486</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREem</td>
<td></td>
<td>1.775</td>
<td></td>
</tr>
<tr>
<td>log 6</td>
<td>$\text{Estimates}$</td>
<td>1.792</td>
<td>1.897</td>
<td>870</td>
</tr>
<tr>
<td>(1.792)</td>
<td>$MC$ se</td>
<td>0.275</td>
<td>0.214</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>$EM$ sd</td>
<td>0.276</td>
<td>0.351</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>946</td>
<td>736</td>
<td>952</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td></td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREem</td>
<td></td>
<td>1.621</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREmse</td>
<td></td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Simulation results for $\phi$, for $N = 100$, $m = 4$, $\beta_1 = 0.7$. 
Table 5.5: Simulation results for $\beta_1$, for $N = 30$, $m = 4$, $\beta_1 = 0.7$.  

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>True Model</th>
<th></th>
<th>Working Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pairwise</td>
<td>full</td>
<td>$N_{good}$</td>
<td>pairwise</td>
<td>full</td>
</tr>
<tr>
<td>log 1.2</td>
<td>Estimates</td>
<td>0.717</td>
<td>0.714</td>
<td>998</td>
</tr>
<tr>
<td>(0.182)</td>
<td>MC se</td>
<td>0.277</td>
<td>0.294</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>EM sd</td>
<td>0.285</td>
<td>0.286</td>
<td>0.305</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>952</td>
<td>945</td>
<td>933</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td>1.126</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>log 0.8</td>
<td>Estimates</td>
<td>0.716</td>
<td>0.690</td>
<td>987</td>
</tr>
<tr>
<td>(-0.223)</td>
<td>MC se</td>
<td>0.265</td>
<td>0.287</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>EM sd</td>
<td>0.274</td>
<td>0.292</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>942</td>
<td>925</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td></td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREem</td>
<td>1.132</td>
<td></td>
<td>1.127</td>
</tr>
<tr>
<td>log 6</td>
<td>Estimates</td>
<td>0.692</td>
<td>0.719</td>
<td>890</td>
</tr>
<tr>
<td>(1.792)</td>
<td>MC se</td>
<td>0.290</td>
<td>0.255</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>EM sd</td>
<td>0.310</td>
<td>0.256</td>
<td>0.304</td>
</tr>
<tr>
<td></td>
<td>Coverage rate</td>
<td>937</td>
<td>839</td>
<td>946</td>
</tr>
<tr>
<td></td>
<td>AREmc</td>
<td></td>
<td>0.770</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREem</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AREmse</td>
<td></td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### Table 5.6: Simulation results for $\phi$, for $N = 30$, $m = 4$, $\beta_1 = 0.7$.

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>True Model</th>
<th></th>
<th>Working Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates</td>
<td>MC se</td>
<td>EM sd</td>
<td>Coverage rate</td>
</tr>
<tr>
<td>log 1.2</td>
<td>0.151</td>
<td>0.401</td>
<td>0.427</td>
<td>920</td>
</tr>
<tr>
<td>(0.182)</td>
<td>0.109</td>
<td>0.558</td>
<td>0.526</td>
<td>948</td>
</tr>
<tr>
<td>log 0.8</td>
<td>-0.260</td>
<td>0.359</td>
<td>0.383</td>
<td>914</td>
</tr>
<tr>
<td>(-0.223)</td>
<td>-0.481</td>
<td>2.952</td>
<td>0.779</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>987</td>
<td>0.365</td>
<td>0.388</td>
<td>914</td>
</tr>
<tr>
<td>log 6</td>
<td>1.824</td>
<td>0.511</td>
<td>0.553</td>
<td>922</td>
</tr>
<tr>
<td>(1.792)</td>
<td>2.030</td>
<td>1.014</td>
<td>0.669</td>
<td>838</td>
</tr>
<tr>
<td></td>
<td>890</td>
<td>0.593</td>
<td>0.643</td>
<td>917</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>pairwise</th>
<th>full</th>
<th>$N_{good}$</th>
<th>pairwise</th>
<th>full</th>
<th>$N_{good}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates</td>
<td>0.152</td>
<td>0.098</td>
<td>996</td>
<td>0.152</td>
<td>0.098</td>
<td>996</td>
</tr>
<tr>
<td>MC se</td>
<td>0.408</td>
<td>141.3</td>
<td></td>
<td>0.446</td>
<td>0.653</td>
<td></td>
</tr>
<tr>
<td>EM sd</td>
<td>0.427</td>
<td>0.526</td>
<td></td>
<td>0.446</td>
<td>0.653</td>
<td></td>
</tr>
<tr>
<td>Coverage rate</td>
<td>920</td>
<td>948</td>
<td></td>
<td>907</td>
<td>949</td>
<td></td>
</tr>
<tr>
<td>AREmc</td>
<td>1.931</td>
<td></td>
<td></td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREem</td>
<td>1.514</td>
<td></td>
<td></td>
<td>2.142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage rate</td>
<td>914</td>
<td>940</td>
<td></td>
<td>903</td>
<td>953</td>
<td></td>
</tr>
<tr>
<td>AREmc</td>
<td>n/a</td>
<td></td>
<td></td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREem</td>
<td>4.134</td>
<td></td>
<td></td>
<td>4.842</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage rate</td>
<td>922</td>
<td>838</td>
<td></td>
<td>917</td>
<td>804</td>
<td></td>
</tr>
<tr>
<td>AREmc</td>
<td>3.932</td>
<td></td>
<td></td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREem</td>
<td>1.463</td>
<td>1.888</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AREmse</td>
<td>-</td>
<td>2.179</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix A. Derivation of asymptotic normality of the maximum composite likelihood estimate using techniques of approximation

\( \tilde{l}(\theta) \) denotes the overall composite log-likelihood function for a random sample of size \( n \). \( \tilde{l}_r \) is \( O_p(n^{-1/2}) \), and \( \tilde{l}_{rs} \) and so on are \( O_p(n) \). The corresponding expectations \( \nu_r \) and \( \nu_{rs} \) and so on are \( O(n^{-1/2}) \) and \( O(n) \), respectively. For convenience, let \( \tilde{\delta} = (\tilde{\theta} - \theta) \), and \( \tilde{\delta}^r = (\tilde{\theta} - \theta)^r \).

According to the expansion (2.21), we find

\[
E(\tilde{\delta}^r) = u^{rs} u^{tu} (\nu_{st,u} + \frac{1}{2} \nu_{stu} u^{vw} \nu_{v,w}) + O(n^{-2}) \tag{A1}
\]

\[
\text{Cov}(\tilde{\delta}^r, \tilde{\delta}^s) = u^{rt} \nu_{t,u} u^{su} + O(n^{-2}) \tag{A2}
\]

\[
\text{cum}(\tilde{\delta}^r, \tilde{\delta}^s, \tilde{\delta}^t) = u^{rs} u^{tu} u^{vw} \nu_{u,v,w} - 3 u^{rs} u^{tu} (\nu_{st,u} + \frac{1}{2} u^{vw} \nu_{stu} \nu_{v,w}) u^{sx} \nu_{x,q} u^{qt} + O(n^{-3}) \tag{A3}
\]

Higher-order cumulants are \( O(n^{-3}) \) or smaller.

In order to achieve the first order normal approximation for the composite likelihood estimate, we actually only need to keep (A1) to error of order \( O(n^{-1}) \), i.e. \( E(\tilde{\delta}^r) = O(n^{-1}) \), and require the third- and higher-order cumulants to be \( O(n^{-2}) \) or smaller, which are fully satisfied here. Note that the notations here have no conflicts with the usual ones, as \( \tilde{l} \) is the total likelihood and so are the other quantities. Hence, in summary, we reach the conclusion on the first-order asymptotic of the maximum composite likelihood estimator consistent with
the earlier result that

\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, G^{-1}) \]

where \( G = HJ^{-1}H \) with \( H \) and \( J \) are defined by (2.3) and (2.4), respectively.

**Appendix B. Derivation of the asymptotic distribution of composite likelihood ratio statistic**

The proof for likelihood ratio tests under model misspecification was given in Kent (1982) which also applies to composite likelihood inference. The problem is assumed regular, and so both \( \hat{\theta} \) and \( \hat{\theta}_0 \) are consistent estimates of \( \theta \). We start with the composite Wald and composite score statistics. Since \( \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, G^{-1}) \), in particular

\[ \sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, G^{\psi\psi}), \]

where \( G^{\psi\psi} \) is evaluated at \((\psi_0, \lambda)\), it follows from Johnson and Kotz (1970) that the asymptotic distribution of \( n(\hat{\psi} - \psi_0)^TH_{\psi\psi,\lambda}(\hat{\psi} - \psi_0) \) is the same as \( \sum_{j=1}^{p} \nu_j V_j \), where \( V_j \) denotes independent \( \chi^2_1 \) variate and \( \nu_j \) are the eigenvalues of \((H^{\psi\psi})^{-1}G^{\psi\psi}\), evaluated under \( H_0 \). That is,

\[ n(\hat{\psi} - \psi_0)^TH_{\psi\psi,\lambda}(\hat{\psi} - \psi_0) \xrightarrow{d} \sum_{j=1}^{p} \nu_j V_j, \]
Similarly, from \( n^{-\frac{1}{2}} U(\theta) \rightarrow_d N(0, J(\theta)) \), we have

\[
n^{-1} U^T(\hat{\theta}_0) H^{-1}_{\psi,\lambda} U^T(\tilde{\theta}_0) \rightarrow_d \sum_{j=1}^p \nu_j V_j.
\]

Next, we derive the asymptotic distribution of the composite likelihood ratio test. Note that by using the notation \( \hat{H}(\theta), \tilde{l}''(\theta^*) = -n \hat{H}(\theta^*) \), and moreover the regularity conditions ensure that \( \hat{H}(\theta^*) = H(\theta) + o_p(1) \). Expanding \( \tilde{l}(\theta) \) around \( \hat{\theta} \) yields

\[
\tilde{l}(\theta) = \tilde{l}(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^T (-n H(\theta))(\theta - \hat{\theta}) + o_p(1).
\]

We further write it in terms of \( \theta = (\psi, \lambda) \) specifically

\[
2\{\tilde{l}(\psi, \lambda) - \tilde{l}(\psi, \lambda)\} = n((\tilde{\psi} - \psi)^T, (\tilde{\lambda} - \lambda)^T) \begin{pmatrix} H_{\psi\psi}(\theta) & H_{\psi\lambda}(\theta) \\ H_{\lambda\psi}(\theta) & H_{\lambda\lambda}(\theta) \end{pmatrix} \begin{pmatrix} \tilde{\psi} - \psi \\ \tilde{\lambda} - \lambda \end{pmatrix} + o_p(1). \quad (B1)
\]

Similarly, for \( \tilde{\theta}_0 = (\psi_0, \lambda_0) \), we have

\[
2\{\tilde{l}(\psi_0, \lambda_0) - \tilde{l}(\psi, \lambda)\} = n((\psi_0 - \psi)^T, (\lambda_0 - \lambda)^T) \begin{pmatrix} H_{\psi\psi}(\theta) & H_{\psi\lambda}(\theta) \\ H_{\lambda\psi}(\theta) & H_{\lambda\lambda}(\theta) \end{pmatrix} \begin{pmatrix} \psi_0 - \psi \\ \lambda_0 - \lambda \end{pmatrix} + o_p(1)
\]

under \( H_0 : \psi = \psi_0 \)

\[
= n((\lambda_0 - \lambda)^T H_{\lambda\lambda}(\lambda_0 - \lambda) + o_p(1). \quad (B2)
\]

Expand the score function \( U.(\hat{\theta}_0) \) around \( (\psi_0, \lambda) \), and use the fact \( U'(\theta^*) = -n \hat{H}(\theta^*) \) and
\[ \hat{H}(\theta^*) = H(\theta) + o_p(1). \] We find

\[
0 = U.(\psi_0, \tilde{\lambda}_0) = U.(\psi_0, \lambda) + (-nH(\theta)) \begin{pmatrix} \psi_0 - \psi_0 \\ \tilde{\lambda}_0 - \lambda \end{pmatrix} + o_p(1),
\]

and simplify it to

\[
U.(\psi_0, \lambda) = n \begin{pmatrix} H_{\psi\lambda}(\tilde{\lambda}_0 - \lambda) \\ H_{\lambda\lambda}(\tilde{\lambda}_0 - \lambda) \end{pmatrix} + o_p(1),
\]

which gives

\[
(\tilde{\lambda}_0 - \lambda) = n^{-1}H_{\lambda\lambda}^{-1}U.(\psi_0, \lambda) + o_p(1).
\]

Meanwhile, from \( U.(\psi, \lambda) = nH(\tilde{\theta} - \theta) + o_p(1), \) we have

\[
U._\lambda(\psi, \lambda) = n(\lambda_{\psi\psi}(\tilde{\psi} - \psi) + \lambda_{\lambda\lambda}(\tilde{\lambda} - \lambda)) + o_p(1).
\]

Then under the null hypothesis \( H_0 : \psi = \psi_0, \) combining the above two equations yields

\[
(\tilde{\lambda}_0 - \lambda) = H_{\lambda\lambda}^{-1}H_{\psi\psi}(\tilde{\psi} - \psi_0) + (\tilde{\lambda} - \lambda) + o_p(1).
\] (B3)

Finally, using equations (B1), (B2) and (B3), the composite likelihood ratio test is ex-
pressed by

\[ \tilde{w}(\psi_0) = 2\{\tilde{l}(\tilde{\theta}) - \tilde{l}(\tilde{\theta}_0)\} \]

\[ = 2\{\tilde{l}(\tilde{\psi}, \tilde{\lambda}) - \tilde{l}(\psi_0, \lambda)\} - 2\{\tilde{l}(\psi_0, \tilde{\lambda}_0) - \tilde{l}(\psi_0, \lambda)\} \]

\[ = n(\tilde{\psi} - \psi_0)(H_{\psi\psi} - H_{\psi\lambda}H_{\lambda\lambda}^{-1}H_{\lambda\psi})(\tilde{\psi} - \psi_0) + o_p(1). \]

Therefore,

\[ \tilde{w}(\psi_0) = n(\tilde{\psi} - \psi_0)H_{\psi\psi, \lambda}(\tilde{\psi} - \psi_0) + o_p(1) \xrightarrow{d} \sum_{j=1}^{p} \nu_j V_j. \]

In addition, the first-order asymptotic equivalence of \( \tilde{w}_e, \tilde{w}_u \) and \( \tilde{w} \) is obtained as a result of the proof, i.e.

\[ \tilde{w} = \tilde{w}_u + O_p(n^{-1/2}) = \tilde{w}_e + O_p(n^{-1/2}). \]

Appendix C. Derivation of the skewness of the signed composite likelihood root statistic

First, we know that

\[ \kappa_3(\tilde{r}) = E(\tilde{r}^3) - 3E(\tilde{r})E(\tilde{r}^2) + 2E^3(\tilde{r}), \]

and by using (2.44) and (2.46), we have

\[ E(\tilde{r}) = \frac{1}{6} u_{11}^{-3/2}(3\nu_{11,1} + u_{11}^{-1}\nu_{11,1}) + O(n^{-3/2}), \]
where the leading term is of order $n^{-1/2}$. Hence, $E^3(\tilde{r}) = O(n^{-3/2})$.

Moreover, (2.37) shows

$$E(\tilde{r}^2) = u_{11}^{-1}\nu_{1,1} + \tilde{R} + O(n^{-2}),$$

where $\tilde{R}$ is of order $n^{-1}$, and so

$$E(\tilde{r})E(\tilde{r}^2) = \frac{1}{6} u_{11}^{-5/2} \nu_{1,1}(3\nu_{11,1} + \nu_3 u_{11}^{-1}\nu_{1,1}) + O(n^{-3/2}).$$

We take the leading term in $\tilde{r}^3$ to the necessary error of order as

$$\tilde{r}^3 = \tilde{r} \cdot \tilde{r}^2 = u_{11}^{-1/2}\tilde{l}_\theta[1 + \frac{1}{6} u_{11}^{-1}(3H_2 + u_{11}^{-1}\nu_3\tilde{l}_\theta)] \cdot (u_{11}^{-1}\tilde{l}_\theta^2 + \frac{1}{3} ((u_{11}^{-1})^3 \nu_3^3 \tilde{l}_\theta^3 + 3(u_{11}^{-1})^2 H_2 \tilde{l}_\theta^2)) + O_p(n^{-3/2})$$

$$= u_{11}^{-3/2}\tilde{l}_\theta + \frac{1}{2} u_{11}^{-5/2}\tilde{l}_\theta^2(u_{11}^{-1}\nu_3\tilde{l}_\theta + 3H_2) + O_p(n^{-3/2}),$$

and then its expectation is given by

$$E(\tilde{r}^3) = u_{11}^{-3/2}\nu_{1,1,1} + \frac{3}{2} u_{11}^{-5/2}(u_{11}^{-1}\nu_3^2 \nu_{1,1,1} + 3\nu_{11,1}\nu_{1,1}) + O(n^{-3/2}).$$

Assembling these terms together, the third order cumulant of $\tilde{r}$ is obtained

$$\kappa_3(\tilde{r}) = u_{11}^{-3/2}\{\nu_{1,1,1} + u_{11}^{-1}\nu_{1,1}(\nu_3 u_{11}^{-1}\nu_{1,1} + 3\nu_{11,1})\} + O(n^{-3/2}). \quad (C1)$$
Furthermore, the standardized third cumulant is the skewness divided by $\kappa_2^{3/2}(\tilde{r})$

$$\rho_3(\tilde{r}) = \nu_{1,1}^{3/2}\{\nu_{1,1} + u_{11}^{-1}\nu_{1,1}(\nu_{1,1}^{-1}\nu_{1,1} + 3\nu_{11,1})\} + O(n^{-3/2}). \quad (C2)$$

**Appendix D. Derivation of the skewness of $\tilde{r}$ for symmetric normal distribution**

Directly using the formulae for the third derivative of the pairwise likelihood for the symmetric normal distribution, we have

$$\nu_3 = \mathbb{E}(l''_2(\rho))$$

$$= \frac{nq(q - 1)(3\rho + \rho^3)}{(1 - \rho^2)^3} - \frac{3}{(1 - \rho^2)^4}\{1 + 4(q - 1)\rho + 6\rho^2 + 4(q - 1)\rho^3 + \rho^4\} \cdot n(q - 1)(1 - \rho)$$

$$+ \frac{3(q - 1)(1 - \rho)^4}{(1 - \rho^2)^4} \cdot n(1 + (q - 1)\rho)$$

$$= -2nq(q - 1)\rho(3 + \rho^2) \frac{1}{(1 - \rho^2)^3}. \quad (D1)$$

Since $SS_W = \sum_{q=1}^{n} \chi^2_{q-1}(1-\rho)$, $\kappa_3(SS_W) = 8n(q-1)(1-\rho)^3$, and $SS_B/q = \sum_{q=1}^{n} \chi^2_{q}(1+(q-1)\rho)$ implies $\kappa_3(SS_B/q) = 8n(1 + (q - 1)\rho)^3$. Further, $SS_W$ and $SS_B$ are independent. Hence, we have
\[ \nu_{1,1,1} = \mathbf{E}(l_2'(\rho))^3 = \kappa_3(l_2') \]
\[ = -\left( \frac{1 + \rho^2 + 2(q-1)\rho}{2(1 - \rho^2)^2} \right)^3 \kappa_3(SS_W) + \left( \frac{(q-1)(1-\rho)^2}{2(1 - \rho^2)^2} \right)^3 \kappa_3\left( \frac{SS_B}{q} \right) \]
\[ = n(q-1)(1-\rho)^3 \{ -(1 + \rho^2 + 2(q-1)\rho)^3 + (q-1)^2(1-\rho)^3(1 + (q-1)\rho)^3 \} \]
\[ = nq(q-1) (1-\rho)^3 d(q,\rho) \]
\[ = nq(q-1) \frac{(1-\rho)^3 d(q,\rho)}{(1 - \rho^2)^6}, \tag{D2} \]

where

\[ d(q,\rho) = -\rho^3(\rho - 1)^3q^4 + \rho^2(\rho - 1)^3(5\rho - 3)q^3 - \rho(-47\rho^2 + 69\rho^3 - 42\rho^4 + 10\rho^5 + 21\rho - 3)q^2 \]
\[ + (\rho - 1)^2(10\rho^4 - 28\rho^3 + 15\rho^2 - 10\rho + 1)q - (\rho - 1)^4(5\rho^2 - \rho + 2). \]

Similarly, using the facts that \( \text{Var}(SS_W) = 2n(q-1)(1-\rho)^2 \) and \( \text{Var}(SS_B/q) = 2n(1 + (q-1)\rho)^2 \), and \( SS_W \) and \( SS_B \) are independent, we have

\[ \mu_{1,1} = \mathbf{E}(l_2''l_2'') = \text{Cov}(l_2',l_2'') \]
\[ = \left( -\frac{1 + \rho^2 + 2(q-1)\rho}{2(1 - \rho^2)^2} \right) \left( -\frac{1}{(1 - \rho^2)^3} \{ q - 1 + 3\rho + 3(q-1)\rho^2 + \rho^3 \} \right) \text{Var}(SS_W) \]
\[ + \left( \frac{(q-1)(1-\rho)^2}{2(1 - \rho^2)^2} \right) \left( -\frac{(q-1)(1-\rho)^3}{(1 - \rho^2)^3} \right) \text{Var}(SS_B/q) \]
\[ = \frac{n(q-1)(1-\rho)^2}{(1 - \rho^2)^5} \{ (1 + \rho^2 + 2(q-1)\rho)(q - 1 + 3\rho + 3(q-1)\rho^2 + \rho^3) \}
\[ - (q-1)(1-\rho)^3(1 + (q-1)\rho)^2 \}
\[ = nq(q-1) \frac{(1-\rho)^2 \rho e(q,\rho)}{(1 - \rho^2)^5},\tag{D3} \]
where

\[ e(q, \rho) = \rho(\rho - 1)^3 q^2 - \rho(\rho - 3)(3\rho^2 - 2\rho + 3)q + (\rho - 1)^2(3\rho^2 - 2\rho + 3), \]

and expressions for \( d(q, \rho) \) and \( e(q, \rho) \) are calculated using Maple 11.

Appendix E. Correlations in the MVB distribution

We assume a common odds ratio \( \psi \) for all pairs. The value \( \psi = 1 \) corresponds to independence. The expectation of the pair \( \mu_{ijk} = E(y_{ij}, y_{ik}) \) is given by

\[
\mu_{ijk} = \frac{[1 + (\psi - 1)(\mu_{ij} + \mu_{ik})] - \sqrt{[1 + (\psi - 1)(\mu_{ij} + \mu_{ik})]^2 - 4\psi(\psi - 1)\mu_{ij}\mu_{ik}}}{2(\psi - 1)}, \quad (E1)
\]

for \( \psi > 0, \psi \neq 1 \). The correlations \( \rho_{ijk} \) can be written as a function of \( \psi \) through

\[
\rho_{ijk} = \frac{\mu_{ijk} - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}\mu_{ik}}}. \quad (E2)
\]

Intuitively, for \( \psi > 1 \), the observations \( (y_{ij}, y_{ik}) \) are concentrated on \((0, 0)\) and \((1, 1)\), which yields larger values of \( \mu_{ijk} \), and in turn larger \( \rho_{ijk} \) from (E2). For \( 0 < \psi < 1 \), the observations are concentrated on \((0, 1)\) and \((1, 0)\), and result in negative correlations.

More specifically, let \( \mu_{ijk} = C_\psi(\mu_{ij}, \mu_{ik}) \), the function \( C_\psi(u, v) \) can be seen as a discrete analogy of the continuous Plackett distribution. As given in Nelsen (1999), the limits of \( C_\psi \)
as $\psi$ goes to 0 and to $\infty$ are the bounds $W$ and $M$, respectively

\[
C_\psi(u, v) = \lim_{\psi \to 0^+} C_\psi(u, v) = \frac{(u + v - 1) + |u + v - 1|}{2} = W(u, v),
\]

\[
C_\psi(u, v) = \lim_{\psi \to \infty} C_\psi(u, v) = \frac{(u + v) - |u - v|}{2} = M(u, v).
\]

**Theorem E.1** The bounds of $\rho_{i:j:k}(\psi)$ are given by

\[
\rho_{i:j:k}(0) = \lim_{\psi \to 0^+} \rho_{i:j:k}(\psi) = \frac{(\mu_{ij} + \mu_{ik} - 1) + |\mu_{ij} + \mu_{ik} - 1| - 2\mu_{ij}\mu_{ik}}{2\sqrt{\nu_{ij}\nu_{ik}}},
\]

(E3)

\[
\rho_{i:j:k}(\infty) = \lim_{\psi \to \infty} \rho_{i:j:k}(\psi) = \frac{(\mu_{ij} + \mu_{ik}) - |\mu_{ij} - \mu_{ik}| - 2\mu_{ij}\mu_{ik}}{2\sqrt{\nu_{ij}\nu_{ik}}}.\]

(E4)

The proof follows immediately from the equation (E2) and results of the limits of $\mu_{i:j:k}$ based on $W$ and $M$. Therefore, under the model setup and assumptions here, theoretically we have $\rho_{i:j:k} \in (-0.3499377, 0.7046881)$. However, as the full joint distribution is misspecified by omitting higher order correlations ($r \geq 3$), the actual interval for $\rho$ working for both likelihoods is narrower. We generate data under three correlation structures: almost no correlation, weak negative correlation, medium positive correlation, selecting $\psi = 1.2, 0.8, 6.$, and the corresponding correlations are given in Table 5.2.
Appendix F. Derivatives of pairwise and full log-likelihood functions for the MVB distribution

Optimizing the full likelihood is based on the log-likelihood derivatives for each cluster $i$

$$ l_i = \sum_{j=1}^{m} (y_{ij} \log \mu_{ij} + (1 - y_{ij}) \log(1 - \mu_{ij})) + \log \{1 + P_i\}, \quad (F1) $$

where

$$ P_i = \sum_{j < k} (\mu_{ijk} - \mu_{ij} \mu_{ik}) (y_{ij} - \mu_{ij}) (y_{ik} - \mu_{ik}) / \nu_{ij} \nu_{ik}. \quad (F4) $$

The partial derivatives of $l_i$ are given by

$$ \frac{\partial l_i(\theta)}{\partial \beta_1} = \sum_{j=1}^{m} \left( x_{ij} y_{ij} - x_{ij} \mu_{ij} \right) \frac{P'_i(\beta_1)}{1 + P_i}, \quad (F2) $$

$$ \frac{\partial l_i(\theta)}{\partial \phi} = \frac{P'_i(\phi)}{1 + P_i} = \frac{1}{1 + P_i} \sum_{j < k} \frac{\mu'_{i;jk}(\phi)}{\nu_{ij} \nu_{ik}} (y_{ij} - \mu_{ij}) (y_{ik} - \mu_{ik}), \quad (F3) $$

where

$$ P'_i(\beta_1) = \sum_{j < k} \frac{\nabla(\beta_1)}{\nu_{ij} \nu_{ik}}, \quad (F4) $$

and

$$ \nabla(\beta_1) = \mu'_{i;jk}(\beta_1)(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}) $$

$$ \quad - \mu_{i;jk}(x_{ij} \nu_{ij} (y_{ik} - \mu_{ik}) + x_{ik} \nu_{ik} (y_{ij} - \mu_{ij} + (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}) [x_{ij} + x_{ik} - 2(x_{ij} \mu_{ij} + x_{ik} \mu_{ik})]) $$

$$ \quad - (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}) \mu_{ij} \mu_{ik} (x_{ij} \mu_{ij} + x_{ik} \mu_{ik}) + \mu_{ij} \mu_{ik} [x_{ij} \nu_{ij} (y_{ik} - \mu_{ik}) + x_{ik} \nu_{ik} (y_{ij} - \mu_{ij})]. $$
In the pairwise likelihood estimation procedure, the derivatives of $l_{ijk}$ are given by

$$
\frac{\partial l_{ijk}(\theta)}{\partial \beta_1} = y_{ij}y_{ik} \frac{\mu'_{ijk}(\beta_1)}{\mu_{ijk}} + y_{ij}(1-y_{ik}) \frac{x_{ij}\nu_{ij} - \mu'_{ijk}(\beta_1)}{\mu_{ijk} - \mu_{ijk}} + (1-y_{ij})y_{ik} \frac{x_{ik}\nu_{ik} - \mu'_{ijk}(\beta_1)}{\mu_{ijk} - \mu_{ijk}} + (1-y_{ij})(1-y_{ik}) \frac{x_{ij}\nu_{ij} - x_{ik}\nu_{ik} + \mu'_{ijk}(\beta_1)}{1 - \mu_{ijk} - \mu_{ijk} + \mu_{ijk}},
$$

(F5)

$$
\frac{\partial l_{ijk}(\theta)}{\partial \phi} = y_{ij}y_{ik} \frac{\mu'_{ijk}(\phi)}{\mu_{ijk}} + y_{ij}(1-y_{ik}) \frac{-\mu'_{ijk}(\phi)}{\mu_{ijk} - \mu_{ijk}} + (1-y_{ij})y_{ik} \frac{-\mu'_{ijk}(\phi)}{\mu_{ijk} - \mu_{ijk}} + (1-y_{ij})(1-y_{ik}) \frac{-\mu'_{ijk}(\phi)}{1 - \mu_{ijk} - \mu_{ijk} + \mu_{ijk}},
$$

(F6)

where

$$
\nu_{ij} = \mu_{ij}(1-\mu_{ij}),
$$

$$
\mu'_{ijk}(\beta_1) = \frac{1}{2} \left( (x_{ij}\nu_{ij} + x_{ik}\nu_{ik} - \frac{1}{\sqrt{b_{ijk}}} [x_{ij}\nu_{ij} + x_{ik}\nu_{ik}] + (e^\phi - 1)(x_{ij}\mu_{ij}(\nu_{ij} + x_{ik}\mu_{ik}\nu_{ik}) - (e^\phi + 1)(x_{ij}\mu_{ik}\nu_{ij} + x_{ik}\mu_{ij}\nu_{ik})) \right),
$$

$$
\mu'_{ijk}(\phi) = \frac{e^\phi}{2(e^\phi - 1)^2 \sqrt{b_{ijk}}} [1 - \sqrt{b_{ijk}} + (e^\phi - 1)(\mu_{ij} + \mu_{ik} - 2\mu_{ij}\mu_{ik})],
$$

$$
b_{ijk} = \left[ 1 - (1-e^\phi)(\mu_{ij} + \mu_{ik})^2 - 4e^\phi(e^\phi - 1)\mu_{ij}\mu_{ik},
$$

and we have used $\mu'_{ij}(\beta_1) = x_{ij}\nu_{ij}$. 


Bibliography


