Inference for clustered data using the independence log-likelihood

Richard E. Chandler and Steven Bate
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT, UK

Abstract

Independence estimating equations (IEEs) are commonly used for the analysis of clustered data. Here, we use the properties of IEEs to adjust the “independence” log-likelihood function in the presence of clustering. The proposed adjustment relies on the robust sandwich estimator of the parameter covariance matrix, which is easy to calculate. We show that the adjusted log-likelihood competes favourably with a number of established techniques based on IEEs, and provide some insight as to why this is so. The adjustment is applied to practical examples relating to the modelling of wind speed in Europe, and annual maximum temperatures in the UK.

Key words: Generalized estimating equations, inference from margins, multivariate extremes, spatial-temporal modelling.

1 Introduction

Dependent data structures arise in many application areas. In the analysis of data from longitudinal studies (e.g. Diggle et al. 2002), dependence arises because repeated measurements are made on the same individuals over a period of time. The resulting data are often described as “clustered”; observations from the same cluster (i.e. measurements made on the same individual) are mutually dependent, but independent of those from other clusters. Similar structures arise in the analysis of multiple time series from different spatial locations (e.g. Chandler et al. 2006); here, each time point contributes a cluster of dependent observations and, providing the time series structure is modelled appropriately, these clusters can be treated as independent.

Broadly speaking, there are two possible approaches to the analysis of clustered data. The first is to model the dependence explicitly: this underlies techniques such as Generalized Least Squares (Harvey, 1989, Section 3.4.2), Generalized Estimating Equations...

Research report No. 255, Department of Statistical Science, University College London.
Date: August 2005.
(GEEs) (Liang and Zeger, 1986), copula-based multivariate models (Joe, 1997) and, al-
beit from a slightly different perspective, hierarchical or multilevel models (Goldstein,
1995; Snijders and Bosker, 1999). GEEs are suitable for analyses based on Generalized
Linear Models (GLMs); the basic idea is to replace the diagonal covariance matrix in the
GLM score equations with a “working” covariance matrix that reflects the within-cluster
dependencies. In the statistical literature, GEEs have been widely adopted as a means of
dealing with clustered data in analyses based on GLMs.

The second approach is to fit models as though the observations are independent and
then to adjust, for example, the estimated standard errors to account for the dependence.
For models based on GLMs, this in fact corresponds to the use of an “independence”
working covariance structure within the GEE framework. The resulting “independence
estimators” are solutions of “independence estimating equations” (IEEs). Another view of
this approach is that it uses only the marginal distributions for each observation, rather
than the full joint distribution; in the multivariate modelling literature, it is therefore
referred to as “inference from margins” (Joe, 1997).

In general, independence estimators are inefficient by comparison with those based
on a correctly specified dependence structure; this inefficiency can be substantial in the
presence of covariates that vary within clusters, when the within-cluster dependence is
high (Fitzmaurice, 1995). However, studies indicate that severe loss of efficiency is the
exception rather than the rule (e.g. Henderson and Shikamura 2003; Sutradhar and Das
1999; Joe 1997, Section 10.1.2; McDonald 1993; Liang and Zeger 1986). Moreover, in some
situations (for example, when there is unmodelled dependence between each observation
and the covariates of other observations in the same cluster) GEE estimators are biased
unless an independence working covariance structure is used (Sullivan Pepe and Anderson,
1994; Park and Kim, 2004); they can also lack robustness to mis-specification of the
working covariance structure and to contamination of the response (Qu and Song, 2004),
and can suffer from parameter non-identifiability (Crowder, 1995). In situations involving
weak or moderate dependence therefore, there are compelling grounds for using IEEs, since
the increase in robustness more than compensates for the slight loss in efficiency.

Another argument for the use of IEEs, at least for the analysis of large and complex
data sets, is computational. Our own motivation for this work stems from an interest in
spatio-temporal modelling for application areas such as climatology, in which data sets
containing tens or hundreds of thousands of observations are common. When large quan-
tities of data are available, some loss of efficiency can be tolerated in routine analyses, and
computation time may be a more important consideration. In such situations the extra
level of iteration involved to implement, for example, GEE estimation is unattractive.

The properties of IEEs, and the estimators they define, are well known. Inference is
usually based upon the asymptotic normality of the estimators, as with Wald statistics
in conventional likelihood-based inference. Alternatives are to base inference upon the
asymptotic normality of the IEEs themselves, by analogy with score statistics; or upon the
“independence” log-likelihood that defines the IEEs. In the latter context Rotnitzky and
Jewell (1990) showed that, in the presence of clustered data and under the usual conditions
permitting the use of quadratic approximations, the asymptotic null distribution of the
“independence” log likelihood ratio test statistic is that of a weighted sum of chi-squared
random variables.

In this paper, we revisit the problem of accounting for clustering in inference based on the independence log-likelihood. However, rather than following the Rotnitzky and Jewell (1990) approach of retaining the test statistic and adjusting the null distribution, we propose to modify the former and preserve the latter. Specifically, we adjust the independence log-likelihood function in such a way that, if the adjusted function is treated as a bona fide log-likelihood, the resulting score and Wald tests correspond with those obtained using IEEs. In some simple situations, our adjustment in fact recovers the true log-likelihood function. The procedure has some theoretical advantages over the Rotnitzky and Jewell (1990) approach, which have been verified in simulations; it also has some computational advantages, since the asymptotic distributional properties are simpler.

In the next section, we define notation and briefly review the asymptotic theory for IEEs. Section 3 develops the proposed adjustment to the independence log-likelihood; results of simulation studies are presented in Section 4, and the methodology is applied to two examples in Section 5. The work is summarised in Section 6.

2 Independence estimating equations

We consider $k$ clusters of observations $\{y_j : j = 1, \ldots, k\}$, where $y_j = (y_{1j}, \ldots, y_{mj})'$ is the vector of $m_j$ observations in the $j$th cluster. For each $j$, denote by $C_j$ a conditioning set relevant to $y_j$. We allow $y_i \in C_j$ for $i < j$, but not for $i \geq j$. The observations are assumed drawn from a parametric family of joint distributions, indexed by parameter vectors $\theta$ and $\alpha$ and with densities factorised as

$$\prod_{j=1}^{k} f_j(y_j | C_j; \theta, \alpha) .$$

(1)

Conditionally upon $C_j$, the densities $\{f_{ij} : i = 1, \ldots, m_j\}$ of the individual observations in the $j$th cluster are assumed to depend on $\theta$ but not on $\alpha$. Hence $\alpha$ parameterises within-cluster dependence, and $\theta$ parameterises the marginal structure. We are solely concerned with inference about $\theta$, which has $p$ elements.

In standard applications such as the analysis of longitudinal studies, clusters are assumed to be independent so that $C_j$ contains only covariate information. However, by allowing $y_i \in C_j$ for $i < j$, more complex data structures can be accommodated. For example, in space-time modelling it is natural to regard ‘clusters’ as groups of observations made at the same time; by including all previous observations into the conditioning set at time $j$, (1) is just the standard time series factorisation of the full joint density.

Given the joint distributions $\{f_j(y_j | C_j; \theta, \alpha) : j = 1, \ldots, k\}$, inference can be based on the log-likelihood function

$$\ell_{FULL}(\theta, \alpha) = \sum_{j=1}^{k} \ln f_j(y_j | C_j; \theta, \alpha) .$$

(2)
Denote by $U^{\text{FULL}}(\theta, \alpha) = \left( [\partial \ell^{\text{FULL}} / \partial \theta]' [\partial \ell^{\text{FULL}} / \partial \alpha]' \right)'$ the score vector of log-likelihood derivatives. Clearly, this can be written as a sum of contributions from each cluster: $U^{\text{FULL}}(\theta, \alpha) = \sum_j U^{(j)}(\theta, \alpha)$, say. Suppose now that the data are generated from the distribution with $\theta = \theta_0$ and $\alpha = \alpha_0$. Then it can be shown (Chandler et al., 2006) that under very general conditions, $E\left[ U^{(i)}_{\text{FULL}}(\theta_0, \alpha_0) \right] = 0$ for $i = 1, \ldots k$, and

$$\text{Cov} \left[ U^{(i)}_{\text{FULL}}(\theta_0, \alpha_0), U^{(j)}_{\text{FULL}}(\theta_0, \alpha_0) \right] = 0 \quad (i \neq j), \quad (3)$$

even if the clusters themselves are dependent. This is a consequence of the between-cluster conditioning built into the factorisation (1).

In general however, the formulation of appropriate joint distributions $\{f_j\}$ is difficult except in a limited number of special cases (for example, when the marginal distributions are normal). Thus, if the goal of the analysis is inference about $\theta$ so that $\alpha$ is of no interest, the use of (2) may be considered impractical. In such situations a natural approach is to proceed as though, given the conditioning sets $\{C_j\}$, the observations are independent: then $\theta$ can be estimated by maximising the “independence” log-likelihood function

$$\ell^{\text{IND}}(\theta) = \sum_{j=1}^k \sum_{i=1}^{m_j} \ln f_{ij}(y_{ij} | C_j; \theta). \quad (4)$$

We restrict attention to problems in which the resulting estimator $\hat{\theta}^{\text{IND}}$ is the unique root of the independence estimating equations

$$U(\theta) = \frac{\partial \ell^{\text{IND}}}{\partial \theta} = \sum_{j=1}^k U_j(\theta) = \sum_{j=1}^k \sum_{i=1}^{m_j} U_{ij}(\theta) = 0, \quad (5)$$

in an obvious notation. Once again, contributions from different clusters to $U(\theta)$ are uncorrelated. To see this, note that the factorisation (1) can be expanded by writing, for example,

$$f_j(y_j | C_j; \theta, \alpha) = f_{ij}(y_{ij} | C_j; \theta) f_j^{(-i)}\left(y_j^{(-i)} | y_{ij}, C_j; \theta, \alpha\right),$$

where $f_j^{(-i)}\left(y_j^{(-i)} | y_{ij}, C_j; \theta, \alpha\right)$ is the joint density, conditionally upon $y_{ij}$, of all the other elements of $y_j$. A full log-likelihood based upon this expanded factorisation will contain a factor $f_{ij}(y_{ij} | C_j; \theta)$, with a corresponding score contribution $U_{ij}(\theta)$; result (3) shows that at $\theta_0$ this must be uncorrelated with the contributions from any other clusters in the expanded factorisation.

The asymptotic properties of the independence estimator $\hat{\theta}^{\text{IND}}$ are well known, and rely on a first-order expansion of the estimating function $U(\theta)$ (equivalently, a quadratic approximation to $\ell^{\text{IND}}(\theta)$). Specifically, in regular problems as $k \to \infty$ (see, for example, Davison 2003, p.147),

$$\hat{\theta}^{\text{IND}} \sim N(\theta_0, H^{-1}VH^{-1}), \quad (6)$$

where $H$ is the expected Hessian of $\ell^{\text{IND}}$ at $\theta_0$ and $V$ is the covariance matrix of $U(\theta_0)$. If there is no within-cluster dependence, $V = -H$ and the covariance matrix in (6) reduces to $V^{-1} = -H^{-1}$. In general however, this simplification does not arise.
In practice, $H$ can be estimated using either its expected or observed value at $\hat{\theta}_{\text{IND}}$; call such an estimate $\hat{H}$. Moreover, since the score contributions from different clusters are uncorrelated, we have $V = \sum_j \text{Var} \left[ U_j (\theta_0) \right] = \sum_j E \left[ U_j (\theta_0) U_j (\theta_0)' \right]$. This suggests estimating $V$ using its empirical counterpart

$$\hat{V} = \sum_{j=1}^k U_j \left( \hat{\theta} \right) U_j' \left( \hat{\theta} \right).$$

(7)

Under the usual regularity conditions (e.g. Davison 2003, p.118) and under the additional assumption that $k^{-1}V$ tends to a finite non-zero limit as $k \to \infty$, both $\hat{H} - H$ and $\hat{V} - V$ are $O(k^{1/2})$ in probability, whereas $H$ and $V$ are $O(k)$. The covariance matrix in (6) can therefore be estimated consistently using

$$R = \hat{H}^{-1} \hat{V} \hat{H}^{-1}.$$  (8)

We follow convention in referring to $R$ as a “robust” variance estimator; it is to be contrasted with the “naive” estimator

$$N = -\hat{H}^{-1},$$

(9)

which fails to account for the within-cluster dependence.

The asymptotic distribution (6) can be used to construct approximate confidence intervals for elements of $\theta$, and to carry out Wald tests of hypotheses. Alternatively, by analogy with classical procedures based on the score function (e.g. Rotnitzky and Jewell 1990), inference can be based directly on $U(\theta_0)$ using the asymptotic result

$$U(\theta_0) \sim N(0, V).$$

(10)

A third possibility is to base inference directly upon the independence log-likelihood (4) itself. The usual procedure for this relies on a modification of the usual chi-squared approximation for the likelihood ratio statistic which, in the context of longitudinal studies, has been given by Rotnitzky and Jewell (1990). The modification can in fact be regarded as an application of standard results on model misspecification (e.g. Kent 1982, Theorem 3.1). Specifically, suppose that $\theta = (\phi' \psi')'$, with target value $(\phi_0' \psi_0')'$, and that $\psi$ contains $q$ elements. Denote by $\hat{\theta}_{\text{IND}}$ the restricted independence estimator of $\theta_0$, obtained by maximising (4) over $\theta$ with $\psi$ held fixed at $\psi_0$. Also, let $[H^{-1}VH^{-1}]_{\psi}$ and $[H^{-1}]_{\psi}$ be the submatrices of $H^{-1}VH^{-1}$ and $H^{-1}$ corresponding to the elements of $\psi$. Then for large $k$, the distribution of

$$\Lambda_{\text{IND}} = 2 \left[ \ell_{\text{IND}} \left( \hat{\theta}_{\text{IND}} \right) - \ell_{\text{IND}} \left( \hat{\theta}_{\text{IND}} \right) \right]$$

(11)

is approximately the same as that of $\sum_{i=1}^q \lambda_i X_i$, where $\lambda_1, \ldots, \lambda_q$ are the eigenvalues of $Q = \left[ H^{-1}VH^{-1} \right]_{\psi} \left\{ -[H^{-1}]_{\psi} \right\}^{-1}$ and $X_1, \ldots, X_q$ are independent $\chi^2_1$ random variables. $Q$ is estimated consistently by $R_{\psi}N_{\psi}^{-1}$, in an obvious notation; this can be used to test hypotheses about, and to construct confidence regions for, $\psi_0$. It is slightly unsatisfactory that the distribution is not the same for all $q$-dimensional subvectors $\{\psi\}$: for example, confidence intervals for individual parameters, computed on the basis of (11), will
each be defined in terms of a different threshold. Moreover, the distribution function of 
\[ \sum_{i=1}^{q} \lambda_i X_i \] is difficult to evaluate exactly, so Rotnitzky and Jewell (1990) suggested treating 
\[ q \Lambda_{IND} / \sum_{i=1}^{q} \lambda_i \] as a \( \chi^2 \) random variable. A better approximation can be obtained 
using results on quadratic forms of normal random variables, as detailed in Bowman and 

A potential concern with the use of independence estimators is that they may be
inefficient by comparison with estimators that account correctly for the within-cluster
dependence. However, as discussed in Section 1 above, there is substantial empirical
evidence that the loss of efficiency is often small. In addition, theoretical results are
known in a few special but important cases. For example, in the linear model

\[ Y = X\beta + u \quad \text{with} \quad E(uu') = \Gamma, \] (12)

the independence estimator of \( \beta \) is identical to the best linear unbiased estimator (and
hence is fully efficient if \( u \) is Gaussian) if and only if the columns of \( X \) span the same
subspace as those of \( \Gamma X \) (Watson, 1972). The worst-case scenario for independence esti-
mators under this model has been studied by Bloomfield and Watson (1975).

3 Adjusting the independence log-likelihood

The previous section summarised the available analogues of Wald, score and likelihood
ratio statistics for carrying out inference based on IEEs. In classical theory, the three
approaches are asymptotically equivalent (Cox and Hinkley, 1974, Section 9.3). However,
Wald tests lack invariance to parameter transformations, and score tests are rarely used
in practice. Wald tests, in the guise of \( t- \) or \( z- \) tests for regression coefficients, can also
be difficult to interpret in problems with highly correlated covariates (Chandler, 1998),
although these difficulties largely disappear if the tests are used solely to compare models
within a nested sequence. Results such as the Neyman-Pearson lemma suggest that in
finite samples, inference based on the likelihood ratio is in some sense optimal. This
motivates further study of likelihood ratios in the context of IEEs.

In this section, we propose an alternative to Rotnitzky and Jewell’s (1990) likelihood
ratio statistic (11). Our approach is similar to theirs insofar as it is based on the inde-
pendence log-likelihood (4). However, whereas they adjusted the null distribution of
the classical test statistic, we propose to adjust the independence log-likelihood itself, in
such a way that the usual asymptotic \( \chi^2 \) null distribution is preserved. The basic idea
is to stretch the independence log-likelihood about the independence estimate, thereby
changing its curvature but leaving other aspects of its shape unchanged.

3.1 Derivation of the adjustment

We start by observing that in the absence of clustering, all of the usual large-sample
properties of the likelihood ratio rely on the identity \( V = -H \) (Davison, 2003, p.118)
or, equivalently, on the fact that the covariance matrix of \( \hat{\theta}_{IND} \) in (6) is \( -H^{-1} \). This
forms the basis for the naive covariance matrix estimator (9) which, in the presence of
clustering, should be replaced by the robust estimator (8). If, therefore, we define an adjusted inference function \( \ell_{\text{adj}}(\theta) \), which is maximized with zero derivative at \( \hat{\theta}_{\text{ind}} \) and which has Hessian \( \hat{H}_{\text{adj}} = -R^{-1} \) there, we will recover the identity between the Hessian and covariance matrix, and can appeal directly to all of the usual theory.

These criteria define the first and second order properties of the adjusted inference function \( \ell_{\text{adj}}(\theta) \). We complete its specification by taking its profile from the independence log-likelihood so as to preserve any marked asymmetries. Specifically, we define

\[
\ell_{\text{adj}}(\theta) = \ell_{\text{ind}}(\theta^*) , \quad \text{where} \quad \theta^* = \hat{\theta}_{\text{ind}} + C (\theta - \hat{\theta}_{\text{ind}})
\]

for some \( p \times p \) matrix \( C \). Since \( \hat{\theta}_{\text{ind}} \) solves (5), it also maximizes \( \ell_{\text{adj}}(\theta) \) with zero derivative. Further, at \( \theta_{\text{ind}} \) the Hessian of \( \ell_{\text{adj}}(\theta) \) is \( \hat{H}_{\text{adj}} = C'HC \). Any desired value of \( \hat{H}_{\text{adj}} \) (here, \( -R^{-1} \)) can therefore be obtained by defining

\[
C = M^{-1}M_{\text{adj}} ,
\]

where \( M'M = \hat{H} \) and \( M'_{\text{adj}}M_{\text{adj}} = \hat{H}_{\text{adj}} \). If \( p > 1 \) the matrix square roots \( M \) and \( M_{\text{adj}} \) are not unique; we discuss their choice below. For the moment, however, note that the Hessian \( \hat{H}_{\text{adj}} \) defines a quadratic approximation to the adjusted inference function, and that this quadratic approximation is independent of the matrix square roots chosen. In particular, if \( \ell_{\text{ind}} \) is globally or locally (in the neighbourhood of \( \hat{\theta}_{\text{ind}} \)) quadratic, the choice of square root is immaterial. This holds for Gaussian models, as well as for suitably large data sets in all regular problems.

### 3.2 Features of the adjustment

Figure 1 illustrates the construction of the adjustment in the case where \( \theta \) is a scalar. In this case, the matrix \( C \) is just the ratio of the robust to naive standard errors, and the adjustment stretches out the independence log-likelihood horizontally by this factor. Since the adjustment is defined so as to preserve the usual asymptotics, large-sample confidence intervals for \( \theta_0 \) can be defined by comparing \( 2 \left[ \ell_{\text{adj}}(\hat{\theta}_{\text{ind}}) - \ell_{\text{adj}}(\theta) \right] \) directly with quantiles of the \( \chi^2_1 \) distribution; Figure 1 illustrates the construction of a 95% confidence interval in this way. This contrasts with the approach of Rotnitzky and Jewell (1990), who work directly with \( \ell_{\text{ind}}(\theta) \) (shown in black in Figure 1) but modify its distribution. In this case, their modification is simply to multiply the \( \chi^2_1 \) distribution by \( \lambda \), the ratio of the robust to naive variances. The two approaches will usually yield similar results for a scalar parameter, and are identical when the independence log-likelihood is quadratic.

The differences become more pronounced, however, for vector parameters. To illustration this, consider \( T \) bivariate normal pairs \( \{y_j = (y_{1j}, y_{2j})' : j = 1, \ldots, k\} \) with unknown mean \( \mu \), unit variances and known correlation matrix \( \Sigma \). In this case the independence log-likelihood for \( \mu \) is

\[
\ell_{\text{ind}}(\mu) = -\frac{1}{2} \sum_{j=1}^{k} (y_j - \mu)'(y_j - \mu) = -\frac{k}{2} \overline{y}'(\overline{y} - \mu) + \text{constant}.
\]


Figure 1: 95% confidence intervals for a scalar parameter $\theta$ derived from the independence log-likelihood. The interval $(L_{RJ}, U_{RJ})$ is obtained by modifying the classical $\chi^2$ distribution of the likelihood ratio as in Rotnitzky and Jewell (1990); the interval $(L_{new}, U_{new})$ is obtained from the adjusted inference function $\ell_{ADJ}(\theta)$.

The score contribution $U_j(\mu)$ is $y_j - \mu$, so that $V = \text{Var}[U(\mu_0)] = k\Sigma$; the independence estimate is $\hat{\mu}_{IND} = \bar{y}$. In this example, the naive and robust covariance matrices of $\hat{\mu}$ are known and do not need to be estimated: the former is $k^{-1}I_{2 \times 2}$, and the latter is $k^{-1}\Sigma^{-1}$. Setting $M = k^{-1/2}I_{2 \times 2}$ (or any other matrix square root of $k^{-1}I_{2 \times 2}$) and $M_{ADJ} = k^{-1/2}\Sigma^{-1/2}$, it is straightforward to show that the adjusted inference function defined by (13) and (14) is

$$\ell_{ADJ}(\mu) = -\frac{k}{2} (\bar{y} - \mu) \Sigma^{-1} (\bar{y} - \mu) + \text{constant}.$$ 

Apart from constant terms, this is in fact the true bivariate log-likelihood for this particular problem. Confidence regions for $\mu$, defined using $\ell_{ADJ}(\mu)$, will therefore have the correct elliptical shape. Notice, however, that confidence regions obtained directly from $\ell_{IND}$ will always be circular. This suggests that the new method may outperform that of Rotnitzky and Jewell (1990) in vector parameter situations.

In the bivariate normal example, the proposed adjustment recovers the correct log-likelihood exactly. The requirements for this in general are (a) that the log-likelihood is quadratic, and (b) that the independence and maximum likelihood estimators are the same. The former condition clearly holds in Gaussian linear models with known covariance matrix, and the latter also holds for a subset of these models as discussed in Section 2.
above. In such models, when the covariance matrix is unknown, the proposed adjustment substitutes an estimate based on (7), which differs from the actual covariance matrix by an amount that is \(O\left(k^{-1/2}\right)\) in probability. It is then straightforward to show that for any \(\theta\) with \(\theta - \hat{\theta}_{IND} = O\left(k^{-1/2}\right)\) (i.e. in the region of interest for most purposes), the corresponding adjusted inference function differs from the full log-likelihood with known covariance by an amount that is also \(O\left(k^{-1/2}\right)\) in probability. Hence, for this class of models, inference using \(\ell_{ADJ}(\theta)\) is asymptotically equivalent to that using the full log-likelihood.

### 3.3 Choice of matrix square roots

So far we have not addressed the non-uniqueness of the matrix square roots \(M\) and \(M_{ADJ}\) in (14). As discussed previously, if the independence log-likelihood is nearly quadratic the exact choice is immaterial, since the second-order approximations involved are unique. In such situations it seems reasonable to use Cholesky square roots, since these are computationally cheap and stable (Press et al., 1992, Section 2.9). When \(\ell_{IND}(\cdot)\) is asymmetric however, it may be worth ensuring that the transformation from \(\theta\) to \(\theta^*\) preserves the directions of asymmetry as far as possible so that \(\ell_{ADJ}(\cdot)\) can still be thought of as measuring the plausibility of different parameter values. It is therefore of interest to identify matrix square roots that will, in some sense, preserve directions in the parameter space by minimising rotation about \(\hat{\theta}_{IND}\).

The mapping from \(\theta\) to \(\theta^*\) can usefully be regarded as a succession of transformations: the ellipsoidal contours of \(\ell_{ADJ}(\cdot)\) are mapped to spheroids, which are then transformed back to the contours of \(\ell_{IND}(\cdot)\) (in this context, the non-uniqueness of the matrix square roots corresponds to the possibility of arbitrarily rotating the spheroids before the back-transformation). A "minimal rotation" transformation can therefore be defined as a pair of dilations along the ellipsoid axes. The corresponding matrix square roots are

\[
M = LDL^1/2L' \quad \text{and} \quad M_{ADJ} = L_{ADJ}D_{ADJ}^{1/2}L_{ADJ}' ,
\]

where \(LDL'\) and \(L_{ADJ}D_{ADJ}L_{ADJ}'\) are the spectral decompositions of of \(\hat{H}\) and \(\hat{H}_{ADJ}\) respectively. \(D^{1/2}\) and \(D_{ADJ}^{1/2}\) here are diagonal matrices containing the square roots of the eigenvalues of \(\hat{H}\) and \(\hat{H}_{ADJ}\) respectively.

In Section 5.2 below, we present an application in which the independence log-likelihood is asymmetric; this provides an opportunity to explore the sensitivity of results to the choice of matrix square roots.

### 3.4 Comparison of nested models

We now consider the use of \(\ell_{ADJ}(\theta)\) to compare nested models. The setup is slightly more general than that considered in Section 2, in that we now define the restricted model via a set of \(q \leq p\) linear constraints: \(\Delta \theta = \delta_0\) say, where \(\Delta\) is a \(q \times p\) matrix of rank \(q\). Recall that the adjustment has been defined so as to preserve all of the usual asymptotics.
Therefore, the restricted and unrestricted models can be compared using a $\chi^2_q$ distribution for the adjusted likelihood ratio statistic

$$\Lambda_{\text{ADJ}} = 2 \left[ \ell_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{ADJ}} \right) \right], \quad (17)$$

where $\tilde{\theta}_{\text{ADJ}}$ maximises $\ell_{\text{ADJ}} (\theta)$ subject to $\Delta \theta = \delta_0$.

Often, however, one may wish to avoid the direct computation of $\tilde{\theta}_{\text{ADJ}}$. We have in mind a situation in which the unrestricted and restricted models have both been fitted using the independence log-likelihood, to yield estimates $\hat{\theta}_{\text{IND}}$ and $\tilde{\theta}_{\text{IND}}$ respectively. In this case, an additional optimisation to find $\tilde{\theta}_{\text{ADJ}}$ may add significantly to the overall computational cost of model comparison, particularly when dealing with large data sets. In such circumstances it is natural to consider substituting $\tilde{\theta}_{\text{IND}}$ for $\tilde{\theta}_{\text{ADJ}}$ in (17), since the evaluation of $\ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right)$ requires only a single extra pass through the data set. However, since by definition $\ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) \leq \ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{ADJ}} \right)$, such a substitution will inflate the value of $\Lambda_{\text{ADJ}}$; this may lead, for example, to liberal tests of hypotheses based on the $\chi^2_q$ distribution.

To overcome this problem, when computation of $\tilde{\theta}_{\text{ADJ}}$ is to be avoided we propose the following alternative to (17):

$$\Lambda^*_\text{ADJ} = 2c \left[ \ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} \right) \right], \quad (18)$$

where

$$c = \frac{\left( \Delta \hat{\theta}_{\text{IND}} - \delta_0 \right)' \left[ \Delta H^{-1}_{\text{ADJ}} \Delta \right]^{-1} \left( \Delta \hat{\theta}_{\text{IND}} - \delta_0 \right)}{\left( \hat{\theta}_{\text{IND}} - \tilde{\theta}_{\text{IND}} \right)' \hat{H}_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} - \tilde{\theta}_{\text{IND}} \right)} \quad (19)$$

is the minimised value of

$$\frac{\left( \tilde{\theta}_{\text{IND}} - \theta^i \right)' \hat{H}_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} - \theta^i \right)}{\left( \hat{\theta}_{\text{IND}} - \tilde{\theta}_{\text{IND}} \right)' \hat{H}_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} - \tilde{\theta}_{\text{IND}} \right)} \quad (20)$$

with respect to $\theta^i$, subject to the restriction $\Delta \theta^i = \delta_0$ (this can be shown as a standard exercise in constrained minimisation using Lagrange multipliers). To justify this proposal, note that (20) is a ratio of quadratic approximations to $\ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} (\theta^i)$ and $\ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} \right)$ respectively. The minimised value of (20), under the constraint $\Delta \theta^i = \delta_0$, therefore approximates the ratio

$$\left[ \ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} \right) \right] / \left[ \ell_{\text{ADJ}} \left( \tilde{\theta}_{\text{IND}} \right) - \ell_{\text{ADJ}} \left( \hat{\theta}_{\text{IND}} \right) \right].$$

Hence $\Lambda^*_\text{ADJ}$, as defined by (18), is an approximation to (17). The order of magnitude of the approximation error is the same as that of the quadratic approximations in (20) i.e. $k^{-1/2}$ in regular problems.
In most cases of interest, the restricted model will be defined by a constraint of the form \( \psi = \psi_0 \) as in Section 2. In such cases the numerator of (19) takes the form

\[
\left( \hat{\psi}_{IND} - \psi_0 \right)' [-R_\psi]^{-1} \left( \hat{\psi}_{IND} - \psi_0 \right),
\]

where \( \hat{\psi}_{IND} \) is the unrestricted independence estimate of \( \psi \) and \( R_\psi \) is the corresponding submatrix of \( R \) as before.

## 4 Simulation studies

In this section the performance of the adjustment, in a hypothesis testing setting, is compared with established techniques using simulation. The techniques selected for comparison are the robust Wald test based on (6), the robust score test based on (10) and Rotnizky and Jewell’s test based on the independence likelihood ratio statistic (11). In all computations below, the null distribution for the latter test has been computed using the approximation given in Bowman and Azzalini (1997, pp.86–88). The adjustment has been implemented via the test statistic (18), and Cholesky square roots have been used to define \( C \) in (14).

The simulation environment is similar to that used by Fitzmaurice (1995), and more recently by Pan (2001). There are \( k \) clusters of data, each consisting of three measurements at times 0, 1 and 2. The responses \( \{ y_{ij} : i = 1, \ldots, 3; j = 1, \ldots, k \} \) are binary, with marginal means \( \{ \mu_{ij} \} \) given by a logistic regression model of the form

\[
\logit (\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}.
\]

Here, \( x_{ij1} \) is a Bernoulli random variable, and \( x_{ij2} = (j - 1) \) represents a linear trend within each cluster. We consider two different mechanisms for generating \( x_{ij1} \), as follows:

1. \( x_{ij1} \) is fixed within each cluster, with success probability 0.5. This represents group membership in a randomised trial, for example.

2. \( x_{ij1} \) has success probability \( 0.3j - 0.1 \), independently for each \( j = 1, 2, 3 \). This induces a correlation of \( 0.6 \sqrt{2/3} = 0.490 \) between the two covariates \( x_{.1} \) and \( x_{.2} \), while preserving the marginal distribution of \( x_{.1} \).

An exchangeable correlation structure is assumed for the responses, with all inter-cluster correlations set to 0.5. In all of the simulations reported below, the representation of Bahadur (1961) has been used to generate the correlated binary responses and the value of \( \beta_0 \) has been held fixed at 0.25.

The various test procedures are compared using power curves for testing nested hypotheses at level 0.05. Two specific tests are considered: \( H_0 : \beta_2 = 0 \) versus \( H_1 : \beta_2 \neq 0 \), and \( H_0 : \beta_1 = \beta_2 = 0 \) versus its complement. The power curves are computed by varying \( \beta_1 = \beta_2 \) together, over the range -0.6 to 0.6 in steps of 0.05. At each value of \( (\beta_1, \beta_2) \), 1000 simulations are performed and the proportion of rejections of \( H_0 \) is recorded. The robust Wald and score tests are implemented throughout in the context of the full model with both covariates included. Similarly, the adjusted likelihood ratio tests all use comparisons between the full model and the corresponding reduced models — hence the hypothesis \( H_0 : \beta_2 = 0 \) is never tested by fitting the model \( \logit (\mu_{ij}) = \beta_0 + \beta_2 x_{ij2} \), for example.
Figure 2: Power curves for each of the four competing tests at level 0.05, for testing (top) \(H_0 : \beta_2 = 0\); (bottom) \(\beta_1 = \beta_2 = 0\), in model (21) when \(x_{ij1}\) represents group membership. Here there are \(k = 100\) clusters of size 3. In the top plot, when \(\beta_2 = 0\), the simulated rejection rates are 0.054, 0.054, 0.054 and 0.053 respectively. The corresponding figures in the bottom plot are 0.041, 0.048, 0.041 and 0.040.

Results, for the case when there are \(k = 100\) clusters and \(x_{ij1}\) is fixed within each cluster, are shown in Figure 2; these are almost exactly as expected from the theory. Note first that when \(\beta_2 = 0\) in both plots, all of the tests yield a rejection rate extremely close to the nominal level of 0.05. In the top plot, where a single parameter is being tested, the four procedures are indistinguishable; it is likely that 100 clusters yield an independence log-likelihood that is almost quadratic, so that the asymptotic equivalence of the procedures is effectively achieved here. In the bottom plot, however, the Rotnitzky and Jewell test based on (11) clearly has much lower power than the other three. An explanation for this can be found in the bivariate normal example of Section 3: when more than one parameter is being tested, the acceptance region for a test based on (11) will usually be the “wrong” shape.

Figure 3 shows more results, for a case in which the covariates are correlated and there are fewer \((k = 40)\) clusters. The most obvious change from Figure 2 is that there are now small differences between the new test procedure, robust Wald and robust score tests in the bottom plot. These result from a combination of smaller sample size and correlated covariates. In this particular example, the robust score test loses power slightly relative
Figure 3: Power curves for each of the four competing tests at level 0.05, for testing (top) $H_0 : \beta_2 = 0$ (bottom) $\beta_1 = \beta_2 = 0$, in model (21) when $x_{ij1}$ is correlated with $x_{ij2}$. Here there are $k = 40$ clusters of size 3. In the top plot, when $\beta_2 = 0$, the simulated rejection rates are 0.044, 0.043, 0.041 and 0.045 respectively. The corresponding figures in the bottom plot are 0.052, 0.047, 0.046 and 0.038.

to both the new procedure and the robust Wald test.

Further simulation studies have been carried out (Bate, 2004) involving more complex data structures, for example those that arise in the space-time modelling of daily climatological data (see Section 5 below). The conclusions are identical to those reported above: in particular, the proposed adjustment outperforms the test based on (11) when testing more than one parameter simultaneously. We have not found a case in which it performs worse than any of the other methods considered.

5 Examples

In this section we present two applications of the proposed adjustment to problems arising in climatology. The first involves the space-time modelling of daily wind speed in northwestern Europe, and the second involves an extreme value analysis of annual maximum temperatures at two locations in England.
5.1 Modelling of European wind speed

In the developed world, wind storms are responsible for substantial economic losses. It is therefore important to understand how wind speed in a particular region is associated with large-scale atmospheric structures, so as to be able to translate the likely effect of changes in such structures into statements about a change in the risk of extreme wind speed. Motivated by such considerations, Yan et al. (2002) developed a model in which daily wind speeds at 120 locations in Europe were related to 19 different indices of large-scale climatic structure. These indices included the North Atlantic Oscillation index (a measure of the atmospheric pressure gradient over the North Atlantic, and hence in some sense a measure of dynamical instability in this region), the Southern Hemisphere temperature and the Southern Oscillation Index (an index that is closely related to the El Niño phenomenon). The data are for the period 1958–1998; there are around 1.9 million observations in total in the data set.

For this example, each day yields a cluster of 120 dependent measurements at the different spatial locations. This structure falls within the framework considered in Section 2, providing the conditioning set $C_j$ for day $j$ includes previous days’ wind speeds so as to account for the space-time nature of the data. Yan et al. (2002) fitted a GLM with gamma marginal distributions and a log link, to relate the daily wind speeds to the atmospheric indices (see also Chandler 2005 for a review of the use of GLMs to represent the complex data structures often found in such problems). Their final model contained 110 terms, representing a combination of previous days’ wind speeds, seasonal effects, terms representing the effect of spatial location, large-scale climate indices and interactions (to account for the fact that the effect of large-scale climate indices will often vary with location and with time of year, for example). Most of the covariates related to the representation of regional variability using orthogonal series.

In Yan et al. (2002), model selection was carried out through the use of a combination of “naive” likelihood ratio tests (i.e. without adjusting for inter-site dependence) and climatological judgement. The justification for this was that when analysing huge data sets, formal tests are almost guaranteed to reject a simple model in favour of a more complex one, even if in practical terms the improvement in fit is negligible; hence the naive likelihood ratio tests were used to guide, rather than to dictate, the model-building process. Nonetheless, it is of interest to revisit this model using the adjusted likelihood ratio test proposed here. We therefore take the final model of Yan et al. (2002) and, for each climate index individually, delete all terms associated with that index. We then carry out formal tests of the reduced versus extended model in each case, using both naive and adjusted likelihood ratio tests. For the adjusted test we again use (18), with Cholesky square roots to define $C$ in (14).

Table 1 shows the results. The adjustment makes a big difference to the test statistics in many cases, but the effects of most indices remain significant by any reasonable standard. This might be expected given the huge sample size: however, note that the last few indices are no longer significant after the adjustment. Most of these non-significant indices are geographically remote from the study area of northwestern Europe, so their elimination makes sense on physical grounds — the arguments offered by Yan et al. (2002) in support of their inclusion were somewhat speculative. However, some apparently remote
Table 1: Effect of dropping climate indices from the European wind speed model of Yan et al. (2002), according to naive and adjusted likelihood ratio test procedures. \( p \)-values less than \( 10^{-15} \) are reported as zero.

indices, such as the Southern Oscillation index and Southern Hemisphere temperature (SHT), remain highly significant so that these associations, and the mechanisms conjectured to be responsible for these associations, are to some extent confirmed.

5.2 Extreme value modelling of maximum temperatures

For our second example, we consider a bivariate time series of annual maximum temperatures, from Oxford and Worthing in the UK, from 1901 to 1980. These data have previously been considered by various authors in the extreme value literature; see, for example, Coles et al. (1999). The data are in degrees Fahrenheit. Neither series exhibits any obvious trends or autocorrelation, but they are moderately interdependent (the correlation between them is 0.62).

In the absence of temporal structure, these data can be considered as 80 independent observations \( y_1, \ldots, y_{80} \) from a bivariate distribution. Since the data are annual maxima, it is natural to consider modelling using generalized extreme value (GEV) distributions; see Coles (2001) for example. Accordingly, for \( i = 1, 2 \) and \( j = 1, \ldots, 80 \) we consider \( y_{ij} \) to be drawn from a GEV distribution with parameters \((\mu_{ij}, \sigma_{ij}, \xi_{ij})\). To accommodate differences between the two sites, we set

\[
\begin{align*}
\mu_{ij} &= \mu_0 + \mu_1 x_i \\
\sigma_{ij} &= \sigma_0 + \sigma_1 x_i \\
\xi_{ij} &= \xi_0 + \xi_1 x_i ,
\end{align*}
\]

where \( x_1 = 1 \) (for data from Oxford) and \( x_2 = -1 \). Thus \( \mu_0, \sigma_0 \) and \( \xi_0 \) are “central”
Table 2: Estimated parameters and standard errors for GEV models fitted to annual maximum temperatures from Oxford and Worthing. The dependence parameter in the bivariate model is excluded from this table.

parameters, while \( \mu_1, \sigma_1 \) and \( \xi_1 \) represent differences between sites.

This particular example has been chosen, partly because there is currently a substantial amount of research into the analysis of dependent extremes (e.g. Heffernan and Tawn 2004 and references therein; also Chavez-Demoulin and Davison 2005, who adopt a bootstrap-based approach to inference in such situations), but also because it provides an opportunity to explore the use of the proposed methodology in a situation where the independence log-likelihood is asymmetric. Recall that the non-uniqueness of the matrix square roots in (14) is potentially problematic unless the independence log-likelihood is close to quadratic: in this example, it is therefore of interest to compare the results obtained using Cholesky square roots with those using the “minimal rotation” square roots defined in (16). Furthermore, the availability of tractable models for bivariate extremes enables a comparison between the adjusted inference function proposed here and the log-likelihood based on a fully specified bivariate model.

To visualise clearly any differences between the methods, we focus upon profile-based confidence regions for selected pairs of parameters in the marginal model. For each pair we compute profiles from (a) the unadjusted independence log-likelihood; (b) the new inference function defined using Cholesky square roots; (c) the new inference function defined using (16); (d) the log-likelihood from a bivariate logistic distribution with GEV margins, which has been used by Coles et al. (1999) to model these same data. All of the optimisation is carried out using the general-purpose optimisation routine \texttt{nlm} in \texttt{R} (R Development Core Team, 2003). We then plot 95% confidence regions, based on the usual \( \chi^2_2 \) approximations.

The estimates and standard errors for the marginal model parameters are presented in Table 2. There is generally close agreement between the independence estimates and those based on the bivariate log-likelihood; the largest discrepancy, in relative terms, is for the parameter \( \xi_0 \). As expected, the naive standard errors appear too low for \( \mu_0 \) and \( \sigma_0 \), and too high for \( \mu_1, \sigma_1 \) and \( \xi_1 \); the robust standard errors agree closely with those from the bivariate log-likelihood. It is surprising that the robust standard error for \( \xi_0 \) is lower than its naive counterpart. However, a “naive empirical” standard error, calculated using (8) but without accounting for the clustering so that \( \hat{\Sigma} \) is defined as
Figure 4: Parameter estimates and 95% confidence regions for selected parameter pairs in the GEV model fitted to annual maximum temperatures from Oxford and Worthing.

\[ \sum_{j=1}^{80} \sum_{i=1}^{2} U_{ij} (\hat{\theta}) U'_{ij} (\hat{\theta}), \] yields the value 0.0379. The difference between this and the model-based naive standard error in Table 2 suggests either that there may be a problem with the model, or that the sample size is not large enough to appeal to asymptotics regarding the parameter \( \xi_0 \).

Figure 4 shows the parameter estimates and confidence regions for pairs \((\mu_0, \mu_1)\), \((\sigma_0, \sigma_1)\) and \((\sigma_0, \xi_0)\). For the first pair, the independence estimate is very close to that obtained from the bivariate model. All of the confidence regions are elliptical and, as expected, there is virtually no difference between the Cholesky and rotation-free adjustments. Moreover, the regions defined by the new inference function are very similar to that from the bivariate model. The results for \((\sigma_0, \sigma_1)\) are similar, even though the confidence regions are no longer elliptical. The pair \((\sigma_0, \xi_0)\) shows marked asymmetry in the profiles, and the difference between the independence and bivariate estimates of \(\xi_0\) is clear. Despite this, the difference between the regions defined by Cholesky and rotation-free adjustments remains negligible. This suggests that even in moderately asymmetric cases the choice of matrix square root may be relatively unimportant.

6 Summary and conclusions

We have proposed a simple adjustment to the independence log-likelihood, for use with cluster correlated data when interest lies in the parameters of the marginal distributions. We have also shown that standard techniques can be applied to dependent clusters, providing the dependence is accounted for via the factorisation (1). This provides an opportunity to use clustered data techniques for the analysis of space-time data, for example.
The idea behind the new inference function is to stretch out the independence log-likelihood, until its Hessian is consistent with the robust covariance matrix at the independence estimate. This procedure recovers all of the usual likelihood asymptotics by construction, and is asymptotically equivalent to analysis based on a fully specified likelihood in some (admittedly special) situations. For the comparison of nested models, an easily computed secondary adjustment can be applied to avoid carrying out a separate constrained maximisation. Simulations confirm that in a hypothesis testing situation, the new inference function performs at least as well as robust Wald and score tests. It also outperforms procedures based on adjusting the critical value of the naive likelihood ratio statistic, when testing more than one parameter simultaneously. This is expected since the contours of the new inference function are designed to be the “correct” shape. A further advantage is ease of computation: the naive and robust variance estimates are widely available in standard packages for the analysis of clustered data, and the null distribution of the adjusted likelihood ratio test statistic is much simpler to calculate than that of the unadjusted statistic.

Potential difficulties with the proposed method include its reliance on independence estimators, which may be inefficient in some situations with strong inter-cluster dependence; also, the non-uniqueness of the matrix square roots in (14) (although this is only likely to be a problem if the independence log-likelihood is markedly asymmetric, and the extremes example in Section 5 suggests that even in this case the results may be insensitive to the square root used).

Outside the rather special class of Gaussian linear models considered in Section 3, in general the adjusted inference function will not recover the full log-likelihood either exactly or asymptotically. It is, however, of interest to determine whether, in any given situation, there exists a joint density of the form (1) for which \( \ell_{\text{adj}} (\theta) \) is asymptotically equivalent to the full log-likelihood for \( \theta \) with \( \alpha \) known. If this is the case and the dependence parameter \( \alpha \) is of no interest, then \( \ell_{\text{adj}} (\theta) \) effectively provides fully likelihood-based inference for \( \theta \), albeit with a very convenient choice of joint distributional structure. Clearly, a necessary condition for this is that the ranges of \( \theta \) and \( \theta^* \) are the same, which in general requires that \( \theta \in \mathbb{R}^p \). A more detailed and complete investigation of this issue will be the subject of future work.

Acknowledgements

Steven Bate’s research was supported by a PhD studentship from the UK Engineering and Physical Sciences Research Council. We are extremely grateful to Valerie Isham for many helpful discussions during the course of this work.

References


