Statistical Modelling 2004; **4**: 161–180

Marginal models for zero inflated clustered data

Daniel B Hall and Zhengang Zhang
Department of Statistics, University of Georgia, Athens, GA, USA

Abstract: Over the last decade or so, there has been increasing interest in 'zero inflated' (ZI) regression models to account for 'excess' zeros in data. Examples include ZI poisson (ZIP), ZI binomial (ZIB), ZI negative binomial and ZI tobit models. Recently, extensions of these models to the clustered data case have begun to appear. For example, Hall considered ZIP and ZIB models with cluster specific random effects. In this paper, we consider an alternative expectation maximization approach on the basis of marginal models and generalized estimating equation (GEE) methodology. In the usual EM algorithm for fitting ZI models, the M step is replaced by the solution of a GEE to take into account within cluster correlation. The details of this approach, including formulas for an asymptotic variance–covariance matrix of parameter estimates, are given for several of the most important ZI regression model classes. Alternatively, GEEs can be applied directly by computing the first two marginal moments of the observed response. We illustrate these two marginal modeling approaches with examples, and compare them via a small simulation study.

Key words: extended generalized estimating equations; finite mixture; generalized linear model; longitudinal data; mixture of experts; repeated measures

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received January 2004; **revised** April 2004; **accepted** June 2004

1 Introduction

Recently, there has been increasing interest in mixture models to account for 'excess' zeros in data. These models, often called zero inflated (ZI) regression models, mix a degenerate distribution with point mass of one at 0 with a simple regression model based on a standard distribution. For example, regression models for zero inflation relative to a Poisson (ZIP models) have been considered by Lambert (1992), Hall (2000), and others; and zero inflated negative binomial (ZINB) models appear in papers by Ridout *et al.* (2001) and others; and zero inflated binomial (ZIB) models are discussed by Hall (2000) and Vieira *et al.* (2000). All of these examples pertain to count data, where zeros occur with positive probability according to the distribution being mixed with zero. The ZI models for count data are useful when zeros are observed more frequently than this probability predicts. Their mixture form accounts for some of the zeros through the non-degenerate distribution (e.g., Poisson) and some through the degenerate (zero) distribution.

Address for correspondence: Daniel B Hall, Department of Statistics, University of Georgia, Athens, GA 30602–1952, USA.

 10.1191/1471082X04st076oa

Of course, large numbers of zeros sometimes occur in continuous data as well, but continuous distributions have a null probability of yielding a zero. Therefore, for independent semi-continuous data, there is little motivation for a model such as a ZI normal, because all observed zeros are unambiguous; they necessarily come from the degenerate distribution, rather than from the nondegenerate continuous distribution. The likelihood for such a model factors into terms for the zero and nonzero data, so that it is equivalent to separately model the nonzero data, and an indicator variable for whether or not the response is zero. Similar comments also apply to count data with excess zeros if the probability of a zero according to the nondegenerate distribution is very small (e.g., the mean is large). In that case it again becomes possible to identify the observed zeros as coming from the zero component so that separate modeling of the positive counts and a zero indicator is justified. The situation in which zero inflation in otherwise continuous data (i.e., semi-continuous data, Olsen and Schafer, 2001) becomes interesting is when there is censoring or truncation in the continuous distribution and/or the data are dependent. For example, Moulton and Halsey (1995) modeled antibody concentrations that were subject to a limit of detection, and were, therefore, left censored with a reported value of zero. However, many more zeros were observed than predicted by left censored versions of standard parametric distributions such as the lognormal. Such a situation motivates the use of a ZI censored regression (tobit) model or ZIT model.

Extensions of basic ZI regression models to correlated data have begun to appear. Hall (2000) extended ZIP and ZIB models to see also the clustered data case by introducing cluster specific random effects into the model (Yau and Lee, 2001). Similarly, a mixed model version of the ZIT model has been proposed by Berk and Lachenbruch (2002). Other mixed models for semi-continuous longitudinal data without censoring have been proposed by Olsen and Schafer (2001) and Tooze *et al.* (2002). As an alternative to the inclusion of random effects, several authors have considered marginal models for clustered data with excess zeros. The approach is either to treat the data as independent during fitting and then use a robust sandwich estimate of the parameter variance–covariance matrix (Moulton *et al.*, 2002), or to incorporate generalized estimating equations (GEEs) with a dependence working correlation matrix into the fitting algorithm (Dobbie and Welsh, 2001).

In this paper, we take the latter approach. However, we differ from Dobbie and Welsh in several respects. These authors consider unbounded count data, and use a mixture model in which zero is mixed with a truncated count distribution such as the Poisson or negative binomial. Such models have been called zero altered by Heilbron (1994) to distinguish them from the ZI models of Lambert (1992) and others. In addition, Dobbie and Welsh utilize GEEs for the observed response, whereas we incorporate GEEs into the expectation maximization (EM) algorithm so that the assumed correlation matrix pertains to the (partially latent) response prior to zero inflation, that is, the response that would have been observed from the nondegenerate distribution had there been no zero inflation. Our approach is based on the work of Rosen *et al.* (2000), who consider mixtures of marginal GLMs for correlated data. The ZI regression models occur as special cases of the class of models Rosen *et al.* considered. We give the details of estimation and inference for overdispersed ZIP models for independent data, and ZIP, ZIB and ZIT models in the clustered case.

The organization of the paper is as follows. Section 2 gives the appropriate background, briefly summarizing ZI models and the approach of Rosen *et al.* to fitting mixtures of marginal GLMs via the expectation solution (ES) algorithm, a generalization of the EM algorithm. In Section 3 we provide details of fitting ZI regression models with the ES algorithm, with attention to each of several important cases. Section 4 discusses an alternative approach based on direct application of GEEs to the observed response. A simulation study is presented in Section 5, which compares this approach with that of Section 3. Real data examples are given in Section 6 to illustrate the methodology.

2 Background

2.1 ZI regression models

In the ZI regression models, a degenerate distribution with point mass of one at zero is mixed with a nondegenerate distribution. Regression structure is built into the model through the mean of the nondegenerate distribution and, possibly, through the mixing probability. The original and most common example is the ZIP regression model, introduced by Lambert (1992). Let $\mathbf{y} = (y_1, \ldots, y_K)^{\mathrm{T}}$ be a vector of observed counts corresponding to independent random variables $Y_1, \ldots, Y_K$. In a ZIP model we assume

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i, \\ \text{Poisson}(\lambda_i) & \text{with probability} 1 - p_i, \end{cases}$$

where the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^{\mathrm{T}}$ and $\mathbf{p} = (p_1, \ldots, p_K)^{\mathrm{T}}$ are assumed to depend on covariates through GLM-like regression specifications. In particular, it is typically assumed that $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$, and $\text{logit}(\mathbf{p}) = \mathbf{G}\boldsymbol{\gamma}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression parameters, and $\mathbf{B}$ and $\mathbf{G}$ are the corresponding model matrices.

Maximum likelihood (ML) has typically been the method of choice for fitting ZI regression models. A particularly convenient method of obtaining the ML estimates is provided by the EM algorithm. The EM algorithm takes advantage of the mixture structure of the model, and allows it to be fit by iteratively fitting weighted versions of simpler generalized linear models. Let $u_i = 1$ when Y_i is drawn from the degenerate zero distribution, and $u_i = 0$ otherwise. Then if we regard $\mathbf{u} = (u_1, \ldots, u_K)^{\mathrm{T}}$ as 'missing data', the 'complete data' loglikelihood for the ZIP model is

$$\begin{aligned} \ell^{\mathrm{c}}(\boldsymbol{\beta}, \boldsymbol{\gamma}\,;\, \mathbf{y}, \mathbf{u}) &= \sum_{i=1}^{K} [u_i \log p_i(\boldsymbol{\gamma}) + (1 - u_i) \log \{1 - p_i(\boldsymbol{\gamma})\}] + \sum_{i=1}^{K} (1 - u_i) \log f_2\{y_i;\, \lambda_i(\boldsymbol{\beta})\} \\ &\equiv \ell^{\mathrm{c}}(\boldsymbol{\gamma};\, \mathbf{y}, \mathbf{u}) + \ell^{\mathrm{c}}(\boldsymbol{\beta};\, \mathbf{y}, \mathbf{u}), \end{aligned} \tag{2.1}$$

where $f_2\{y_i;\lambda_i(\boldsymbol{\beta})\} = \mathrm{e}^{-\lambda_i}\lambda_i^{y_i}/(y_i!)$ is the Poisson probability mass function. The subscript 2 here is to denote that the Poisson distribution is the second component in the mixture model. At the $(h + 1)$th iteration of the EM algorithm, we compute

$Q(\boldsymbol{\beta}, \gamma|\boldsymbol{\beta}^{(h)}, \gamma^{(h)}) = E\{\ell^c(\boldsymbol{\beta}, \gamma; \mathbf{y}, \mathbf{u})|\mathbf{y}, \boldsymbol{\beta}^{(h)}, \gamma^{(h)}\}$ where the expectation is with respect to the distribution of $\mathbf{u}$ given $\mathbf{y}$, $\boldsymbol{\beta}^{(h)}$ and $\gamma^{(h)}$. Since $\ell^c(\boldsymbol{\beta}, \gamma; \mathbf{y}, \mathbf{u})$ is linear with respect to $\mathbf{u}$, this expectation is simply $\ell^c(\boldsymbol{\beta}, \gamma; \mathbf{y}, \mathbf{u}^{(h)})$ where $\mathbf{u}^{(h)} = E(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}^{(h)}, \gamma^{(h)})$, which has ith element

$$u_i^{(h)} = \Pr(u_i = 1|y_i, \boldsymbol{\beta}^{(h)}, \gamma^{(h)}) = 1_{\{y_i=0\}}[1 + \{1 - p_i(\gamma^{(h)})\}f_2(y_i; \boldsymbol{\beta}^{(h)})/p_i(\gamma^{(h)})]^{-1}.$$

Plugging this quantity into Equation (2.1) yields $Q(\boldsymbol{\beta}, \gamma|\boldsymbol{\beta}^{(h)}, \gamma^{(h)}) = \ell^c(\gamma; \mathbf{y}, \mathbf{u}^{(h)}) + \ell^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}^{(h)})$, which is then maximized with respect to $\boldsymbol{\beta}$ and γ in the M step of the algorithm. Note that $Q(\boldsymbol{\beta}, \gamma|\boldsymbol{\beta}^{(h)}, \gamma^{(h)})$ has a particularly convenient form for maximization because γ appears only in the first term $\ell^c(\gamma; \mathbf{y}, \mathbf{u}^{(h)})$, and $\boldsymbol{\beta}$ appears only in the second term $\ell^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}^{(h)})$. The first of these terms has the form of a binomial loglikelihood with response vector $\mathbf{u}^{(h)}$, and the second has the form of a weighted Poisson loglikelihood with weights $1 - u_1^{(h)}, \ldots, 1 - u_K^{(h)}$. Therefore, the M step can be accomplished by fitting two GLMs. Updating of $\mathbf{u}^{(h)}$ (the E step) and fitting of these two GLMs (the M step) are iterated until convergence.

Other ZI regression models have the same structure as the ZIP model, but where the second component density f_2 is that of the negative binomial (yielding the ZINB), binomial (yielding the ZIB), or some other distribution appropriate for the range and scale of the response vector. For example, in the ZIB model we assume $Y_i \sim 0$, with probability p_i, $Y_i \sim \text{Binomial}(\pi_i, n_i)$, with probability $1 - p_i$, where the parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{\mathrm{T}}$ and $\mathbf{p} = (p_1, \ldots, p_K)^{\mathrm{T}}$ are related to covariates via $\text{logit}(\boldsymbol{\pi}) = \mathbf{B}\boldsymbol{\beta}$, and $\text{logit}(\mathbf{p}) = \mathbf{G}\gamma$, and where y_i is a bounded count representing the number of 'successes' out of n_i trials. The EM algorithm for ML estimation remains essentially the same, with the M step consisting of fitting an unweighted binomial GLM to the pseudo-response $\mathbf{u}^{(h)}$ and a weighted GLM based on the density f_2.

For semi-continuous response variables, the loglikelihood under independence is

$$\ell(\boldsymbol{\beta}, \gamma; \mathbf{y}) = \sum_{i=1}^{K} \log[p_i(\gamma)^{1_{\{y_i=0\}}}\{1 - p_i(\gamma)\}^{1-1_{\{y_i=0\}}}] + \sum_{i=1}^{K}(1 - 1_{\{y_i=0\}})\log f_2(\boldsymbol{\beta}).$$

The ML estimation for such a model can be accomplished by separately fitting a binomial regression model to the indicator $1_{\{y_i=0\}}$, $i = 1, \ldots, K$, and a model based on f_2 to the nonzero elements of $\mathbf{y}$. Therefore, under independence there is no need to introduce missing data into the problem to formulate the EM algorithm. Note that this is also true for the zero altered count model which attributes all of the zeros to the first component of the mixture and the nonzero counts to a distribution truncated at zero. The reason for the simplification in these cases is that it is possible to tell from which distribution in the mixture each response comes simply from its value. That is, u_i, the component indicator, is equal to $1_{\{y_i=0\}}$, an observed variable, in the continuous case, but not in the discrete case. However, for clustered data or in the presence of censoring this simplification does not occur even for continuous data.

Suppose that Y_i is a non-negative response variable that is subject to left censoring at a known value $\alpha > 0$. For example, as in Moulton and Halsey (1995), Y_i might be an antibody concentration that is measured subject to a limit of detection so that responses

less than α are recorded as zero. In such a situation, if there is a positive probability for a subject to have a true antibody concentration of zero, then observing $y_i = 0$ is once again ambiguous. Zero inflation in this case leads to a likelihood that does not factor into terms for $\boldsymbol{\beta}$ and γ. In particular, suppose that prior to censoring and zero inflation, the ith response is lognormal with mean μ_i and constant standard deviation σ. Then the loglikelihood for a sample of K observations censored at α (recorded as zero) and subject to excess zeros is

$$\ell(\boldsymbol{\beta}, \sigma, \alpha, \gamma; \mathbf{y}) = \sum_{i=1}^{K} 1_{\{y \le \alpha\}} \log\left(p_i(\gamma) + \{1 - p_i(\gamma)\}\Phi\left[\frac{\log \alpha - \mu_i(\boldsymbol{\beta})}{\sigma}\right]\right) + \sum_{i=1}^{K} (1 - 1_{\{y \le \alpha\}}) \log\left(\frac{\{1 - p_i(\gamma)\}\phi[\{\log y_i - \mu_i(\boldsymbol{\beta})\}/\sigma]}{y_i}\right),$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and probability density functions, respectively. However, if we introduce missing data $\mathbf{u}$, where $u_i = 1$ if y_i is an extra zero and $u_i = 0$ if y_i is drawn from the censored lognormal distribution, then the complete data loglikelihood $\ell^c(\boldsymbol{\beta},\sigma,\gamma;\mathbf{y},\mathbf{u})$ has the same form as in Equation (2.1), but where now f_2 has the form of a censored lognormal density function

$$f_2(y_i; \boldsymbol{\beta}, \sigma, \alpha) = \left[\Phi\left\{\frac{\log \alpha - \mu_i(\boldsymbol{\beta})}{\sigma}\right\}\right]^{1_{\{y_i \le \alpha\}}} \left[y_i^{-1}\phi\left\{\frac{\log y_i - \mu_i(\boldsymbol{\beta})}{\sigma}\right\}\right]^{1 - 1_{\{y_i \le \alpha\}}}. \tag{2.2}$$

Again, the EM algorithm simplifies estimation, allowing the model to be fit by iteratively fitting a binomial GLM and a weighted version of a censored lognormal regression model.

Berk and Lachenbruch (2002) extended the ZI censored lognormal model to the clustered data case by introducing cluster specific random effects into the model. Olsen and Schafer (2001) consider ZI transformed normal (e.g., lognormal) regression models for clustered semi-continuous data without censoring. They also use cluster specific random effects in the model to account for within cluster correlation. This idea was also used by Hall (2000) in the context of the ZIP and ZIB regression models. An alternative approach is to use a marginal model in which nonzero within-cluster correlation is assumed directly, rather than indirectly through the inclusion of random effects. Mixture models based on such an approach were introduced by Rosen *et al.* (2000). We discuss their methodology in the following section.

2.2 Mixtures of marginal GLMs

In the machine learning literature, finite mixtures of GLMs are called mixtures of experts models. In this class of models, the response variable Y_i is assumed to have been generated from one of g component densities of the exponential dispersion family form, with probabilities $p_1, \ldots, p_g$ where $\sum_{i=1}^{g} p_i = 1$. The (conditional) means in the

component densities as well as the mixing probabilities are assumed to depend on covariates through GLM type specifications, that is, through some link function and linear predictor. Rosen *et al.* (2000) describe this class of models and the EM algorithm for fitting such models.

The EM algorithm for the mixtures of experts model has the same outline as described in Section 2.1 for the ZI regression models. Missing data are introduced consisting of a $(g-1)\times 1$ vector of indicators $\mathbf{u}_i$ for the component densities, for each observation $i=1,\ldots,K$. The E step of the algorithm consists of computing the expectation of these indicator variables, and the M step consists of fitting weighted versions of standard GLMs. Note that the score equation in a GLM for independent data has the same form as a GEE (Liang and Zeger, 1986) with independence working correlation matrix. Therefore, to extend the mixtures of experts model to the clustered data setting, Rosen *et al.* suggest replacing the weighted GLM score equations in the M step of the EM algorithm with weighted GEEs with working correlation matrices not necessarily of the independence form. This approach generalizes the EM algorithm to become an expectation-solution (ES) algorithm (Rosen *et al.*, 2000; see also Breckling *et al.*, 1994). They present theory to establish that if the algorithm converges, the resulting parameter estimators solve an unbiased estimating equation, and are therefore consistent and asymptotically normal under suitable regularity conditions.

Of course, the ZIP and ZIB models are special cases of mixtures of experts models with $g=2$ and where the first component density is degenerate. In the following section, we describe in detail the ES algorithm incorporating GEEs to fit marginal ZIP, ZIB, and other ZI regression models for the clustered data.

3 ZI regression via the ES algorithm

To generalize to the clustered data context, we now let $\mathbf{y}_i$ be an $n_i\times 1$ vector of responses for the ith cluster, $i=1,\ldots,K$. In a marginal ZI regression model we assume that Y_{ij}, the random variable associated with observation y_{ij}, follows a ZI distribution: $Y_{ij}\sim 0$ with probability p_{ij}, $Y_{ij}\sim F_2(y_{ij};\theta_{ij},\phi)$, with probability $1-p_{ij}$, where $F_2(y_{ij};\theta_{ij},\phi)$ is an exponential dispersion family distribution with density $f_2(y_{ij};\theta_{ij},\phi)=h_2(y_{ij},\phi)\exp[\{\theta_{ij}y_{ij}-\kappa(\theta_{ij})\}w_{ij}/\phi]$. Here, θ_{ij} is the canonical location parameter, ϕ is a scale parameter and the w_{ij}s are known constants (e.g., binomial denominators). The function κ is a cumulant generating function, so F_2 has (conditional) mean $\zeta_{ij}=\kappa'(\theta_{ij})$ and (conditional) variance $v(\zeta_{ij})\phi/w_{ij}$, where $v(\zeta_{ij})=\kappa''(\theta_{ij})$ is a variance function. We assume the canonical parameters $\boldsymbol{\theta}_i=(\theta_{i1},\ldots,\theta_{in_i})^{\mathrm{T}}$ are related to covariates via GLM type specifications. That is, for canonical link function we have $\boldsymbol{\theta}_i(\boldsymbol{\zeta}_i)=\boldsymbol{\eta}_i=\mathbf{B}_i\boldsymbol{\beta}$, or $\boldsymbol{\zeta}_i=\boldsymbol{\theta}_i^{-1}(\boldsymbol{\eta}_i)$, where $\mathbf{B}_i$ is an $n_i\times p$ model matrix and $\boldsymbol{\beta}$ a $p\times 1$ unknown regression parameter. Although canonical links are convenient, they are not necessary. In general, we assume $g(\boldsymbol{\zeta}_i)=\boldsymbol{\eta}_i$ for some link function g. In addition, we assume that the mixing probabilities $\mathbf{p}_i=(p_{i1},\ldots,p_{in_i})^{\mathrm{T}}$ are related to covariates through $g_p(\mathbf{p}_i)=\mathbf{G}_i\boldsymbol{\gamma}$, where g_p is a link function (e.g., logit), $\mathbf{G}_i$ is an $n_i\times q$ model matrix, and $\boldsymbol{\gamma}$ an unknown parameter vector.

Let $u_{ij} = 0$ if $Y_{ij} \sim F_2$, $u_{ij} = 1$ otherwise, for $i = 1, \ldots, K$, and $j = 1, \ldots, n_i$. Then under independence, the complete data loglikelihood based on $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{u} = (u_{11}, \ldots, u_{Kn_K})^{\mathrm{T}}$ is given by

$$\begin{aligned}\ell^{\mathrm{c}}(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i,j} [u_{ij} \log p_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log \{1 - p_{ij}(\boldsymbol{\gamma})\}] + \sum_{i,j} (1 - u_{ij}) \log f_2(y_{ij}; \boldsymbol{\beta}, \phi) \\ &= \ell^{\mathrm{c}}(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + \ell^{\mathrm{c}}(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u})\end{aligned}$$

At the $(h+1)$th step of the EM algorithm we maximize $Q(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma} | \boldsymbol{\beta}^{(h)}, \phi^{(h)}, \boldsymbol{\gamma}^{(h)}) = \ell^{\mathrm{c}}(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}^{(h)}) + \ell^{\mathrm{c}}(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}^{(h)})$, where $\mathbf{u}^{(h)}$ has components $u_{ij}^{(h)} = 1_{\{y_{ij}=0\}}[1 + \{1 - p_{ij}(\boldsymbol{\gamma}^{(h)})\} f_2(y_{ij}; \boldsymbol{\beta}^{(h)}, \phi^{(h)}) / p_{ij}(\boldsymbol{\gamma}^{(h)})]^{-1}$, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. Maximization with respect to $\boldsymbol{\gamma}$ leads to solving the equation

$$\sum_{i=1}^{K} \left\{ \frac{\partial \mathbf{p}_i(\boldsymbol{\gamma})^{\mathrm{T}}}{\partial \boldsymbol{\gamma}} \right\} [\mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\} \mathbf{I} \mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\}]^{-1} \{\mathbf{u}_i^{(h)} - \mathbf{p}_i(\boldsymbol{\gamma})\} = \mathbf{0}, \tag{3.1}$$

where $\mathbf{A}_i(\mathbf{p}_i) = \mathrm{diag}\{p_{i1}(1 - p_{i1}), \ldots, p_{in_i}(1 - p_{in_i})\}$. Maximization of Q with respect to $\boldsymbol{\beta}$ leads to solving the equation

$$\sum_{i=1}^{K} \left\{ \frac{\partial \boldsymbol{\zeta}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \right\} [\mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\} \mathbf{I} \mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}]^{-1} \mathbf{U}_i^{(h)} \{\mathbf{y}_i - \boldsymbol{\zeta}_i(\boldsymbol{\beta})\} = \mathbf{0}, \tag{3.2}$$

where $\mathbf{U}_i^{(h)} = \mathrm{diag}\{(1 - u_{i1}^{(h)}), \ldots, (1 - u_{in_i}^{(h)})\}$, and $\mathbf{D}_i(\boldsymbol{\zeta}_i) = \mathrm{diag}\{\phi v(\zeta_{i1}) / w_{i1}, \ldots, \phi v(\zeta_{in_i}) / w_{in_i}\}$.

Notice that both Equation (3.1) and (3.2) have the form of a GEE [a weighted GEE in the case of Equation (3.2)] with working correlation matrix equal to $\mathbf{I}$. As in Rosen *et al.* (2000), we propose to modify the independence EM algorithm by substituting working correlation matrices of the exchangeable, AR(1) or other form in place of $\mathbf{I}$ in Equations (3.1) and (3.2) to account for within cluster correlation. This leads to equations

$$\sum_{i=1}^{K} \left\{ \frac{\partial \mathbf{p}_i(\boldsymbol{\gamma})^{\mathrm{T}}}{\partial \boldsymbol{\gamma}} \right\} [\mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\} \mathbf{R}(\boldsymbol{\delta}) \mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\}]^{-1} \{\mathbf{u}_i^{(h)} - \mathbf{p}_i(\boldsymbol{\gamma})\} = \mathbf{0} \tag{3.3}$$

and

$$\sum_{i=1}^{K} \left\{ \frac{\partial \boldsymbol{\zeta}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \right\} [\mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\} \mathbf{P}(\boldsymbol{\rho}) \mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}]^{-1} \mathbf{U}_i^{(h)} \{\mathbf{y}_i - \boldsymbol{\zeta}_i(\boldsymbol{\beta})\} = \mathbf{0}, \tag{3.4}$$

where $\mathbf{R}(\boldsymbol{\delta})$ and $\mathbf{P}(\boldsymbol{\rho})$ are working correlation matrices. Here $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$ are correlation parameters that must be estimated. In addition, there is a scale parameter ϕ to be estimated.

As in the original GEE approach (Liang and Zeger, 1986), method of moment estimators for these parameters can be used. However, greater efficiency in the estimation of $\boldsymbol{\beta}$ and γ can be achieved by introducing a second set of estimating equations for $\boldsymbol{\delta}$, and for $\boldsymbol{\rho}$ and ϕ (Prentice and Zhao, 1991). To guard against correlation misspecification, we advocate the GEE-1 approach (Liang *et al.*, 1992) in which first and second moment parameters are treated orthogonally. This leads to replacing Equation (3.4) by the following combined estimating equation for $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and ϕ

$$\sum_{i=1}^{K}\begin{pmatrix} \partial\boldsymbol{\zeta}_i^{\mathrm{T}}/\partial\boldsymbol{\beta} & \mathbf{0} \\ \mathbf{0} & \partial\boldsymbol{\sigma}_i^{\mathrm{T}}/\partial\tilde{\boldsymbol{\rho}} \end{pmatrix}\begin{pmatrix} \mathbf{V}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{i22}^{-1} \end{pmatrix}\mathbf{H}_i\begin{pmatrix} \mathbf{y}_i - \boldsymbol{\zeta}_i \\ \mathbf{s}_i - \boldsymbol{\sigma}_i \end{pmatrix} = \mathbf{0}. \tag{3.5}$$

Here, $\tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}^{\mathrm{T}}, \phi)^{\mathrm{T}}$, $\mathbf{V}_{i11} = \mathbf{D}_i^{1/2}\{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}\mathbf{P}(\boldsymbol{\rho})\mathbf{D}_i^{1/2}\{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\}$, $\mathbf{s}_i = \mathrm{vech}\{(\mathbf{y}_i - \boldsymbol{\zeta}_i)(\mathbf{y}_i - \boldsymbol{\zeta}_i)^{\mathrm{T}}\}$, $\boldsymbol{\sigma}_i = E(\mathbf{s}_i) = \mathrm{vech}(\mathbf{V}_{i11})$, $\mathbf{H}_i = \mathrm{diag}[\mathbf{j}_{n_i} - \mathbf{u}_i^{(h)}, \mathrm{vech}\{(\mathbf{j}_{n_i} - \mathbf{u}_i^{(h)})(\mathbf{j}_{n_i} - \mathbf{u}_i^{(h)})^{\mathrm{T}}\}]$, where $\mathbf{j}_{n_i}$ is an $n_i \times 1$ vector of ones. In addition, $\mathbf{V}_{i22}^{-1}$ is a weight matrix for the second moment estimating function. For this matrix, we suggest the 'Gaussian working structure' of Prentice and Zhao (1991). This structure is convenient, and has been found to work well in the GEE-1 context by several authors (Hall, 2001; Hall and Severini, 1998; Lipsitz *et al.*, 2000). The Gaussian working structure sets third and fourth order moments to what they would be if the response vector were multivariate normal (Gaussian). In particular, $\mathbf{V}_{i22}$ has elements given by the relation $\mathrm{cov}(s_{ijk}, s_{i\ell m}) = \sigma_{ij\ell}\sigma_{ikm} + \sigma_{ijm}\sigma_{ik\ell}$ where $s_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})$ and $\sigma_{ijk} = E(s_{ijk})$ are the elements of $\mathbf{s}_i$ and $\boldsymbol{\sigma}_i$, respectively.

Similarly, a second estimating equation can be added to Equation (3.3) for the estimation of γ and $\boldsymbol{\delta}$. In particular, we replace Equation (3.3) by the combined estimating equation

$$\sum_{i=1}^{K}\begin{pmatrix} \partial\mathbf{p}_i^{\mathrm{T}}/\partial\gamma & \mathbf{0} \\ \mathbf{0} & \partial\boldsymbol{\tau}_i^{\mathrm{T}}/\partial\boldsymbol{\delta} \end{pmatrix}\begin{pmatrix} \mathbf{W}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i22}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{u}_i^{(h)} - \mathbf{p}_i \\ \mathbf{t}_i - \boldsymbol{\tau}_i \end{pmatrix} = \mathbf{0}, \tag{3.6}$$

where $\mathbf{W}_{i11} = \mathbf{A}_i^{1/2}\{\mathbf{p}_i(\gamma)\}\mathbf{R}(\boldsymbol{\delta})\mathbf{A}_i^{1/2}\{\mathbf{p}_i(\gamma)\}$, $\mathbf{t}_i = \mathrm{vech}\{(\mathbf{u}_i^{(h)} - \mathbf{p}_i)(\mathbf{u}_i^{(h)} - \mathbf{p}_i)^{\mathrm{T}}\}$, $\boldsymbol{\tau}_i = E(\mathbf{t}_i) = \mathrm{vech}(\mathbf{W}_{i11})$, and $\mathbf{W}_{i22}^{-1}$ is a weight matrix. Again, we suggest the Gaussian working form for $\mathbf{W}_{i22}$, so that that elements of $\mathbf{W}_{i22}$ are functions of the elements of $\boldsymbol{\tau}_i$.

Estimation proceeds by iteratively computing $\mathbf{u}^{(h)}$ in the E step and solving Equations (3.5) and (3.6) in the S step. By the results of Rosen *et al.* (2000), the estimators at convergence are consistent and asymptotically normal with variance–covariance matrix that can be estimated from the estimating functions used in the S step. In particular, let $\psi = (\gamma^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\rho}^{\mathrm{T}}, \phi)^{\mathrm{T}}$ be the combined parameter vector with corresponding ES estimator $\hat{\psi}$. Then $\mathrm{var}(\hat{\psi})$ is consistently estimated by $\hat{\boldsymbol{\xi}}^{-1}\hat{\boldsymbol{\zeta}}\hat{\boldsymbol{\xi}}^{-T}$, where

$$\hat{\boldsymbol{\xi}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\nabla s_{ij}(\hat{\psi}), \quad \hat{\boldsymbol{\zeta}} = \sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i} s_{ij}(\hat{\psi})\right\}\left\{\sum_{j=1}^{n_i} s_{ij}(\hat{\psi})\right\}^{\mathrm{T}} \tag{3.7}$$

and $s_{ij}(\hat{\psi})$ is a $\dim(\psi) \times 1$ vector which contains the estimating functions of $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and ϕ used in the S-step. Let $[\nabla \mathbf{x}]_{kl} = \partial x_k / \partial \psi_l$ for $k, l = 1, \ldots, \dim(\psi)$. In $s_{ij}(\hat{\psi})$, the first $\dim(\boldsymbol{\gamma})$ components are

$$\left(\sum_{m=1}^{n_i} \left[\frac{\partial \mathbf{p}_i(\boldsymbol{\gamma})^{\mathrm{T}}}{\partial \boldsymbol{\gamma}} \right]_{ml} \left[\mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\} \mathbf{R}(\boldsymbol{\delta}) \mathbf{A}_i^{1/2} \{\mathbf{p}_i(\boldsymbol{\gamma})\} \right]^{-1}_{mj} \{ u_{ij}^{(\infty)}(\psi) - p_{ij}(\boldsymbol{\gamma}) \} \right)_{\psi = \hat{\psi}},$$

for $l = 1, \ldots, \dim(\boldsymbol{\gamma})$ and $j = 1, \ldots, n_i$. The next $\dim(\boldsymbol{\delta})$ components in $s_{ij}(\hat{\psi})$ are

$$\left(\sum_{m=1}^{n_i} \left[\frac{\partial \boldsymbol{\tau}_i^{\mathrm{T}}}{\partial \boldsymbol{\delta}} \right]_{ml} [\mathbf{W}_{i22}^{-1}]_{mj} (t_{ij} - \tau_{ij}) \right)_{\psi = \hat{\psi}},$$

for $l = 1, \ldots, \dim(\boldsymbol{\delta})$. For parameter $\boldsymbol{\beta}$, the $\dim(\boldsymbol{\beta})$ components in $s_{ij}(\hat{\psi})$ are

$$\left(\sum_{m=1}^{n_i} \left[\frac{\partial \boldsymbol{\zeta}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \right]_{ml} \left[\mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\} \mathbf{P}(\boldsymbol{\rho}) \mathbf{D}_i^{1/2} \{\boldsymbol{\zeta}_i(\boldsymbol{\beta})\} \right]^{-1}_{mj} \{1 - u_{ij}^{(\infty)}(\psi)\} \{y_{ij} - \zeta_{ij}(\boldsymbol{\beta})\} \right)_{\psi = \hat{\psi}},$$

for $l = 1, \ldots \dim(\boldsymbol{\beta})$ and $j = 1, \ldots, n_i$. Finally, $s_{ij}(\hat{\psi})$ also contains $\dim(\tilde{\boldsymbol{\rho}})$ components for $\tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}, \phi)$, which are

$$\left(\sum_{m=1}^{n_i} \left[\frac{\partial \boldsymbol{\sigma}_i^{\mathrm{T}}}{\partial \tilde{\boldsymbol{\rho}}} \right]_{ml} \left[\mathbf{V}_{i22}^{-1} \right]_{mj} [\mathbf{h}_i]_c (s_{ij} - \sigma_{ij}) \right)_{\psi = \hat{\psi}},$$

for $l = 1, \ldots, \dim(\tilde{\boldsymbol{\rho}})$ and $c = 1, \ldots, n_i(n_i + 1)/2$, where $\mathbf{h}_i = \mathrm{diag}[\mathrm{vech}\{(\mathbf{j}_{n_i} - \mathbf{u}_i^{(\infty)})(\mathbf{j}_{n_i} - \mathbf{u}_i^{(\infty)})^{\mathrm{T}}\}]$. On the basis of Equations in (3.7), $\mathrm{var}(\hat{\psi})$ can be easily calculated using numerical derivatives.

3.1 ZIP and ZIB models with overdispersion for independent data

Because the Poisson, binomial and normal distributions are all in the exponential dispersion family, it is clear that ZIP, ZIB and ZI (possibly transformed) normal regression models for clustered data can all be accommodated in the framework described above. However, this methodology also applies to ZI overdispersed Poisson (ZIOP) and ZI overdispersed binomial regression models for independent data. Such models can be thought of as simple alternatives to ZINB and ZI beta binomial regression models. For example, in the independent (or nonclustered) data case, $n_i = 1$ for all i, and the ZIP model assumes $Y_i \sim 0$ with probability $p_i(\boldsymbol{\gamma})$ and $Y_i \sim \mathrm{Poisson}\{\lambda_i(\boldsymbol{\beta})\}$ otherwise. This assumption can be relaxed so that $Y_i \sim 0$ with probability $p_i(\boldsymbol{\gamma})$ and $Y_i \sim F_2$ otherwise, where F_2 has mean $\lambda_i(\boldsymbol{\beta})$ and variance $\phi \lambda_i(\boldsymbol{\beta})$ where ϕ is an unknown parameter to be estimated. This can be easily accommodated in the ES algorithm by solving Equations (3.1) and (3.5) in the S step, where $\mathbf{P} = \mathbf{I}$ and $\tilde{\boldsymbol{\rho}} = \phi$. Or, more simply, we can substitute a MOM estimator $\hat{\phi}$ into Equation (3.2) and

solve Equations (3.1) and (3.2) in the S step. The approach here is akin to quasi-likelihood estimation in the overdispersed Poisson GLM.

3.2 ZI censored lognormal regression

Because the density function (2.2) does not belong to the exponential dispersion family, the methodology described above does not directly apply to the ZI censored lognormal model. However, by introducing missing data corresponding to the uncensored lognormal response as well as for the zero inflation mechanism, we can construct a complete data loglikelihood that decomposes into a term for $\boldsymbol{\gamma}$ and one for $(\boldsymbol{\beta}, \sigma)$. Given the missing data, the latter term is lognormal, rather than censored lognormal, and hence is in the exponential dispersion family. Thus, by augmenting the data correctly, we can handle the ZI censored lognormal regression problem with the ES methodology described above.

In particular, let $\mathbf{u}$ and $\mathbf{e}$ be the missing data, where $u_{ij} = 1$ if y_{ij} is an extra zero, and $u_{ij} = 0$ if y_{ij} comes from a censored lognormal distribution, and $\mathbf{e} = (e_{11}, \ldots, e_{Kn_K})^{\mathrm{T}}$ is the lognormal response vector 'prior to' censoring and zero inflation. Note that $\mathbf{e}$ is partially observed, since $y_{ij} = e_{ij}$ when $y_{ij} > \alpha$. We assume the response y_{ij} corresponds to the random variable $Y_{ij} = (1 - U_{ij})Z_{ij}$ where $U_{ij} = 1$ with probability p_{ij} and $U_{ij} = 0$ otherwise. Here, Z_{ij} is censored lognormal, following density function (2.2), and is related to the partially latent variable E_{ij} via $Z_{ij} = \mathbf{1}_{\{E_{ij} > \alpha\}} E_{ij}$. We also assume U_{ij} and E_{ij} are independent. Under these assumptions, the joint density of (y_{ij}, u_{ij}, e_{ij}) is

$$f(y_{ij}, u_{ij}, e_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) = p_{ij}^{u_{ij}} (1 - p_{ij})^{(1-u_{ij})} \frac{1}{e_{ij}\sqrt{2\pi\sigma^2}} \exp\{ -\frac{1}{2\sigma^2} (\log e_{ij} - \mu_{ij})^2 \}.$$

Under independence both between and within-clusters, the complete data loglikelihood for the entire sample is

$$\begin{aligned} \ell^{\mathrm{c}}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma; \mathbf{y}, \mathbf{u}, \mathbf{w}) &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} [u_{ij} \log p_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log \{1 - p_{ij}(\boldsymbol{\gamma})\}] \\ &\quad - \left[\frac{1}{2\sigma^2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \{\log w_{ij} - \mu_{ij}(\boldsymbol{\beta})\}^2 + \sum_{i=1}^{K} \sum_{j=1}^{n_i} \log w_{ij} + \frac{N}{2} \log (2\pi\sigma^2) \right] \\ &\equiv \ell^{\mathrm{c}}(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}, \mathbf{w}) + \ell^{\mathrm{c}}(\boldsymbol{\beta}, \sigma; \mathbf{y}, \mathbf{u}, \mathbf{w}), \end{aligned}$$

where $N = \sum_{i=1}^{K} n_i$.

Again, let $\psi = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$ be the combined vector of parameters. At the $(h+1)$th iteration of the EM algorithm, we compute $Q(\psi|\psi^{(h)}) = E_{\mathbf{e},\mathbf{u}|\mathbf{y},\psi^{(h)}}\{\ell^{\mathrm{c}}(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}, \mathbf{e})|\mathbf{y},\psi^{(h)}\} + E_{\mathbf{e},\mathbf{u}|\mathbf{y},\psi^{(h)}}\{\ell^{\mathrm{c}}(\boldsymbol{\beta}, \sigma; \mathbf{y}, \mathbf{u}, \mathbf{e})|\mathbf{y},\psi^{(h)}\}$, where $\psi^{(h)}$ is the parameter vector from the previous

step. The expectation is with respect to the joint distribution of $\mathbf{u}$ and $\mathbf{e}$ given $\mathbf{y}$ and $\psi^{(h)}$. For the first term,

$$E_{\mathbf{e},\mathbf{u}|\mathbf{y},\psi^{(h)}}\{\ell^{c}(\gamma;\,\mathbf{y},\,\mathbf{u},\,\mathbf{e})\} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}[u_{ij}^{(h)}\log p_{ij}(\gamma) + (1-u_{ij}^{(h)})\log\{1-p_{ij}(\gamma)\}],$$

where $u_{ij}^{(h)} = 1_{\{y_{ij}=0\}}[1+\Phi((\log\alpha - \mathbf{B}_i\boldsymbol{\beta}^{(h)})/\sigma^{(h)})/\exp(\mathbf{G}_i\gamma^{(h)})]^{-1}$. For the second term in $Q(\psi|\psi^{(h)})$,

$$\begin{aligned}E_{\mathbf{w},\mathbf{u}|\mathbf{y},\psi^{(h)}}\{\ell^{c}(\boldsymbol{\beta},\,\sigma;\,\mathbf{y},\,\mathbf{u},\,\mathbf{w})\} &= E_{\mathbf{u}|\mathbf{y},\psi^{(h)}}[E_{\mathbf{w}|\mathbf{u},\psi^{(h)}}\{\ell^{c}(\boldsymbol{\beta},\,\sigma;\,\mathbf{y},\,\mathbf{u},\,\mathbf{w})|\mathbf{y},\,\mathbf{u},\,\psi^{(h)}\}]\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}E_{\mathbf{u}|\mathbf{y},\psi^{(h)}}[E_{\mathbf{w}|\mathbf{u},\psi^{(h)}}\{\log w_{ij} - \mu_{ij}(\boldsymbol{\beta})\}^2]\\ &\quad -\sum_{i=1}^{K}\sum_{j=1}^{n_i}E_{\mathbf{u}|\mathbf{y},\psi^{(h)}}\{E_{\mathbf{w}|\mathbf{u},\psi^{(h)}}(\log w_{ij})\} - \frac{N}{2}\log(2\pi\sigma^2).\end{aligned}$$

In the earlier equation, only the first term involves $\boldsymbol{\beta}$. Thus, maximization of Q with respect to $\boldsymbol{\beta}$ leads to solving the equation

$$\sum_{i=1}^{K}\left\{\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial\boldsymbol{\beta}}\right\}\mathbf{I}[E_{\mathbf{u}|\mathbf{y},\psi^{(h)}}\{E_{\mathbf{e}|\mathbf{u},\psi^{(h)}}(\log\mathbf{e}_i)\} - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}. \tag{3.8}$$

Once again, this equation has the form of a GEE with independence working correlation matrix [compare Equation (3.2)]. The expection term in Equation (3.8), which plays the role of $\mathbf{y}_i$ in Equation (3.2), is easily calculated:

$$\begin{aligned}E_{\mathbf{u}|\mathbf{y},\psi^{(h)}}\{E_{\mathbf{w}|\mathbf{u},\psi^{(h)}}(\log\mathbf{w}_{ij})\} &= 1_{\{y_{ij}>\alpha\}}\log(y_{ij}) + (1-u_{ij}^{(h)})1_{\{y_{ij}=0\}}E_{\mathbf{w}|\psi^{(h)}}\{\log(\mathbf{w}_{ij})|w_{ij}\le\alpha\}\\ &\quad + u_{ij}^{(h)}1_{\{y_{ij}=0\}}E_{\mathbf{w}|\psi^{(h)}}(\log\mathbf{w}_{ij})\end{aligned}$$

In addition, we have assumed that e_{ij} is from lognormal(μ_{ij},σ^2) distribution, so $E_{\mathbf{e}|\psi^{(h)}}(\log e_{ij}) = \mu_{ij}^{(h)} = \mathbf{B}_i\boldsymbol{\beta}^{(h)}$, and

$$E_{\mathbf{e}|\psi^{(h)}}(\log e_{ij}|e_{ij}\le\alpha) = -\frac{\sigma^{(h)}}{\sqrt{2\pi}}\exp\{-\frac{1}{2(\sigma^{(h)})^2}(\log\alpha - \mathbf{B}_i\boldsymbol{\beta}^{(h)})^2\} + \mathbf{B}_i\boldsymbol{\beta}^{(h)}\Phi\left(\frac{\log\alpha - \mathbf{B}_i\boldsymbol{\beta}^{(h)}}{\sigma^{(h)}}\right).$$

Now, since maximization of Q with respect to γ and $\boldsymbol{\beta}$ leads to solving GEE-like equations (3.1) and (3.8), repectively, the method described in Section 3 can be applied to the ZI censored lognormal model. That is, to account for within-cluster correlation, the identity working correlation matrix $\mathbf{I}$ in these equations can be replaced by an exchangeable, AR(1) or other form matrix to account for within-cluster correlation.

4 Direct application of GEEs to the observed response

An alternative approach for estimation in the marginal model for clustered ZI data that is natural to consider is to apply the GEE methodology to the observed response vector $\mathbf{y}$. That is, here we consider GEEs formed from assumptions on the first and second moments of $\mathbf{y}_i$, $i = 1, \ldots, K$, where the variance–covariance assumption accounts for correlation among the elements of $\mathbf{y}_i$. These assumptions could be made directly on $\mathbf{y}_i$ or indirectly, through assumptions on the unobserved ('pre-zero inflation') response $\mathbf{z}_i$ and on the relationship between $\mathbf{y}_i$ and $\mathbf{z}_i$. We adopt the indirect approach, because it seems easier and more natural to hypothesize a variance–covariance structure, say, on the unmixed response (e.g., the correlated Poisson vector $\mathbf{z}_i$) than on the mixture response (e.g., the correlated ZIP vector $\mathbf{y}_i$).

Suppose $\mathbf{z}_i$ is a vector of correlated responses specific to the ith cluster, and suppose $\zeta_{ij} \equiv E(z_{ij}) = g^{-1}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$, $\mathrm{var}(z_{ij}) = \phi v(\zeta_{ij})/w_{ij}$, where $g^{-1}(\cdot)$ is the inverse link function, $\boldsymbol{\beta}$ is an unknown regression parameter corresponding to the vector of explanatory variables $\mathbf{x}_{ij}$, $v(\cdot)$ is a variance function, ϕ is a scale parameter (possibly unknown) and w_{ij} are known weights (e.g., binomial denominators). In addition, we assume a working correlation structure for $\mathbf{z}_i$: $\mathrm{var}(\mathbf{z}_i) = \phi \mathbf{D}_i(\boldsymbol{\zeta}_i)^{1/2}\mathbf{P}(\boldsymbol{\rho})\mathbf{D}_i(\boldsymbol{\zeta}_i)^{1/2}$, where $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{in_i})^{\mathrm{T}}$, $\mathbf{D}_i = \mathrm{diag}\{v(\zeta_{i1})/w_{i1}, \ldots, v(\zeta_{in_i})/w_{in_i}\}$ and $\mathbf{P}(\boldsymbol{\rho})$ is a working (i.e., not necessarily correct) correlation matrix for $\mathbf{z}_i$ parameterized by $\boldsymbol{\rho}$. We assume independence between clusters.

However, we do not observe $\mathbf{z}_i$. Instead, we observe $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, a ZI version of $\mathbf{z}_i$. We assume $y_{ij} = 0$, with probability p_{ij}, $y_{ij} = z_{ij}$, with probability $1 - p_{ij}$, where we assume the mixing mechanism is independent both between and within-clusters. More explicitly, let $u_{ij} = 1$ with probability p_{ij} and $u_{ij} = 0$ with probability $1 - p_{ij}$, where $u_{11}, \ldots, u_{Kn_K}$ are independent of each other and of the $\mathbf{z}_i$s. Then the observed response is $y_{ij} = (1 - u_{ij})z_{ij}$, with $\mu_{ij} \equiv E(y_{ij}) = (1 - p_{ij})\zeta_{ij}$, $\mathrm{var}(y_{ij}) = (1 - p_{ij})\{\phi v(\zeta_{ij})/w_{ij} + p_{ij}\zeta_{ij}^2\}$. As in the usual ZIP model, we allow that the mixing probabilities $\{p_{ij}\}$ may depend on covariates through a logit link: $\mathrm{logit}(\mathbf{p}_i) = \mathbf{G}_i\boldsymbol{\gamma}$, $i = 1, \ldots, K$. On the basis of the working model for $\mathrm{corr}(\mathbf{z}_i)$, the marginal variance–covariance matrix of $\mathbf{y}_i$ is given by $\mathbf{V}_i(\boldsymbol{\chi}, \boldsymbol{\rho}, \phi) \equiv \mathrm{var}(\mathbf{y}_i) = \phi \mathbf{Q}_i\mathbf{D}_i(\boldsymbol{\zeta}_i)^{1/2}\mathbf{P}(\boldsymbol{\rho})\mathbf{D}_i(\boldsymbol{\zeta}_i)^{1/2}\mathbf{Q}_i + \mathbf{C}_i$, where $\mathbf{C}_i = \mathrm{diag}\{p_{i1}(1 - p_{i1})(\phi v(\zeta_{i1})/w_{i1} + \zeta_{i1}^2), \ldots, p_{in_i}(1 - p_{in_i})(\phi v(\zeta_{in_i})/w_{in_i} + \zeta_{in_i}^2)\}$, $\mathbf{Q}_i = \mathrm{diag}(1 - p_{i1}, \ldots, 1 - p_{in_i})$ and we have made the dependence of $\mathrm{var}(\mathbf{y}_i)$ on the regression parameter $\boldsymbol{\chi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$, correlation parameter $\boldsymbol{\rho}$, and dispersion parameter ϕ explicit in the notation $\mathbf{V}_i(\boldsymbol{\chi}, \boldsymbol{\rho}, \phi)$.

Under this marginal model, the GEEs for the first moment parameter $\boldsymbol{\chi}$ take the form

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^{\mathrm{T}}}{\partial \boldsymbol{\chi}} \mathbf{V}_i^{-1}(\boldsymbol{\chi}, \hat{\boldsymbol{\rho}}, \hat{\phi})(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{4.1}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^{\mathrm{T}}$. Here $\mathbf{V}_i$ is evaluated at $\hat{\boldsymbol{\rho}}$ and $\hat{\phi}$, which are $\sqrt{K}$ consistent estimators of the corresponding parameters. However, there is a serious problem with this approach because $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ will typically be confounded in Equation (4.1), and hence not identifiable. Essentially the same problem was pointed out by Crowder (1987), who used several examples, including a mixture model, to illustrate that linear estimating

functions such as Equation (4.1) can fail. To alleviate the problem, he suggested the use of a quadratic estimating function, which essentially introduces an additional moment condition to solve the problem of p_{ij} and ζ_{ij} being confounded in μ_{ij}. The idea here is basically the same as the requirement of having as many moment equations as parameters for the method of moments.

Let $\mathbf{s}_i^{(1)} = \text{diag}\{(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\}$ be the vector of 'empirical' variances of the y_{ij}s, with expectation $\boldsymbol{\sigma}_i^{(1)} = E(\mathbf{s}_i^{(1)}) = \text{diag}(\mathbf{V}_i)$, and let $\mathbf{s}_i^{(2)} = \text{vech}^*\{(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\}$ be the empirical covariances, with corresponding expectation $\boldsymbol{\sigma}_i^{(2)}$. Here $\text{vech}^*(\cdot)$ denotes the vector valued function that returns the columns of its matrix argument, including only those elements below the diagonal. Note that this differs from the usual $\text{vech}(\cdot)$, which returns only those elements on or below the diagonal (Searle, 1982: 332). Let $\mathbf{s}_i = \{(\mathbf{s}_i^{(1)})^{\mathrm{T}}, (\mathbf{s}_i^{(2)})^{\mathrm{T}}\}^{\mathrm{T}}$ and $\boldsymbol{\sigma}_i = \{(\boldsymbol{\sigma}_i^{(1)})^{\mathrm{T}}, (\boldsymbol{\sigma}_i^{(2)})^{\mathrm{T}}\}^{\mathrm{T}}$. A quadratic estimating function for $\boldsymbol{\chi}$ could be built from the elementary estimating function $\mathbf{s}_i - \boldsymbol{\sigma}_i$. However, we wish to avoid the use of information in the covariances (the elements of $\boldsymbol{\sigma}_i^{(2)}$) for estimation of the mean parameter $\boldsymbol{\chi}$, so that our estimators of this parameter are robust to correlation structure misspecification. Therefore, instead we propose an estimating function of the form

$$\sum_{i=1}^{K} \left\{ \frac{\partial \boldsymbol{\mu}_i^{\mathrm{T}}}{\partial \boldsymbol{\chi}}, \frac{\partial (\boldsymbol{\sigma}_i^{(1)})^{\mathrm{T}}}{\partial \boldsymbol{\chi}} \right\} \mathbf{W}_i \begin{pmatrix} \mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{s}_i^{(1)} - \boldsymbol{\sigma}_i^{(1)} \end{pmatrix}, \tag{4.2}$$

where $\mathbf{W}_i$ is a weight matrix. Optimally, $\mathbf{W}_i$ would be set equal to $\left\{ \text{var} \begin{pmatrix} \mathbf{y}_i \\ \mathbf{s}_i^{(1)} \end{pmatrix} \right\}^{-1}$. However, this choice requires us to specify third and fourth order moments of $\mathbf{y}_i$ (or $\mathbf{z}_i$), which we would like to avoid. A natural alternative is to use a working structure for $\text{cov}(\mathbf{y}_i, \mathbf{s}_i^{(1)})$ and $\text{var}(\mathbf{s}_i^{(1)})$. That is, we set

$$\mathbf{W}_i = \begin{pmatrix} \mathbf{V}_i & \tilde{\text{cov}}(\mathbf{y}_i, \mathbf{s}_i^{(1)}) \\ \tilde{\text{cov}}(\mathbf{s}_i^{(1)}, \mathbf{y}_i) & \tilde{\text{var}}(\mathbf{s}_i^{(1)}) \end{pmatrix}^{-1},$$

where '~' indicates that a working structure is used. The estimating equation (4.2) for $\boldsymbol{\chi}$ could be augmented by method of moment estimators for $\boldsymbol{\rho}$ and ϕ, or by estimating equations for these parameters. As in Section 3, we adopt the latter approach, with estimating functions of the form

$$\sum_{i=1}^{K} \frac{\partial (\sigma_i^{(2)})^{\mathrm{T}}}{\partial \boldsymbol{\rho}} \{\tilde{\text{var}}(\mathbf{s}_i^{(2)})\}^{-1} (\mathbf{s}_i^{(2)} - \boldsymbol{\sigma}_i^{(2)}), \quad \sum_{i=1}^{K} \frac{\partial \boldsymbol{\sigma}_i^{\mathrm{T}}}{\partial \phi} \{\tilde{\text{var}}(\mathbf{s}_i)\}^{-1} (\mathbf{s}_i - \boldsymbol{\sigma}_i),$$

for $\boldsymbol{\rho}$ and ϕ, respectively.

For the working third and fourth order moment structures, we again adopt the Gaussian working structure. In this context, we can either use the structure implied by the assumption that $\mathbf{y}_i$ is Gaussian, or by the assumption that $\mathbf{z}_i$ is Gaussian. The former choice simplifies the formulas for $\tilde{\text{cov}}(\mathbf{y}_i, \mathbf{s}_i^{(1)})$ and $\tilde{\text{var}}(\mathbf{s}_i^{(1)})$. The latter option, however,

is consistent with our approach of making direct moment assumptions on $\mathbf{z}_i$ rather than $\mathbf{y}_i$. In addition, since $\mathbf{y}_i$ is the ZI version of $\mathbf{z}_i$, we would typically expect $\mathbf{z}_i$ to be more nearly normal than $\mathbf{y}_i$. On the basis of a working Gaussian assumption on $\mathbf{z}_i$ and the relationship $y_{ij} = (1 - u_{ij})z_{ij}$ it is possible to derive formulas for $\text{cov}(\mathbf{y}_i, \mathbf{s}_i)$ and $\text{var}(\mathbf{s}_i)$. Because these formulas are somewhat complex and for the sake of brevity, they are omitted here, but may be obtained by request from the first author.

5 A simulation study

To investigate the relative performance of the ES algorithm with GEEs and direct GEE approaches to fitting marginal GLMs with excess zeros for clustered data, we conducted a small simulation study on the basis of the now famous epilepsy data set of Thall and Vail (1990). These data consist of seizure counts in each of four consecutive two week periods in addition to an eight week baseline seizure count, a treatment indicator (active drug progabide versus placebo), the subject's age, and other covariates, measured on each of $K = 59$ epilepsy patients.

We randomly generated 500 data sets of the structure of the epilepsy data, where the 4×1 response vector $\mathbf{y}_i$ for the ith subject was generated as a correlated Poisson vector subject to zero inflation. That is, for each subject, we generated $\mathbf{z}_i \sim \text{Poisson}(\boldsymbol{\mu}_i)$ where $\mu_{ij} = \exp\{\beta_0 + \beta_1 \log(\text{base}_i) + \beta_2 \text{trt}_i + \beta_3 \log(\text{base}_i)\text{trt}_i\}$, $i = 1, \ldots, 59$, $j = 1, \ldots, 4$, where base_i is the eight week baseline seizure count divided by four for the ith subject, and $\text{trt}_i = 1$ if treated with progabide, $\text{trt}_i = 0$ if treated with placebo. Then zero inflation was added, giving the response vector $\mathbf{y}_i$ with $y_{ij} = (1 - u_{ij})z_{ij}$, where the u_{ij}s are independent Bernoulli random variables with parameters $p_{ij} = \text{logit}^{-1}(\gamma_0 + \gamma_1 \text{trt}_i)$, $i = 1, \ldots, 59$, $j = 1, \ldots, 4$.

The true parameter values used to generate the data were $\boldsymbol{\beta} = (2.12, 0.43, -0.49, 0.21)^{\text{T}}$ and $\boldsymbol{\gamma} = (-2.20, 0.46)^{\text{T}}$, which imply zero inflation probabilities of 0.10 and 0.15 for the placebo and progabide groups, respectively. Three correlation structures, the exchangeable, AR(1) and Toeplitz (i.e., banded, where $\mathbf{R}_{ii} = 1$, $\mathbf{R}_{ij} = \rho_{|i-j|}$, $i \neq j$), were used to generate and model the data. In addition, we examined three strengths of correlation among the elements of $\mathbf{z}_i$. The correlated Poisson vectors $\{\mathbf{z}_i\}$ were generated using the algorithm of Sim (1993). Note that generation of four dimensional correlated Poisson vectors with $\rho > 0.5$ is not possible for this algorithm under the AR(1) and Toeplitz structures, so we only examined two levels of correlation for these working structures. The low, medium and high levels of correlation corresponded to $\rho = 0.25$, $\rho = 0.5$ and $\rho = 0.75$ for the exchangeable structure; low and medium levels of correlation were given by $\rho = 0.25$ and $\rho = 0.5$ for the AR(1) structure and $\rho = (0.25, 0.15, 0.05)^{\text{T}}$ and $\rho = (0.5, 0.4, 0.3)^{\text{T}}$ for Toeplitz. Models were fit using the ES algorithm and the direct GEE approaches with software routines written in Matlab by the authors.

Ratios of mean squared errors from this simulation study are summarized in Table 1. The results in Table 1 are as expected; the ES algorithm approach is substantially more efficient than direct GEE, and efficiency gains are greatest for the correlation parameter $\boldsymbol{\rho}$. In addition, efficiency gains increase as the degree of within cluster correlation

Table 1 Simulation results comparing the ES algorithm with GEEs in the S step (ES-GEE) and with direct GEE for fitting marginal ZIP model to clustered data

Correlation structures	Parameter	Ratio 1			Ratio 2		
		$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
Exchangeable	γ_1	0.781	0.529	0.621	0.822	0.593	0.704
	γ_2	0.826	0.635	0.655	0.862	0.696	0.747
	β_1	0.926	0.860	0.845	0.950	0.882	0.862
	β_2	0.921	0.893	0.858	0.946	0.914	0.874
	β_3	0.849	0.834	0.811	0.893	0.860	0.843
	β_4	0.835	0.849	0.796	0.881	0.877	0.837
	ρ	0.541	0.493	0.477	0.556	0.500	0.480
AR(1)	γ_1	0.731	0.758	–	0.790	0.803	–
	γ_2	0.782	0.812	–	0.870	0.844	–
	β_1	0.837	0.883	–	0.865	0.907	–
	β_2	0.847	0.875	–	0.877	0.902	–
	β_3	0.771	0.815	–	0.819	0.860	–
	β_4	0.760	0.803	–	0.812	0.847	–
	ρ	0.295	0.333	–	0.302	0.343	–
Toeplitz	γ_1	0.758	0.663	–	0.869	0.773	–
	γ_2	0.786	0.732	–	0.887	0.827	–
	β_1	0.893	0.874	–	0.926	0.893	–
	β_2	0.888	0.898	–	0.927	0.920	–
	β_3	0.830	0.883	–	0.888	0.911	–
	β_4	0.807	0.855	–	0.872	0.896	–
	ρ_1	0.327	0.247	–	0.327	0.247	–
	ρ_2	0.310	0.237	–	0.314	0.242	–
	ρ_3	0.460	0.409	–	0.462	0.410	–

Ratio 1 is the MSE of ES-GEE divided by that of direct GEE with the Gaussian working structure assumed on y.
Ratio 2 is the MSE of ES-GEE divided by that of direct GEE with the Gaussian working structure assumed on z.

increases. Also as expected, direct GEE is more efficient when working assumptions are made on **z** rather than **y**.

6 Examples

6.1 Apple tree roots

Ridout *et al.* (1998) used ZINB regression to model data concerning the number of roots produced by shoots of a certain apple tree cultivar. Two hundred and seventy shoots were microprogagated under eight treatments corresponding to a 2×4 completely randomized design. The treatment factors were length of photoperiod (8 and 16h) and concentration of the cytokinin BAP (2.2, 4.4, 8.8 and 17.6 μM) in the growing medium. In two of the treatments 40 shoots were propagated and 30 shoots in each of the other treatments. Roots were then grown under identical conditions, and the number of roots per shoot measured as the response variable. The data appear in Ridout *et al.* (1998).

A large proportion of the shoots grown under the 16 hour photoperiod produced no roots. Therefore, ZI regression models are appropriate to consider here. Because the experimental design is completely randomized, it is reasonable to assume that the

responses are independent across shoots. That is, there is no clustering here. Therefore, we consider ZIP, ZINB, and ZIOP models for these data.

Let y_{ijk} be the number of roots on the kth shoot at the ith level of photoperiod and jth level of BAP. In addition, let λ_{ijk} be the mean from the nondegenerate component of a ZI regression model for y_{ijk}, and let p_{ijk} be the mixing probability. Ridout *et al.* (1998) consider various ZIP and ZINB models for these data, but we focus on models in which $\log(\lambda_{ijk}) = \beta_{1i} + \beta_{2i}\log(\text{BAP}_{ij})$, $\text{logit}(p_{ijk}) = \gamma_i$, for all i, j and k. The ZINB model with this specification for λ_{ijk} and p_{ijk} was among the best fitting models found by Ridout *et al.* (1998).

Table 2 contains results from fitting ZIP, ZINB and ZIOP models to these data. A score test of overdispersion relative to the ZIP model was given by Ridout *et al.* (2001) and Hall and Berenhaut (2002). For the ZIP model fit here, this test yields a value of 3.872 on 1 degree of freedom ($P = 0.0491$), suggesting that the ZINB model or, alternatively, the ZIOP model will provide a better fit to these data. This is borne out by the lower AIC value for the ZINB model. Alternatively, the ZIOP model can be used.

Notice that the parameter estimates and their standard errors agree closely for the ZINB and ZIOP models. ZINB assumes that the variance in the negative binomial component is $\lambda_{ij}(1 + \lambda_{ij}\omega) \equiv \lambda_{ij}\phi_{ij}$, whereas the ZIOP model assumes a variance function of $\lambda_{ij}\phi$ in the nonzero component. The average of the estimated values of the overdispersion index ϕ_{ij} across the eight treatments here is $\hat{\bar{\phi}} = 1.43$, which is quite close to value $\hat{\phi} = 1.41$ from the ZIOP model. Thus, the ZINB and ZIOP models also agree concerning the degree of overdispersion relative to a ZIP model in these data.

A nice feature of the ZIOP model is that the parameter estimates are identical to those from the ZIP model. However, the standard errors for the $\hat{\beta}$s are inflated so that $\text{se}(\hat{\beta}_{\text{ZIOP}}) \approx \hat{\phi}^{1/2}\text{se}(\hat{\beta}_{\text{ZIP}})$ to account for overdispersion. This relationship between the standard errors of $\hat{\beta}_{\text{ZIOP}}$ and $\hat{\beta}_{\text{ZIP}}$ is exact if the 'model based' variance–covariance matrix $\hat{\boldsymbol{\xi}}^{-1}$ is used. This parallels the situation in ordinary (non-ZI) Poisson loglinear models when quasi-likelihood is used to allow a non-unity dispersion parameter in the variance function (McCullagh and Nelder, 1989: 199–200).

Table 2 Parameter estimates (standard errors) and model fit criteria for the ZIP, ZIOP and ZINB models fit to apple tree root data

	Estimates		
Parameter	ZIP	ZIOP	ZINB
β_{11}	1.79(0.0883)	1.79(0.117)	1.79(0.108)
β_{12}	2.01(0.132)	2.01(0.168)	2.01(0.158)
β_{21}	0.0909(0.0419)	0.0909(0.0512)	0.0898(0.0517)
β_{22}	− 0.160(0.0662)	− 0.160(0.0813)	− 0.163(0.0784)
γ_1	− 4.26(0.733)	− 4.26(0.733)	− 4.37(0.826)
γ_2	− 0.0726(0.178)	− 0.0726(0.178)	− 0.0844(0.179)
ϕ	–	1.41(0.130)	–
ω	–	–	0.0680(0.0243)
AIC	1242.8	–	1231.8

6.2 Whitefly data

To illustrate our methodology for fitting the ZI models to clustered data with the ES algorithm, we consider a data set from a horticultural experiment in which several methods of applying pesticide to control silverleaf whiteflies on greenhouse raised poinsettia were examined. These data set were used by Hall (2000) to illustrate the ZIP and ZIB models with random effects to account for clustering in the data (van Iersel *et al.*, 2000). Here we fit marginal ZIB models, but account for within-cluster correlation through working correlation structures in the GEEs that appear in the S step of the ES algorithm rather than through cluster specific random effects.

The experimental design that generated the data was a randomized complete block design with weekly repeated measures over 12 weeks. The experimental unit in this study was a trio of poinsettia plants, and 18 such units (54 plants) were randomized to six treatments in three complete blocks. Each week, clip-on leaf cages were used on one leaf per plant to hold a fixed number of whiteflies on or near the plant. The number of surviving insects (S) was measured two days after enclosure in the leaf cage and following treatment. Thus, the experiment generated a bounded count in a randomized complete block design with repeated measures.

Let S_{ijk} denote the number of surviving insects out of m_{ijk} insects placed on all three plants in the experimental unit assigned to the ith treatment in the jth block measured at week k. To facilitate presentation of results, we consider relatively simple ZIB models for these data. In particular, we assume only main effects in the binomial portion of the mixture and we assume a constant mixing probability. Thus, for S_{ijk} we assume a marginal ZIB model with $\text{logit}(\pi_{ijk}) = \mu + \text{block}_j + \text{trt}_i + \beta\text{week}_k$, $\text{logit}(p_{ijk}) = \gamma$, where here week is treated as a continuous covariate and the other terms represent factor effects.

Results from fitting this model with an ordinary ZIB model (ignoring within-cluster correlation) and marginal ZIB models with various choices of the working correlation matrix **P** appear in Table 3. In all cases, we set the working correlation matrix for the

Table 3 Parameter estimates (standard errors) for ZIB and ZIB-GEE models fit to whitefly data

		ZIB-GEE		
Parameter	ZIB	Indep.	Exchangeable	AR(1)
μ	− 1.21(0.122)	− 1.21(0.167)	− 1.21(0.169)	− 1.18(0.172)
block 1	− 0.460(0.105)	− 0.460(0.191)	− 0.457(0.193)	− 0.469(0.197)
block 2	− 0.0483(0.100)	− 0.0483(0.166)	− 0.0491(0.166)	− 0.0539(0.170)
trt 1	− 0.496(0.136)	− 0.496(0.113)	− 0.497(0.113)	− 0.487(0.124)
trt 2	− 0.302(0.132)	− 0.302(0.213)	− 0.304(0.213)	− 0.312(0.223)
trt 3	− 0.545(0.156)	− 0.545(0.0972)	− 0.539(0.0921)	− 0.546(0.106)
trt 4	− 0.269(0.137)	− 0.269(0.242)	− 0.269(0.243)	− 0.292(0.254)
trt 5	3.18(0.123)	3.18(0.207)	3.17(0.218)	3.16(0.217)
week	0.0130(0.0112)	0.0130(0.0260)	0.0129(0.0260)	0.0088(0.0258)
γ	− 1.14(0.164)	− 1.14(0.267)	− 1.14(0.267)	− 1.14(0.268)
ρ	–	–	− 0.0199(0.0284)	− 0.129(0.163)
ϕ	–	3.61(0.359)	3.60(0.357)	3.61(0.359)

The response variable here is S, the number of surviving whiteflies following treatment.

mixing mechanism, **R**, to the identity matrix. As in the previous example, parameter estimates for the ZIB model and for the ZIB-GEE model with independence working structure are identical. However, the standard errors differ substantially because the latter model takes into account the overdispersion in these data relative to a ZIB model ($\hat{\phi} = 3.61$) in the computation of the standard errors. In addition, the ZIB-GEE model with independence working structure also takes into account the within-cluster correlation in the computation of the standard errors through the use of the sandwich variance–covariance estimator. The ZIB-GEE models with other working structures account for within-cluster correlation and overdispersion in the estimation of the parameters and their standard errors. However, in this example very little within-cluster correlation exists. Therefore, differences between the results for the three working correlation structures are very small here.

7 Discussion

In this paper we have explored two approaches to fitting marginal zero inflated regression models to longitudinal data with excess zeros. Direct application of GEEs to the marginal ZI model seems a natural, although perhaps naive, approach to take in the clustered data context in which a full likelihood based approach is infeasible but moment assumptions can be made. It has been seen that a first order, or linear, GEE approach is unproductive since it is not possible to identify regression parameters pertaining to the mixing probability and second component mean from the marginal mean of the observed response. It is necessary to introduce second order, or quadratic, estimating functions to fit the ZI model with this approach, but efficiency is still low. The problem here seems to be that the mixture structure of the model is not captured adequately by one or two marginal moments. Undoubtedly, greater efficiency can be achieved by introducing estimating functions based on higher order marginal moments. However, this will require difficult and unwanted assumptions concerning the third and higher order within-cluster dependence structure in the data.

We feel that incorporation of GEEs in the ES algorithm is a much better alternative because it is closer, in some sense, to the ML estimation. This approach is an application of methodology of Rosen *et al.* (2000), although those authors limit their discussion to exponential dispersion family components in the mixtures they consider. We point out that models outside this class, such as the ZI censored lognormal regression models, can also be handled. An important issue not addressed here is selection of a suitable working correlation structure for the GEEs that appear in the ES algorithm. One possible approach is to develop a model selection criterion similar to that of Pan (2001), appropriate for this context. This idea will be pursued elsewhere.

As in the use of GEEs in non-ZI contexts, the choice of working correlation structures in the ES algorithm will affect the efficiency but not the consistency of our parameter estimators. Exactly how much efficiency is lost is a subject for future research, but we can expect similar results to those found for GEEs in non-ZI contexts (Davis, 2002: Hall and Severini, 1998; Section 9.5.7; and references therein). Of course, the amount of efficiency loss depends on the degree of working correlation misspecification. In the

whitefly example, we chose to use the working independence structure for **R**, the within subject correlation matrix for the mixing mechanism. This was done for simplicity and because, intuitively, we expect that the missing mechanism **u** will often be subject to less correlation than the prezero inflation response **z**. Furthermore, the missing mechanism is completely latent, whereas **z** is only partially unobserved. This fact, combined with the usual situation in which the mixing probability is small, implies that there will typically be less information in the data concerning **u** than **z**. Therefore, we believe that it is generally good practice to be especially parsimonious in the modeling of **p** and **R** relative to the modeling of parameters of the nondegenerate component of the mixture.

Acknowledgements

The authors wish to thank two anonymous referees whose comments led to improvements in the paper.

References

Berk KN and Lachenbruch PA (2002) Repeated measures with zeros. *Statistical Methods in Medical Research* **11**, 303–16.

Breckling JU, Chambers RL, Dorfman AH, Tam SM and Welsh AH (1994) Maximum likelihood inference from sample survey data. *International Statistical Review* **62**, 349–63.

Crowder M (1987) On linear and quadratic estimating functions. *Biometrika* **74**, 591–97.

Davis CS (2002) *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Dobbie MJ and Welsh AH (2001) Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* **43**, 431–44.

Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 56, 1030–39.

Hall DB (2001) On the application of extended quasilikelihood to the clustered data case. *Canadian Journal of Statistics* **29**, 77–97.

Hall DB and Berenhaut KS (2002) Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Canadian Journal of Statistics* **30**, 77–97.

Hall DB and Severini TA (1998) Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association* **93**, 1365–75.

Heilbron DC (1994) Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**, 531–47.

van Iersel M, Oetting R, and Hall DB (2000) Imidacloprid applications by subirrigation for control of silverleaf whitefly (Homoptera: Aleyrodidae) on poinsettia. *Journal of Economic Entomology* **93**, 813–19.

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Liang KY and Zeger SL (1986) Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.

Liang KY, Zeger SL and Qaqish B (1992) Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

Lipsitz SR, Molenberghs G, Fitzmaurice GM and Ibrahim J (2000) GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* **56**, 528–36.

McCullagh P and Nelder JA (1989) *Generalized linear models; 2nd edition*. Boca Raton: Chapman & Hall/CRC.

Moulton LH, Curriero FC and Barroso PF (2002) Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* **11**, 317–25.

Moulton LH and Halsey NA (1995) A mixture model with detection limits for regression analyses of quantitative assay data. *Biometrics* **51**, 1570–78.

Olsen MK and Shafer JL (2001) A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–45.

Pan W (2001) Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–25.

Prentice RL and Zhao LP (1991) Estimating equations for parameters in mean and covariances in multivariate discrete and continuous responses. *Biometrics* **47**, 825–39.

Ridout M, Hinde J and Demétrio CGB (1998) *Models for count data with many zeros*. Invited Paper, The XIXth International Biometric Conference, Cape Town, South Africa, 179–92.

Ridout M, Hinde J and Demétrio CGB (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57**, 219–23.

Rosen O, Jiang W and Tanner MA (2000) Mixtures of marginal models. *Biometrika* **87**, 391–404.

Searle SR (1982) *Matrix algebra useful for statistics*. New York: John Wiley & Sons.

Sim CH (1993) Generation of Poisson and gamma random vectors with given marginals and covariance matrix. *Journal of Statistical Computation and Simulation* **47**, 1–10.

Thall PF and Vail SC (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–71.

Tooze JA, Grunwald GK and Jones RH (2002) Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* **11**, 341–55.

Vieira AMC, Hinde JP, and Demétrio CGB (2000) Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics* **27**, 373–89.

Yau KKW and Lee AH (2001) Zero-inflated Poisson regression with random effects to evaluate an occupations injury prevention programme. *Statistics in Medicine* **20**, 2907–20.