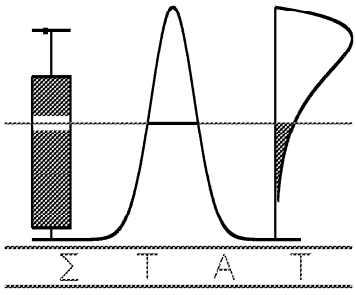
T E C H N I C A L

R E P O R T

0224

Comparing DNA Sequences using Generalized Estimating Equations and Pseudo-Likelihood

K. Van Steen, G. Molenberghs, M. De Wit, and M. Peeters



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Comparing DNA Sequences using Generalized Estimating Equations and Pseudo-Likelihood.

Kristel Van Steen,[1] **Geert Molenberghs,**[1]
Mieke De Wit,[2] **and Monika Peeters** [2]**.**

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

[2] Tibotec-Virco, B-2800 Mechelen, Belgium.

Summary

Proficiency testing is a key part of a laboratory's quality control activities and often takes the form of comparing a new query DNA sequence with sequences generated by well-established reference labs. In this paper, we will show how generalized estimating equations and pseudo-likelihood estimation can be used to set up an equivalence test that assesses the "closeness" of such biological sequences. At the basis lies a marginal model for the probability of observing a match between a new sequence and a reference consensus sequence at a particular locus, using a single (intercept) mean parameter. Because of the simplicity of this model, closed forms of the variance of the intercept parameter are easily derived. When the association parameter lies on the boundary of the parameter space, numerical problems with standard statistical software may give rise to unreliable variance estimates.

Key words: DNA Sequence Comparison, Clustered Binary Data, Generalized Estimating Equations, Pseudo-Likelihood.

1 Introduction

Nowadays, comparative sequence analysis is one of the cornerstones of modern molecular biology. Probably the earliest type of these analyses date back to the 1960's with the work

of Zuckerlandl and Pauling (1965). The title of their work "Molecules as documents of evolutionary history" suggests the usefulness of biological sequence comparison in studying gene (family) evolution. Although it is still an important task to understand organismal diversity and to reconstruct historical events that led to the observed diversity (Maddison 1994), the increasing number of fully sequenced genomes and the Human Genome Project seems to cause a shift towards large-scale functional annotation.

With the development of dynamic programming theory and with the availability of high-speed computers, alignment algorithms became a popular and widely used tool in biological sequence comparisons. The prototype of a global algorithm is the classic Needleman-Wunsch algorithm (Needleman and Wunsch 1970). The Smith-Waterman algorithm is the best known local alignment algorithm (Smith and Waterman 1981). Both alignment methods assign scores to insertions, deletions and replacements, and compute an alignment of two sequences that corresponds to the least costly set of such mutations, hereby maximizing the similarity between the two sequences. However, it is not always straightforward how to choose scoring matrices and gap penalties. Nevertheless, the latter plays an important role in determining the alignment (Vingron and Waterman 1994). Moreover, it seems to be fairly difficult to detect and correct for deviations from the i.i.d. assumption in DNA sequences. In addition, statistical estimates from DNA comparisons are generally less reliable than similar comparisons with protein sequences (Pearson and Wood 2001).

The management of HIV positive people is becoming increasingly complex. The availability of a wide number of antiretroviral drugs and diagnostic tools has significantly improved the survival of people with HIV infection but has also increased the complexity of the management for caring physicians. However, if the quality of the laboratory contribution to patient care (e.g., DNA sequenced samples) is not monitored objectively nor evaluated systemati-

cally, the whole care process breaks down.

Note that many viruses, including that of the human immunodeficiency virus HIV-1, have genomes made of RNA that encode reverse transcriptase. This is an enzyme that makes DNA copies of the RNA genome and integrates them into the genome of the host. One of the first steps in analyzing such an HIV-containing sample is to sequence it. The necessity emerges to create standards in order to guarantee a satisfactory quality level. Solutions will have to be sought in the twilight zone of statistics and bioinformatics.

This paper has the following organization. Section 2 provides a brief description of the data and introduces the format of the data for further reference. In Section 3 we introduce marginal models that will be applied to the data at hand. Section 4 gives an overview of the relevant theory underlying generalized estimating equations and pseudo-likelihood estimation, respectively. In Section 5 it is shown how generalized estimating equations and pseudo-likelihood theory can be applied to the comparison (agreement) problem of a single DNA query sequence with a collection of reference sequences (these are: DNA sequences generated by licensed reference labs that have proven to give consistent results). Results of the practical implementation are presented in Section 6. The latter is followed by concluding remarks in Section 7.

2 Data Description

Proficiency testing is a key part of laboratory's quality control activities. It offers to laboratory customers independent evidence of the laboratory's performance. The purpose of the proficiency testing programme, as set up by Tibotec-Virco, is to enable ongoing monitoring of a laboratory's competence in the genotypic sequencing of HIV-containing samples. A battery of seven samples was selected which contained all relevant genotypic resistance profiles.

These samples were first sent to four reference laboratories for replicate (5 times) sequencing. Participating laboratories sequenced the samples only once ($M = 1$). All nucleotide sequences were summarized using IUPAC-IUB Ambiguity Codes, allowing for mixtures of nucleotides at certain positions.

Practically, each genotype sequence exists of (i) nucleotide/amino acid of HIV-1 PR (codons 1 through 99 $\times$ 3) and (ii) nucleotide/amino acid of HIV-1 RT (codons 1 through 250 $\times$ 3). This results in a total of L=1047 nucleotides. Only one of the samples (sample 002) is selected for illustrative purposes. Since one reference lab sequenced all samples only once, the pool of reference sequences comprises 16 DNA sequences.

All information within a reference lab (subject i) is summarized in a consensus sequence (of length 1047) by selecting the most frequent nucleotide per locus. Ties are broken arbitrarily. Moreover, if every sequence generated within a lab is non-informative with respect to a particular locus j, then the derived consensus sequence shows a missingness code at locus j. In addition, mixtures of nucleotides are treated as missing observation. Note that there are $N = 4$ reference labs.

To assess the quality of a query (new) DNA sequence, we construct for each of the 4 query-reference comparisons a binary string, indicating matching and mismatching loci. If at a particular locus j consistency is observed in the measured or sequenced nucleotide, the locus gives rise to a "match" (coded as 1). In addition, also a missing observation at locus j in both the query sequence and the consensus sequence of a reference lab is regarded as a match. All other cases are referred to as "mismatches" (coded as 0).

3 Model Formulation

The final strings of zeros and ones obtained in the previous section, can be seen as binary time series of length 1047. Extending the theory of generalized linear models (McCullagh and Nelder 1989) to the longitudinal setting, the so-resulting model needs to account for correlations among the multiple observations for an individual. Let $Y_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, N$ represent the jth measurement on the ith subject (here: sequence comparison). Hence, there are n_i measurements on subject i and $\sum_{i=1}^{N} n_i$ total measurements. When inferences about the population average are the focus, the generalized linear model should be expressed as a marginal model where the marginal expectation $E(y_{ij}) = \mu_{ij}$ ($i = 1, \cdots, N$ refers to an experimental unit and $j = 1, \cdots, n_i$ refers to a measurement time) is directly modeled in terms of covariates of interest. The marginal expectation represents the average response over the subpopulation that shares a common value of the covariate vector $\boldsymbol{\beta}$. Associations among the repeated observations are modeled seperately. More specifically, the following assumptions are made (Diggle et al. 1994):

(i) The marginal expectation of the response $E(y_{ij}) = \mu_{ij}$ depends on explanatory variables $\boldsymbol{x_{ij}}$ via $g(\mu_{ij}) = \boldsymbol{x'_{ij}\beta}$, with g a known link function (e.g., logit link for binary responses).

(ii) The marginal variance depends on the marginal mean according to $\mathrm{Var}(Y_{ij}) = v(\mu_{ij})\phi$, where v is a known variance function and ϕ is a scale parameter which may need to be estimated (e.g., $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and $\phi = 1$ for binary responses).

(iii) The correlation between Y_{ij} and Y_{ik} is a function $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\delta})$ of the marginal means and perhaps of additional parameters $\boldsymbol{\delta}$, with $\rho(.)$ a known function (e.g., $\rho(.)$ can indicate an independent, exchangeable, autoregressive AR(1) or unstruc-

tured (working) correlation structure, amongst others).

Alternatively, the association between pairs of responses (Y_{ij}, Y_{ik}) can be modeled via log odds ratios instead of correlations. For example, the multivariate Dale model (Molenberghs and Lesaffre 1994) is a marginal model that uses marginal means and describes the association structure via (multivariate) marginal odds ratios (of second and higher orders). It extends the bivariate global cross-ratio model described by Dale (1986) and McCullagh and Nelder (1989).

Note that the odds ratio

$$\psi_{ijk} = \frac{\Pr(Y_{ij}=1, Y_{ik}=1)\Pr(Y_{ij}=0, Y_{ik}=0)}{\Pr(Y_{ij}=1, Y_{ik}=0)\Pr(Y_{ij}=0, Y_{ik}=1)} \tag{3.1}$$

is not constrained by the means μ_{ij} and μ_{ik}, unlike the correlation $\mathrm{Corr}(Y_{ij}, Y_{ik})$. (Prentice 1988, Lipsitz 1989, Diggle et al. 1994).

In the subsequent section, we will discuss GEE-related features in the context of a marginal model using the correlation between pairs of responses as association function. Pseudo-likelihood estimation will be described using a bivariate Plackett distribution (Plackett 1965) and pairwise odds ratios. Note that it is also possible to apply GEE's in conjunction with a marginal model using log odds ratios specifying the association structure instead of correlations (e.g., Lipsitz et al. 1991).

4 Estimation

4.1 Generalized Estimating Equations

The generalized estimating equations approach (Liang and Zeger 1986, Zeger and Liang 1986, Zeger et al. 1988) is one of the most popular approaches to the analysis of correlated count

or binary data. This multivariate analogue of quasi-likelihood (Wedderburn 1974) is semi-parametric in that the estimating equations are derived without full specification of the joint distribution of a subject's observations. Only the likelihood for the (univariate) marginal distributions and a working covariance matrix for the vector of repeated measurements from each subject need to be specified.

For instance, for the marginal model of Section 3 with the association between responses modeled via correlations, the covariance matrix of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ may be modeled as

$$\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\delta}) \boldsymbol{A}_i^{\frac{1}{2}}, \tag{4.2}$$

where $\boldsymbol{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ is the jth diagonal element. In (4.2), $\boldsymbol{R}(\boldsymbol{\delta})$ represents an $n_i \times n_i$ known, hypothesized, or estimated correlation matrix. Although in principle this working correlation matrix can differ from subject to subject, we commonly let $\boldsymbol{R}(\boldsymbol{\delta})$ approximate the average dependence among repeated observations over subjects (Diggle et al. 1994). The beauty of the method is the ability to choose a working correlation structure which differs from the correct one.

In the absence of a convenient likelihood, parameter estimates are obtained by solving the generalized estimating equations

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial(\mu_{i1}, \ldots, \mu_{in_i})}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1}(\boldsymbol{\delta}) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right),$$

while iterating between quasi-likelihood methods for estimating $\boldsymbol{\beta}$ and an empirical based method for estimating $\boldsymbol{\delta}$, using the sandwich estimator, as a function of $\boldsymbol{\beta}$. It yields consistent estimates for $\boldsymbol{\beta}$ and the corresponding variances, even with misspecification of the structure of the covariance matrix. The efficiency loss relative to maximum likelihood methods is often minimal (Geys et al. 1998).

For completion, we mention that the alternating logistic regressions method (ALR) of Carey

et al. (1993) also relies on generalized estimating equations, but not exclusively as in the approach of Lipsitz et al. (1991). The ALR algorithm alternates between a GEE step to update the model for the mean and a logistic regression step to update the model for the log odds ratios.

4.2 Pseudo-Likelihood Estimation

As indicated in the previous section, GEE's differ from likelihood equations in that they only model the first moments (describing the marginal probabilities) of the joint distribution, and apply working assumptions to construct the information needed from the higher order moments. Another approach is to replace the true contribution $f(y_{i1}, \ldots, y_{in_i})$ of a vector of correlated binary data to the full likelihood by the product of all pairwise contributions $f(y_{ij}, y_{ik})$, $1 \leq j < k \leq n_i$ to obtain a so-called pseudo-likelihood function (Le Cessie and Van Houwelingen 1994).

Applied to the multivariate Dale model (Section 3), the joint probabilities $\mu_{ijk} = f(y_{ij}, y_{ik})$ can be specified in terms of marginal probabilities and pairwise odds ratios using a bivariate Plackett distribution (Plackett 1965):

(i) The bivariate marginal means μ_{ijk} satistfy

$$\mu_{ijk} = \begin{cases} \frac{1+(\mu_{ij}+\mu_{ik})(\psi_{ijk}-1)-\mathrm{S}(\mu_{ik},\mu_{ij},\psi_{ijk})}{2(\psi_{ijk}-1)} & \text{if } \psi_{ijk} \neq 1, \\ \mu_{ij}\mu_{ik} & \text{if } \psi_{ijk} = 1, \end{cases} \tag{4.3}$$

where $S(\mu_{ij}, \mu_{ik}, \psi_{ijk}) = \sqrt{[1 + (\mu_{ij} + \mu_{ik})(\psi_{ijk} - 1)]^2 + 4\psi_{ijk}(1 - \psi_{ijk})\mu_{ij}\mu_{ik}}$,

using (3.1) as definition for the pairwise odds ratios ψ_{ijk} (Fitzmaurice et al. 1995).

(ii) The marginal expectation of the response $E(y_{ij}) = \mu_{ij}$ depends on explanatory variables $\boldsymbol{x_{ij}}$ via $g(\mu_{ij}) = \boldsymbol{x'_{ij}\beta}$, with g an appropriate link function (e.g., logit link).

(iii) The correlation structure between Y_{ij} and Y_{ik} is captured by a model relating the bivariate odds ratio ψ_{ijk} with explanatory variables via a known link function (e.g., log link).

The model specification in items (ii) and (iii) above can be combined into $\boldsymbol{\eta}_i = \boldsymbol{X}_i \boldsymbol{\beta}$, with $\boldsymbol{X}_i$ a known design matrix and $\boldsymbol{\beta}$ the parameter vector of interest. By analogy with maximum likelihood estimation, the maximum pseudo-likelihood estimator of $\boldsymbol{\beta}$ is given by the solution of the so-called pseudo-score equations

$$\boldsymbol{S}_{\text{pseudo}}(\boldsymbol{\beta}) = \boldsymbol{0} \tag{4.4}$$

(in which the likelihood in the classical likelihood-score equations is substituted for the pseudo-likelihood function). Two popular fitting algorithms are the Newton-Raphson and the Fisher Scoring algorithms. Pseudo-likelihood estimation yields consistent and asymptotically normal estimates of the parameters of interest (Arnold and Strauss 1991, Geys et al. 1997). An account of the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure is given in Geys et al. (1998).

In the following section, we will show how pseudo-likelihoods and generalized estimating equations can be used to set up an equivalence test for the closeness of two DNA sequences. At the basis lies a marginal model for the probability of observing a match between a new sequence and a reference consensus sequence, using a single (intercept) mean parameter β_0. Apart from an estimate $\hat{\beta}_0$ of β_0, the test statistic requires an estimate of the variance of $\hat{\beta}_0$.

5 Methodology

5.1 Parameter Estimate Covariances

5.1.1 Generalized Estimating Equations

Adopting the notation of Section 3, the empirical (also referred to as robust) estimator of the covariance matrix of the estimated parameter vector $\hat{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\Sigma}_{\text{emp}} = \boldsymbol{\Sigma}_{\text{mod}}\, \boldsymbol{\Sigma}_{\text{betw}}\, \boldsymbol{\Sigma}_{\text{mod}}, \tag{5.1}$$

with

$$\boldsymbol{\Sigma}_{\text{betw}} = \sum_{i=1}^{N} \frac{\partial(\mu_{i1}, \ldots, \mu_{in_i})}{\partial \boldsymbol{\beta}}\, \boldsymbol{V}_i^{-1}\, \text{Cov}(\boldsymbol{Y}_i)\, \boldsymbol{V}_i^{-1}\, \frac{\partial(\mu_{i1}, \ldots, \mu_{in_i})'}{\partial \boldsymbol{\beta}}\Bigg|_{(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}},\, \boldsymbol{\delta} = \hat{\boldsymbol{\delta}})} \tag{5.2}$$

and

$$\boldsymbol{\Sigma}_{\text{mod}} = \left(\sum_{i=1}^{N} \frac{\partial(\mu_{i1}, \ldots, \mu_{in_i})}{\partial \boldsymbol{\beta}}\, \boldsymbol{V}_i^{-1}\, \frac{\partial(\mu_{i1}, \ldots, \mu_{in_i})'}{\partial \boldsymbol{\beta}} \right)^{-1}\Bigg|_{(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}},\, \boldsymbol{\delta} = \hat{\boldsymbol{\delta}})}. \tag{5.3}$$

Even when under the assumptions for $\boldsymbol{V}_i$, $\boldsymbol{V}_i \neq \text{Var}(\boldsymbol{Y}_i)$, the estimator $\boldsymbol{\Sigma}_{\text{emp}}$ of $\text{Var}(\hat{\boldsymbol{\beta}})$ is consistent. However, if $\boldsymbol{V}_i = \text{Var}(\boldsymbol{Y}_i)$, then $\boldsymbol{\Sigma}_{\text{emp}}$ reduced to the so-called model-based estimator $\boldsymbol{\Sigma}_{\text{mod}}$ for $\text{Var}(\hat{\boldsymbol{\beta}})$.

We will show that, in our situation, these formulae reduce to simple expressions. As model for the marginal expectation of the binary responses Y_{ij} we suggest

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} = \beta_0, \tag{5.4}$$

with $g(.)$ the logit link function. The marginal variance $\text{Var}(Y_{ij})$ naturally equals $\mu_{ij}(1-\mu_{ij})$. The (j,k)th element of a working correlation matrix $\boldsymbol{R}(\boldsymbol{\delta})$ is hypothesized to be

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & \text{if } j = k, \\ \delta & \text{if } j \neq k, \end{cases} \tag{5.5}$$

corresponding to exchangeable working conditions. This relies on the plausible assumption that the locationwise comparisons between the query sequence and a reference lab are interchangeable.

Note that $\frac{\partial g(\mu_{ij})}{\partial \beta_0} = g'(\mu_{ij})\Big|_{\mu_{ij}=\frac{\exp\beta_0}{(1+\exp\beta_0)}} \frac{\partial \mu_{ij}}{\partial \beta_0} = 1$ with

$$g'(\mu_{ij})\Big|_{\mu_{ij}=\frac{\exp\beta_0}{(1+\exp\beta_0)}} = \frac{1}{\mu_{ij}(1-\mu_{ij})}\Big|_{\mu_{ij}=\frac{\exp\beta_0}{(1+\exp\beta_0)}} = \frac{(1+\exp\beta_0)^2}{\exp\beta_0}$$

subsequently denoted as $g'(\beta_0)$. Using (5.3) and letting $\boldsymbol{V}_i^{-1} = (\omega_{jk})_{1\leq j,k\leq 1047}$, it follows that

$$\boldsymbol{\Sigma}_{\text{mod}} = \left(\frac{N}{g'^2(\beta_0)} \sum_{k-1}^{L} \sum_{j=1}^{L} \omega_{jk}\right)^{1}. \tag{5.6}$$

Applying (4.2) the (r,s)th element of $\boldsymbol{V}_i$ under exchangeability is determined by

$$(\boldsymbol{V}_i)_{(r,s)} = \frac{\delta I(r \neq s) + I(r=s)}{g'(\beta_0)},$$

with $I(.)$ representing an indicator operator which equals 1 if its argument holds and zero otherwise. Consequently, if $\delta \neq 1, \frac{1}{1-L}$ such that $1+(L-2)\delta-(L-1)\delta^2 \neq 0$,

$$\omega_{jk} = \begin{cases} \frac{g'(\beta_0)(1+(L-2)\delta)}{1+(L-2)\delta-(L-1)\delta^2} & \text{if } j=k, \\ \\ \frac{-g'(\beta_0)\delta}{1+(L-2)\delta-(L-1)\delta^2} & \text{if } j\neq k. \end{cases} \tag{5.7}$$

For more details we refer to Appendix A. Substituting (5.7) in (5.6) leads to

$$\boldsymbol{\Sigma}_{\text{mod}} = \frac{g'(\beta_0)}{N} \frac{1-(1-L)\delta}{L}. \tag{5.8}$$

With $\text{Cov}(\boldsymbol{Y}_i)$ in (5.2) estimated as $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'$ and using (5.7), it is straightforward to show that

$$\boldsymbol{\Sigma}_{\text{betw}} = \sum_{i=1}^{N} \left(\frac{1}{1-(1-L)\delta} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik})\right)^2 \tag{5.9}$$

(Appendix B).

Substituting the expressions for $\mathbf{\Sigma}_{\text{mod}}$ (5.8) and $\mathbf{\Sigma}_{\text{betw}}$ (5.9) in (5.1), we obtain

$$\mathbf{\Sigma}_{\text{emp}} = \frac{g'^2(\beta_0)}{N^2} \sum_{i=1}^{N} \left(\frac{1}{L} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik}) \right)^2 .$$

In particular, as an explicit function of the parameter of interest, β_0,

$$\mathbf{\Sigma}_{\text{emp}} = \frac{(1+\exp\beta_0)^4}{\exp^2\beta_0} \frac{1}{N^2} \sum_{i=1}^{N} \left([\frac{1}{L} \sum_{k=1}^{L} Y_{ik}] - \frac{\exp\beta_0}{1+\exp\beta_0} \right)^2 . \tag{5.10}$$

5.1.2 Pseudo-Likelihood

Building on Section 4.2 and assuming exchangeability of cluster elements, we now denote with π_{i11} the bivariate probability of observing two successes and with π_{i10} the probability of observing two successes, or a success in the first component and a failure in the second component. By doing so, and by applying the Fisher scoring fitting algorithm to solve (4.4), the asymptotic covariance matrix of the parameters $\boldsymbol{\beta}$ is turned into the following simple form:

$$\boldsymbol{W}(\hat{\boldsymbol{\beta}})^{-1} \left(\sum_{i=1}^{N} \boldsymbol{S}_{\text{pseudo},i}(\hat{\boldsymbol{\beta}}) \boldsymbol{S}_{\text{pseudo},i}(\hat{\boldsymbol{\beta}})^T \right) \boldsymbol{W}(\hat{\boldsymbol{\beta}})^{-1}. \tag{5.11}$$

Here, $\boldsymbol{W}(\boldsymbol{\beta}) = \sum_{i-1}^{N} \boldsymbol{X}_i^T (\boldsymbol{T}_i^{-1})^T \boldsymbol{A}_i (\boldsymbol{T}_i^{-1}) \boldsymbol{X}_i$, $\boldsymbol{A}_i$ is the expected value of the matrix of second order derivatives of the log pseudo-likelihood with respect to $\boldsymbol{\pi}_i = (\pi_{i10}, \pi_{i11})'$, $\boldsymbol{T}_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\pi}_i}$, $\boldsymbol{\eta} = (\eta_{i1}, \eta_{i2})'$, $\eta_{i1} = \ln(\pi_{i10}) - \ln(1-\pi_{i10})$ and $\eta_{i2} = \ln(\psi_i) = \ln(\pi_{i11}) + \ln(1 - 2\pi_{i10} + \pi_{i11}) - 2\ln(\pi_{i10} - \pi_{i11})$, $\boldsymbol{\eta} = \boldsymbol{X}_i\boldsymbol{\beta}$, with $\boldsymbol{X}_i$ a known design matrix and $\boldsymbol{\beta}$ a vector of unknown regression parameters (Geys et al. 1998).

We propose the constant marginal odds ratio model:

$$\begin{cases} \text{logit}(\pi_{i10}) &= \beta_0 \\ \ln\psi_i &= \beta_1. \end{cases} \tag{5.12}$$

Note that β_1 above plays the role of the association parameter δ in (5.5).

5.2 Test Procedure

5.2.1 Generalized Estimating Equations

The $(1-\alpha)100\%$ confidence interval of the parameter β_0 is given by

$$[\hat{\beta}_0 - z_{1-\alpha/2}\sqrt{\hat{\mathbf{\Sigma}}_{\text{emp}}}, \hat{\beta}_0 + z_{1-\alpha/2}\sqrt{\hat{\mathbf{\Sigma}}_{\text{emp}}}], \tag{5.13}$$

where $\hat{\beta}_0$ is the GEE estimate of β_0, $\hat{\mathbf{\Sigma}}_{\text{emp}}$ is given by (5.10) evaluated in $\hat{\beta}_0$ and in which $z_{1-\alpha/2}$ symbolizes the upper $100\alpha/2$ percentile of a standard normal distribution. These Wald confidence limits are computed by assuming an asymptotic normal distribution for the parameter estimator $\hat{\beta}_0$. Although alternatives for the classical Wald test for GEE regression parameters have been proposed (Rotnitzky and Jewell 1990), the proposed confidence interval (5.13) serves our purposes well. Indeed, often one focuses on the independence model of genome decomposition. As Wu et al. (1997) mention, the independence of nucleotides is only an approximation to the actual independence in DNA sequences. However, Arratia et al. (1990) evaluated this approximation and found it to be quite good. Note that if the independence model for nucleotides in a DNA sequence is regarded as a true model, our models simplify to a logistic regression model.

However, our main interest lies in finding evidence towards β_0 parameters that are "large enough" (corresponding to high overlap probabilities). The decision rule we adopt is listed in Table 1 and is based on the test statistic

$$z^\star = \frac{\hat{\beta}_0 - \beta_{\text{equiv}}^{\text{GEE}}}{\sqrt{\mathbf{\Sigma}_{\text{emp}}}}, \tag{5.14}$$

with $\mathbf{\Sigma}_{\text{emp}}$ as in (5.10) and $\beta_{\text{equiv}}^{\text{GEE}}$ separating the "equivalence region" from the "inequivalence region". For example, if field workers support an equivalence region of $[\pi_{\text{equiv}} = 0.97, 1]$ (1 is the maximal success rate and corresponds with a perfect match), obviously $\beta_{\text{equiv}}^{\text{GEE}} = \ln\left(\frac{\pi_{\text{equiv}}}{1-\pi_{\text{equiv}}}\right)$.

Table 1: ABOUT HERE

5.2.2 Pseudo-likelihood

Also when modeling the association between pairs of responses via log odds ratios instead of correlations and using a pseudo-likelihood estimation approach, the parameter of interest is β_0 (5.12). However, the interpretation of β_0 here is somewhat different as compared to the previous section: here, β_0 is related to the probability of observing at least one succes in a pair of responses via the logit link function.

Whereas the protocol for the test remains the same (Table 1), the denomenator of the test statistic $z^\star$ defined before is replaced by the square root of the robust variance as in (5.11).

6 Results and Discussion

After delineating the equivalence region, we can perform the equivalence test as specified in Section 5 (Table 1), using $\alpha = 0.01$ as significance level. For illustrative purposes, we set the boundary point specifying the equivalence region within a GEE or pseudo-likelihood framework respectively equal to $\beta_{\text{equiv}}^{\text{GEE}} = \ln\left(\frac{0.97}{1-0.97}\right) = 3.4761$ and $\beta_{\text{equiv}}^{\text{PL}} = \ln\left(\frac{0.98}{1-0.98}\right) = 3.8918$.

6.1 Generalized Estimating Equations

Adopting a generalized estimating equations approach and proposing an exchangeable 1047 by 1047 working correlation matrix (Section 5.1.1), the mean parameter of interest β_0 is estimated as 4.6415. According to (5.10) with δ estimated as 0.0003, the corresponding empirical based standard error we need in (5.14) is estimated as 0.2408. However, since δ

appears to be small, it would make sense to continue working with independence working assumptions (i.e., $\delta = 0$) and hence to use the initial parameter estimate of β_0 in the iterative GEE estimation algorithm under exchangeable working assumptions. In our data set, this initial parameter estimate of β_0 coincides with the earlier obtained estimate of 4.6415. Therefore, since δ only implicitly occurs in (5.10) via β_0, the empirical based standard error for β_0 under $\delta = 0$ remains 0.2408.

Consequently, the test statistic $z^\star = \frac{4.6415-3.4761}{0.2408} = 4.8397$. Since $z^\star > z_{1-\alpha} = 2.3263$, we conclude that the new query sequence is similar enough to the reference consensus sequences, guaranteeing the proficiency of the new lab (assuming that the new lab has successfully completed internal consistency checks).

The estimates for β_0 and δ above were obtained using the SAS procedure GENMOD. There appears to be a discrepancy between standard errors as provided in the SAS output and those provided by (5.8) and (5.10). Summarizing all available reference sequences into a single consensus sequence, and constructing a match/mismatch binary sequence as before by comparing this consensus sequence with the query sequence, the association parameter under an exchangeable working correlation matrix is estimated as $\hat{\delta} = -0.0010$. However, since $\boldsymbol{V}_i$ is a multiple of $(1-\delta)I_{L\times L} + \delta J_{L\times L}$ (the scalar in the product being $\frac{1}{g'(\beta_0)}$, the eigenvalues of $\boldsymbol{V}_i$ can be shown to be $1-\delta$ (multiplicity $L-1$) and $1-(1-L)\delta$ (multiplicity 1). In order for $\boldsymbol{V}_i$ to be positive definite, we have to impose that $1-\delta > 0$ and $1-(L-1)\delta > 0$. Equivalently,

$$\delta \in]\frac{1}{1-L}, 1[.$$

Note that $\delta \neq 1$ and $\delta \neq \frac{1}{1-L}$ are also the conditions for the covariance matrix $\boldsymbol{V}_i$ to be invertible (Appendix A). With $L = 1047$, it is clear that $\hat{\delta} = -0.0010$ is near the lower bound $\frac{1}{1-L} = -0.00096$ of the parameter space for δ. The latter seems to cause numerical

problems resulting in large discrepancies between SAS generated empirical based standard errors under different working assumptions (Table 2).

Table 2: ABOUT HERE

On the other hand, if δ is believed to be small, then the estimate of δ via Pearson residuals (which are in turn functions of β_0) will be close to the lower bound of the parameter space, if (Appendix C)

$$\frac{N(1-L)}{N(1-L)+\frac{1}{L}} \approx \exp \beta_0. \tag{6.1}$$

Hence, with L as large as 1047, this is true whenever β_0 is nearly zero. In other words, if the success rate is nearly 1, the estimate of δ will move towards the boundary of the parameter space. Note that in our situation, the success rate is 0.9876. In general, the success rate will decrease with an increasing number of binary sequences. In our situation, taking up four match/mismatch sequences in the analysis leads to a slightly smaller global success rate of 0.9864. As expected from (6.1), numerical problems are still bound to occur (note the deviation of (5.10) from 0.1785), but at least SAS generates a non-zero empirical based standard error under exchangeable working assumptions (Table 2). The large discrepancy between the model based standard errors under the exchangeability model is reduced (Table 2, column 2).

6.2 Pseudo-Likelihood Estimation

Using pseudo-likelihood estimation and the odds ratio model (5.12), the parameter of interest β_0 is estimated as 4.6415. It is not surprising that this number agrees with the estimate for β_0 (5.4) obtained in Section 6.1. Indeed, for "highly similar" sequences the probability of a succes at some location will be comparable with the probability of having at least one

succes in a pair of locations. However, using pseudo-likelihood estimation, we observe that a more accurate estimation is obtained, since 0.1785 is found as associated robust standard error (5.11), as compared to 0.2408 via the closed form expression (5.10). The association is quantified via β_1 and estimated as 0.0214 (standard error: 0.0672).

Demarcating the equivalence region via $\beta_{\text{equiv}}^{\text{PL}} = \ln\left(\frac{0.98}{1-0.98}\right) = 3.8918$, the test statistic $z^\star$ takes on the value $\frac{4.6415-3.8918}{0.1785} = 4.2003$. Since $z^\star > z_{1-\alpha} = 2.3263$, evidence is found towards proficiency of the new lab.

For reasons of comparability with Section 6.1, we apply the technique to a single match / mismatch binary sequence by comparing the query sequence with the consensus sequence derived from all reference labs in the study. In this particular situation, we obtain a parameter estimate of $\hat{\beta}_0$=3.5235 with associated robust standard error of 0.0001. The association parameter β_1 turns out to be negative in a significant way (-0.0348, with estimated standard error 0.0001). As a side-effect of the high accuracy level in the estimation, the test statistic $z^\star$ blows up in absolute value: $z^\star = \frac{3.5235-3.8918}{0.0001} = -2612.9228$. Note that by using the condensed information of a single consensus sequence the null hypothesis of inequivalent sequences can no longer be rejected!

However, the numerical issues brought up for discussion in Section 6.1 are also relevant in this setting. Therefore it is not surprising that the summarized results within a pseudo-likelihood approach (Table 3, obtained via GAUSS code) are in close agreement with the ones obtained in Table 1 (SAS - e.g., exchangeable working assumptions).

Table 3: ABOUT HERE

As for the reliability of the estimated association parameter, we note that Zhao and Prentice

(1990) and Liang et al. (1992) extended the GEE method (GEE2) for simultaneous estimation of regression parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\delta}$. In practice, this requires modeling the third and fourth moments of y_{ij}, instead of just modeling the mean and variance as in the previous case (also referred to as GEE1). Lipsitz et al. (1994) extended Liang and Zeger's method (1986) to models for the correlation between repeated nominal and ordinal categorical responses; in particular, when the repeated responses are binary, their methods reduce to Liang and Zeger's method.

GEE2 applied to the Dale model for clustered binary data again makes us suspicious about presented accuracy (robust standard errors) of the parameter estimates for β_0 and β_1 (Table 3) when a single consensus reference sequence is used. In case $N = 4$, the estimates for β_0 obtained via pseudo-likelihood or GEE2 are comparable. Whatever approach taken, the association parameter seems to be non-significant.

7 Conclusion

In this paper we proposed a model-based method to assess the agreement between two biological sequences. Unlike with random-effects models and transition models, the use of maximum likelihood methods to estimate unknown parameters of a marginal model based on (4.2) may be infeasible (it may be too laborious or impossible to fully specify the likelihood). In such a situation, where specification of the full likelihood becomes cumbersome, generalized estimating equations (GEE's) or pseudo-likelihoods are a useful alternative.

Within the framework of HIV-proficiency testing, we set up an equivalence test to test the closeness of a query DNA sequence with a set of reference sequences. To this end, we constructed strings of match/mismatch codes, which can be regarded as time series, time

points being the loci in the biological sequence. At first sight, all requisites for the developed test statistic can be estimated via standard statistical software.

However, we illustrated that if the association parameter (describing the association between two responses withing the same cluster or string) is near the boundary of the parameter space, numerical problems may lead to spurious results. Hence, in the current data setting, it may be better to rely on closed form formulae for standard errors or confidence intervals (such as those derived within a GEE framework) to implement the equivalence test.

Acknowledgement

Research supported by a PAI program P5/24 of the Belgian Federal Government (Federal Office for Scientific, Technical, and Cultural Affairs).

References

Arratia, R., Gordon, L. and Waterman, W.S. (1990) The Erdös-Rényi law in distribution, for coin tossing and sequence matching, *Annals of Statistics*, **18**, 539 – 570.

Arnold, B.C. and Strauss, D. (1991) Pseudolikelihood estimation: some examples, *Sankhya B*, **53**, 233 – 243.

Cox, D.R. (1972) The analysis of multivariate binary data, *Applied Statistics*, **21**, 113 – 120.

Dale, J.R. (1986) Global cross-ratio models for bivariate, discret, ordered responses, *Biometrics*, **42**, 909 – 917.

Davis, C.S. (1999) Analysis of repeated measurements using the GEE method. Short course presented at ENAR 1999 Spring Meeting

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994) *Analysis of Longitudinal Data*, New-York: Oxford University Press.

Fitzmaurice, G.M., Laird, N.M. and Rotnitzky, A.G. (1993) Regression models for discrete longitudinal responses (with Discussion), *Statistical SCience*, **8**, 284 – 309.

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

Geys, H., Molenberghs, G. and Lipsitz, S. (1998) A note on the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure, *J. Statist. Comput. Simul.*, **62**, 45 – 71.

Le Cessie, S. and Van Houwelingen, J.C. (1994) Logistic regression for correlated binary data, *Applied Statistics*, **43**, 95 – 108.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13 –22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analysis for categorical data (with discussion), *J. R. Statist. Soc.* B, **54**, 3 – 40.

Lipsitz, S. (1989) Generalized estimating equations for correlated binary data: using odds ratio as a measure of association. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Lipsitz, S.R., Kim, K. and Zhao, L. (1994) Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine*, **13**, 1149 – 1163.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika*, **78**, 153 – 160.

Maddison, D.R. (1994) Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters, *Annu. Rev. Entomol.*, **39**, 267 – 292.

McCullagh, P and Nelder, J.A. (1989) *Generalized Linear Models*, London: Chapman and Hall.

McLachlan, G.J. and Basford, K.E., (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modeling of correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Society*, **89**, 633 – 644.

Needelman, S.B. and Wunsch, Ch.D. (1970) A general method applicable to search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol*, **48**, 443 – 453.

Pearson, W.R. and Wood, T.C. (2001) Statistical significance in biological sequence comparison, *Handbook of Statistical Genetics*. Chapter 2. Edited by D.J. Balding et al. , New York: John Wiley and Sons.

Plackett, R.L. (1965) A class of bivariate distributions, *Journal of the American Statistical Association*, **60**, 516 – 520.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation, *Biometrics*, **44**, 1033 – 1048.

Rotnitzky, A. and Jewell, N.P. (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered correlated data, *Biometrika*, **77**, 485 – 497.

Searle, S.R. (1982) *Matrix algebra useful for statistics*, New York: John Wiley and Sons.

Smith, J.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, *147*, 195 – 197.

Smith, T.F. (1999) The art of matchmaking: sequence alignment methods and their structural implications. *Structure*, **7**, R1 – R12.

Spitzer, R.L., Cohen, J., Fleiss, J.L., and Endicott, J. (1967) Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry*, **17**, 83 – 87.

Vingron, M. and Waterman, M.S. (1994) Review article: Sequence alignment and penalty choice - Review of concepts, case studies and implications, *J. Mol. Biol.*, **235**, 1 – 12.

Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models and the Gaussian method, *Biometrika*, **61**, 685 – 700.

Westlake, W.J. (1974) The use of balanced incomplete block designs in comparative bioavailability trials, *Biometrics*, **30**, 319 – 327.

Wu, T-J., Burke, J.P. and Davison, D.B. (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, *Biometrics*, **53**, 1431 – 1439.

Zeger, S.L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, **42**, 121 – 130.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, **44**, 1049 – 1060.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model, *Biometrika*, **77**, 642 – 648.

Zuckerlandl, E. and Pauling, L.C. (1965) Molecules as documents of evolutionary history, *J. Theoret. Biol.*, **8**, 357 – 358.

Appendix A

The inverse of $\boldsymbol{V}_i$

Since $(\boldsymbol{V}_i)_{(r,s)} = \dfrac{\delta I(r \neq s) + I(r = s)}{\sqrt{g'(\mu_{ir})g'(\mu_{is})}}$ and since $\frac{1}{g'(\mu_{ir})} = \frac{\exp\beta_0}{(1+\exp\beta_0)^2}$ is independent of r, we can simplify

$$\boldsymbol{V}_i = \frac{1}{g'(\mu_{ir})} \begin{pmatrix} 1 & \delta & \dots & \delta \\ \delta & 1 & \dots & \delta \\ \vdots & \vdots & \ddots & \vdots \\ \delta & \delta & \dots & 1 \end{pmatrix} \text{ for all } r,\ 1 \leq r \leq L = 1047.$$

Observing that

$$\begin{pmatrix} 1 & \delta & \dots & \delta \\ \delta & 1 & \dots & \delta \\ \vdots & \vdots & \ddots & \vdots \\ \delta & \delta & \dots & 1 \end{pmatrix}^{-1}_{L\times L} = \left((1-\delta)I_{L\times L} + \delta J_{L\times L}\right)^{-1},$$

with $I_{L\times L}$ an L-dimensional identity matrix and $J_{L\times L}$ an L-dimensional matrix of ones, it follows that for all r, $1 \leq r \leq L = 1047$,

$$\boldsymbol{V}_i^{-1} = \frac{g'(\mu_{ir})}{1+(L-2)\delta-(L-1)\delta^2} \begin{pmatrix} 1+(L-2)\delta & -\delta & \dots & -\delta \\ -\delta & 1+(L-2)\delta & \dots & -\delta \\ \vdots & \vdots & \ddots & \vdots \\ -\delta & -\delta & \dots & 1+(L-2)\delta \end{pmatrix},$$

provided $\delta \notin \{1, \frac{1}{1-L}\}$ (Searle 1982).

Appendix B

Explicit form of $\boldsymbol{\Sigma}_{\mathbf{betw}}$

With $\mathrm{Cov}(\boldsymbol{Y}_i)$ in (5.2) estimated as $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'$ and using (5.7), we can derive

$$
\begin{aligned}
&\boldsymbol{\Sigma}_{\mathrm{betw}} \\
&\quad = \sum_{i=1}^{N} \left(\frac{1}{g'(\beta_0)} \sum_{j=1}^{L} \sum_{k=1}^{L} \omega_{jk} (Y_{ik} - \mu_{ik}) \right)^2 \\
&\quad = \sum_{i=1}^{N} \left(\frac{1}{g'(\beta_0)} \sum_{k=1}^{L} \omega_{kk} (Y_{ik} - \mu_{ik}) + \frac{1}{g'(\beta_0)} \sum_{k=1}^{L} \sum_{\substack{j=1 \\ j \neq k}}^{L} \omega_{jk} (Y_{ik} - \mu_{ik}) \right)^2 \\
&\quad = \sum_{i=1}^{N} \left(\frac{1 + (L-2)\delta}{1 + (L-2)\delta - (L-1)\delta^2} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik}) - \frac{(L-1)\delta}{1 + (L-2)\delta - (L-1)\delta^2} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik}) \right)^2 \\
&\quad = \sum_{i=1}^{N} \left(\frac{1 - \delta}{1 + (L-2)\delta - (L-1)\delta^2} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik}) \right)^2 \\
&\quad = \sum_{i=1}^{N} \left(\frac{1}{1 - (1-L)\delta} \sum_{k=1}^{L} (Y_{ik} - \mu_{ik}) \right)^2
\end{aligned}
$$

Appendix C

Estimation of δ

With the notations and model of Section 5.1.1, we denote the Pearson residual by $e_{ij} = \frac{y_{ij}-\mu_{ij}}{\sqrt{v(\mu_{ij})}}$ The GENMOD procedure in SAS uses these residuals to estimate the association parameter δ in the following way:

$$\begin{aligned}
\hat{\delta} &= \tfrac{1}{N\,L(L-1)-1} \textstyle\sum_{i=1}^{N} \sum_{j\neq k}^{L} e_{ij}e_{ik} \\
&= \tfrac{1}{N\,L(L-1)-1} \textstyle\sum_{i=1}^{N} \sum_{j\neq k}^{L} (\tfrac{y_{ij}}{\mu_{ij}} - 1)(\tfrac{y_{ik}}{\mu_{ij}} - 1) \\
&= \tfrac{-N\,L}{N\,L(L-L)-1} \tfrac{1}{\exp\beta_0}.
\end{aligned}$$

The latter step is based on the assumption that for a small association parameter (this is: δ close to zero), $\mu_{ij} = \frac{\exp\beta_0}{1+\exp\beta_0}$ can be approximated by $\sum_{i=1}^{N}\sum_{j=1}^{L} \frac{y_{ij}}{N\,L}$. Hence, δ will approach the lower bound of its parameter space, if

$$\frac{-N\,L}{N\,L(L-1)-1}\frac{1}{\exp\beta_0} \approx \frac{1}{1-L},$$

or if

$$\frac{N(1-L)}{N(1-L)+\frac{1}{L}} \approx \exp\beta_0.$$

Table 1: Outline of test protocol, with α as pre-specified significance level.

Alternatives	Decision Rule
$H_0 : \beta \leq \beta_{\text{equiv}}$	If $z^\star \leq z_{1-\alpha}$, conclude H_0
$H_A : \beta > \beta_{\text{equiv}}$	If $z^\star > z_{1-\alpha}$, conclude H_A

Table 2: Empirical and model based standard errors for the intercept model parameter β_0, via derived closed form formulae (5.10, 5.8) and SAS proc genmod output. Software computational problems appear to be caused by the fact that the association parameter is near the boundary of the parameter space. Under exchangeable working assumptions, the association parameter is estimated as 0.0003 when four reference consensus sequences are used. When the query sequence is compared to a consensus sequence derived from all available reference sequences (leading to a single "time series"), the association parameter is estimated as -0.0010.

	Exchangeability		Independence	
	Model based s.e. of β_0	Empir. based s.e. of β_0	Model based s.e. of β_0	Empir. based s.e. of β_0
$N = 4$				
SAS	0.1785	0.1785	0.1589	0.1785
Closed Form	0.1821	0.2408	0.1589	0.2408
$N = 1$				
SAS	0.0057	0.0000	0.1852	0.1785
Closed Form	-	0.5834	0.1852	0.5834

Table 3: Parameter estimates for the constant marginal odds ratio model (5.12) with associated robust standard errors according to formula (5.11), taking the exchangeability assumtion into account.

	Parameter	Pscudo-likclihood Estimate (s.e.)	GEE2 Estimate (s.e.)
$N = 4$			
	β_0	4.6415 (0.1785)	4.6415 (0.1785)
	β_1	0.0214 (0.0672)	0.0027 (0.0675)
$N = 1$			
	β_0	3.5235 (0.0001)	3.5234 (3.9e-015)
	β_1	-0.0348 (0.0001)	-0.0349 (1.7e-012)