Published in final edited form as:

J Stat Plan Inference. 2010 July 1; 140(7): 1939–1947. doi:10.1016/j.jspi.2010.01.034.

A Self-consistency Approach to Multinomial Logit Model with Random Effects

Shufang Wang and Alex Tsodikov[†]

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

Abstract

The computation in the multinomial logit mixed effects model is costly especially when the response variable has a large number of categories, since it involves high-dimensional integration and maximization. Tsodikov and Chefo (2008) developed a stable MLE approach to problems with independent observations, based on generalized self-consistency and quasi-EM algorithm developed in Tsodikov (2003). In this paper, we apply the idea to clustered multinomial response to simplify the maximization step. The method transforms the complex multinomial likelihood to Poisson-type likelihood and hence allows for the estimates to be obtained iteratively solving a set of independent low-dimensional problems. The methodology is applied to real data and studied by simulations. While maximization is simplified, numerical integration remains the dominant challenge to computational efficiency.

Keywords

QEM algorithm; Multinomial logit model with random effects

1 Introduction

Multinomial logit model with random effects is a common choice in the analysis of correlated nominal data in biomedical science. Such correlation could come from repeated measures or clustered observations. The complex form of the likelihood function and the presence of random effects make the computation costly. The presence of random effects implies computationally expensive multi-dimensional integrals.

A number of approaches have been proposed in the literature to overcome computational difficulties both in standard multinomial logit model and multinomial logit mixed effects model. Breslow and Clayton (1993) advocated penalized quasi-likelihood estimation approach to avoid the complex form of multinomial likelihood. Chen and Kuo (2001) used the fact that the multinomial distribution can be derived from a set of Poisson random variables conditionally on their total being fixed (McCullagh and Nelder (1989)) and suggested transforming the multinomial problem to Poisson log-linear or non-linear model. Lang (1996) showed that the idea can be interpreted as using Lagrange multipliers to normalize multinomial probabilities. Because normalization needs to be enforced for each distinct

© 2010 Elsevier B.V. All rights reserved.

[†]Tel: 734-764-5450, Fax: 734-763-2215, tsodikov@umich.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

covariate pattern, the Poisson log-linear transformation is restricted to discrete covariates. This introduces high-dimensional augmentation adding nuisance parameters. It is convenient that this method can be implemented with standard software such as SAS NLMIXED. Chen and Kuo (2001) and Kuss and McLerran (2007) also considered a general non-linear modeling approach in conjunction with SAS NLMIXED. A related series of methods used multinomial likelihood directly applying numerical approximation to multidimensional integrals, such as Gaussian quadrature (Rabe-Hesketh et al. (2002), Hedeker (2003) etc.), spherical-radial quadrature (Clarkson and Zhan (2002)), first-order Taylor series expansion of the integrand (Breslow and Clayton (1993)), and Bayesian methods (Daniels and Gatsonis (1997)). Tsodikov and Chefo (2008) introduced artificial mixing variables to transform the multinomial likelihood to Poisson-type likelihood and reduce the complexity of the likelihood function. They apply EM-type algorithm that enjoys factorization of the model dimension at the M step that represents Poisson regressions. In this paper, we extend the method to the multinomial logit model with random effects. Treating random effects and the artificial variables as missing data, we apply the generalized self-consistency approach described in Tsodikov (2003) (Quasi-EM algorithm) to parameter estimation. The key benefit of this approach is that the M-step reduces to a set of low-dimensional problems as described in Section 3.

EM-type methods are advocated because of their stability with complex models. As pointed out in Hartzel et al. (2001), quadrature method for integration is feasible only for integrals with dimensions up to about 5 or 6. There may be a slight increase in the computation capacity these days, while it is still true that multivariate quadrature method is not computationally feasible for evaluating high-dimensional integrals. A number of Monte Carlo(MC) EM procedures were recently proposed in the literature under a more general framework - generalized linear mixed models, such as McCulloch (1997), Booth and Hobert (1999) and Chen et al. (2002), where Monte Carlo approximation for integration and EM algorithm are synthesized to obtain consistent parameter estimates. For ease of evaluating Monte Carlo error and automatically adjusting Monte Carlo sample size, M , Booth and Hobert (1999) proposed to use the MC approximations to high-dimensional integrals with independent random samples. As a consequence, M can be automatically increased until algorithm converges. Brian S. Caffo (2005) developed a more general method to increase the MC sample size which requires less computation, while it roughly preserves the ascent property of the observed likelihood function.

Unlike MCEM in McCulloch (1997), automated MCEM in Booth and Hobert (1999) and ascent-based MCEM in Brian S. Caffo (2005), where the missing data contain only the random effects, in this paper, we introduce artificial random variables and treat them, together with the random effects, as missing data. We then use MC approximation to high-dimensional integrals with independent random samples in E-step. Our approach is different in the treatment of the M-step - the introduction of artificial variables simplifies and factorizes the complex likelihood function in the M-step which contributes to greater computational efficiency in a mixed multinomial subclass of generalized linear mixed models. To accelerate the MCEM algorithm, we propose a simple method to increase M , which does not require us to evaluate the MC error. At the same time, it roughly preserves the ascent property of observed likelihood function, since it is very similar to a special case of Brian S. Caffo (2005), as shown later in the paper.

In Section 2, we specify the multinomial logit model with mixed effects. In Section 3, we specify estimation steps. Variance estimation is described in Section 4. We summarize the detailed estimation procedure in Section 5, followed by simulation studies in Section 6. The method is applied to a real dataset in Section 7.

2 Model specification

Suppose the categorical response, Y_{ij}^* , has R categories, indexed as $r(r = 1, \dots, R)$. Clusters of correlated data are indexed by i ($i = 1, \dots, I$). Repeated measures within a cluster are indexed by j ($j = 1, \dots, T_i$). Let $Y_{ijr} = I(Y_{ij}^* = r)$. Hence, response category probability

$p_{ijr} = E[Y_{ijr}] = P(Y_{ij}^* = r)$. Let $\mathbf{X}_{ij}$ denote the column vector of exploratory variables for the j -th observation in the i -th cluster, in which the first element is one, corresponding to the intercept in model (2.1). And let $\mathbf{Z}_{ij}$ denote the column vector of random effects for the j -th observation in the i -th cluster.

Assume category R is the reference category. The logits compare any category $r = 1, \dots, R - 1$ with the reference category. The model is called reference-cell logit random effects model as described in Agresti (2002).

$$\log\left(\frac{p_{ijr}}{p_{ijR}}\right) = \mathbf{X}_{ij}^T \boldsymbol{\beta}_r + \mathbf{Z}_{ij}^T \mathbf{b}_{ir}, \quad r = 1, \dots, R - 1, \quad (2.1)$$

where $\boldsymbol{\beta}_r$ is the fixed effects coefficient vector of length $p + 1$, corresponding to an intercept and p covariates. $\boldsymbol{\beta}_R$ is set to be zero. $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{R-1}$ are to be estimated. $\mathbf{b}_{ir}$ is the random effects coefficient vector. The assumption on the distribution of $\{\mathbf{b}_{ir}\}_{r=1}^{R-1}$ is arbitrary and our choice will be specified in Section 3.

Define $\mathbf{Y}_{ij}$ as a column vector with r^{th} element $(Y_{ij})_r$ being Y_{ijr} . Under a GLM setting, p_{ijr} is linked with the linear predictor, $\eta_{ijr} = \mathbf{X}_{ij}^T \boldsymbol{\beta}_r + \mathbf{Z}_{ij}^T \mathbf{b}_{ir}$, through the reference-cell logit function

$$g(\cdot) = \log\left(\frac{p_{ijr}}{p_{ijR}}\right).$$

$$p_{ijr} = \frac{\theta_{ijr}}{1 + \sum_{l=1}^{R-1} \theta_{ijl}}, \quad r = 1, \dots, R - 1 \quad (2.2)$$

In the rest of the paper, we use the following notations: $\theta_{ijr} = \exp\{\eta_{ijr}\}$. $\mathbf{b}_i$ is a vector consisting of all elements in $\{\mathbf{b}_{ir}\}_{r=1}^{R-1}$. $\mathbf{Y}_i = \{\mathbf{Y}_{ij}\}_{j=1}^{T_i}$. E_X indicates statistical expectation taken with respect to random variable X .

3 Estimation

3.1 Artificial Mixture

Tsodikov and Chefo (2008) introduced artificial variables $\{U_{ij}\} \stackrel{iid}{\sim} \text{Exp}(1)$ to standard multinomial logit model, to transform the multinomial likelihood to Poisson-type likelihood. This approach also works when repeated measures are present. For standard exponentially distributed variable, U , the Laplace transform has the form

$$\mathcal{L}(s) = E_U \{e^{-Us}\} = \frac{1}{1+s}. \quad (3.1)$$

Observing the similarity between (3.1) and the multinomial probabilities (2.2), we can write p_{ijr} in the artificial mixture form

$$p_{ijr} = \theta_{ijr} E_{U_{ij}} \left\{ \exp \left[-U_{ij} \sum_{l=1}^{R-1} \theta_{ijl} \right] \right\}.$$

If we pretend that $U_{ij}(\forall i, j)$ are observed, the expectation sign, E , can be removed in the above equation. Denote the complete-data expression by $\tilde{p}_{ijr}(\cdot, U_{ij})$.

$$\tilde{p}_{ijr} = \theta_{ijr} \exp \left\{ -U_{ij} \sum_{l=1}^{R-1} \theta_{ijl} \right\}. \quad (3.2)$$

It is clear that $\tilde{p}_{ijr}(\cdot, U_{ij})$ is no longer a legitimate probability. This indicates that the likelihood function constructed from $\tilde{p}_{ijr}(\cdot, U_{ij})$ is not a legitimate probabilistic model. The self-consistency theory developed in Tsodikov (2003) justifies the approach. Denote $\mathbf{U}_i$ as a column vector, $\mathbf{U}_i = (U_{i1}, \dots, U_{iT_i})^T$.

3.2 Complete-data Likelihood

For model (3.2) with the artificial variables $\{U_{ij}\}$, the complete data include $\{\mathbf{X}_{ij}, \mathbf{Y}_{ij}, \mathbf{b}_{ir}, U_{ij}, \forall i, j, r\}$. The complete data likelihood function is

$$L_{CD}(\beta, \Sigma) = \prod_i f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{U}_i; \beta, \Sigma) = \prod_i f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{U}_i; \beta) \times f(\mathbf{b}_i) \times f(\mathbf{U}_i),$$

where $f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{U}_i; \beta) = \prod_{j,r} \left\{ \tilde{p}_{ijr} \right\}^{y_{ijr}}$ is the conditional multinomial density for the i -th cluster with illegitimate probabilities $\left\{ \tilde{p}_{ijr} \right\}$. It can be derived that

$$\log \{f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{U}_i; \beta)\} = \sum_{j,r} \{y_{ijr} \log(\theta_{ijr}) - U_{ij} \theta_{ijr}\}$$

Under the following assumptions: (1) $\{U_{ij}\} \stackrel{iid}{\sim} \text{Exp}(1)$; (2) U_{ij} and $\mathbf{b}_{ir}$ are independent for any j, r ; and (3) $\mathbf{b}_i$ follows multivariate normal distribution, $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \Sigma)$, the complete data likelihood can be written as

$$L_{CD}(\beta, \Sigma) = \prod_i \{f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{U}_i; \beta) \times f(\mathbf{b}_i; \Sigma) \times f(\mathbf{U}_i)\}. \quad (3.3)$$

Denote the corresponding log-likelihood function as ℓ_{CD} and the contribution of the i -th cluster to the log likelihood function $\ell_{CD,i}$.

In the likelihood function (3.3), $\{\mathbf{b}_i\}$ and artificial variables $\{U_{ij}\}$ can be treated as missing data, hence EM-type algorithm can be used for parameter estimation.

3.3 Estimation steps

Let $\phi = (\beta, \Sigma)$ be the combined parameter vector. The EM algorithm is an iterative process. Given parameter estimates from the k -th iteration, the $(k+1)$ -th iteration can be formulated as $\text{Argmax}_{\phi} \{Q(\phi | \phi^{(k)})\}$, where $Q(\phi | \phi^{(k)})$ forms the E-step.

$$Q(\phi | \phi^{(k)}) = E \left\{ \log (L_{\text{CD}}(\phi)) | y; \phi^{(k)} \right\}. \quad (3.4)$$

The M-step finds $\phi^{(k+1)}$ such that

$$Q(\phi^{(k+1)} | \phi^{(k)}) \geq Q(\phi | \phi^{(k)}), \quad (3.5)$$

for any ϕ in the parameter space.

To evaluate the expectation in E-step, we need to evaluate the multi-dimensional integrals with respect to b_i and U_i .

$$Q(\phi | \phi^{(k)}) = E \left\{ \ell_{\text{CD}} | y; \phi^{(k)} \right\} = \sum_i \left\{ \int \ell_{\text{CD},i} \times f(b_i, U_i | y_i, \phi^{(k)}) db_i dU_i \right\},$$

where $f(b_i, U_i | y_i, \phi^{(k)}) \propto f(y_i | b_i, U_i, \beta^{(k)}) \times f(b_i; \Sigma^{(k)}) \times f(U_i)$. The normalizing constant is $\int f(y_i | b_i, U_i; \beta^{(k)}) \times f(b_i; \Sigma^{(k)}) \times f(U_i) db_i dU_i$, denoted as $a_i^{(k)}$.

Note that the dimension of the integrals easily exceeds the limit that the computer can handle using quadrature methods when either R or the number of random effects is large. In this case, Monte Carlo (MC) approximations to the intractable integrals are often suggested. Booth and Hobert (1999) and Hartzel et al. (2001) proposed to use rejection sampling or importance sampling based on $f(b_i, U_i | y_i, \phi^{(k)})$ as it's not easy to directly sample from f . By doing so, one does not need to evaluate the normalizing constants $a_i^{(k)}$. In this paper, we randomly select M independent samples from multivariate normal distribution with mean zero and covariance matrix $\Sigma^{(k)}$ for b_i and from standard exponential distribution for U_{ij} . Denote these samples as $b_{i,m}, U_{ij,m}$ for $m = 1, \dots, M$. Hence, $Q(\phi | \phi^{(k)})$ can be approximated by $\tilde{Q}(\phi | \phi^{(k)})$, and $a_i^{(k)}$ by $\tilde{a}_i^{(k)}$.

$$\tilde{Q}(\phi | \phi^{(k)}) = \frac{1}{M} \sum_i \sum_{m=1}^M \left\{ \frac{f(y_i, b_{i,m}, U_{i,m}; \beta^{(k)})}{\tilde{a}_i^{(k)}} \log f(y_i, b_{i,m}, U_{i,m}; \phi) \right\}, \quad (3.6)$$

where $\tilde{a}_i^{(k)} = \frac{1}{M} \sum_{m=1}^M f(y_i | b_{i,m}, U_{i,m}; \beta^{(k)})$. Compared to rejection sampling and/or importance sampling methods widely used in the literature on generalized mixed effects models such as Booth and Hobert (1999) and Hartzel et al. (2001), the main advantage of our sampling scheme is that one does not need to look for a trial distribution as we sample from the exact distributions, $MVN(\mathbf{0}, \Sigma^{(k)})$ and $\text{Exp}(1)$. The method synthesizes MC approximation and QEM algorithm developed in Tsodikov (2003), hence we will call it MCQEM algorithm to differentiate it from a regular M-step procedure.

Denote $w_{i,m}^{(k)} = \frac{f(\mathbf{y}_i | \mathbf{b}_{i,m}, \mathbf{U}_{i,m}; \boldsymbol{\beta}^{(k)})}{\tilde{a}_i^{(k)}}$. The M-step becomes

$$\underset{\boldsymbol{\beta}, \boldsymbol{\Sigma}}{\text{Argmax}} \left\{ \sum_{i,m} w_{i,m}^{(k)} \{ \log f(\mathbf{y}_i | \mathbf{b}_{i,m}, \mathbf{U}_{i,m}; \boldsymbol{\beta}) + \log f(\mathbf{b}_i; \boldsymbol{\Sigma}) + \log f(\mathbf{U}_i) \} \right\}$$

We obtain $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\Sigma}^{(k+1)}$ with a two-stage procedure. First, we estimate $\boldsymbol{\beta}^{(k+1)}$ with $\boldsymbol{\Sigma}$ fixed at $\boldsymbol{\Sigma}^{(k)}$, achieved by maximizing the first term in $\tilde{Q}(\phi | \phi^{(k)})$, which is the log-likelihood function as if $\mathbf{b}_{i,m}$, $\mathbf{U}_{i,m}$ are observed. Plugging in $\log f(\mathbf{y}_i | \mathbf{b}_i, \mathbf{U}_i; \boldsymbol{\beta})$ specified early, the objective function in this step can be written as

$$\sum_{i,m} w_{i,m}^{(k)} \{ \log f(\mathbf{y}_i | \mathbf{b}_{i,m}, \mathbf{U}_{i,m}; \boldsymbol{\beta}) \} = \sum_r \sum_{i,m} w_{i,m}^{(k)} \sum_j \{ y_{ijr} (\mathbf{X}_{ij}^T \boldsymbol{\beta}_r + \mathbf{Z}_{ij}^T \mathbf{b}_{ir}) - U_{ij} \exp(\mathbf{Z}_{ij}^T \mathbf{b}_{ir}) \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}_r) \}.$$

It's obvious that we can view the objective function as a sum of $R-1$ components, each corresponding to the contribution of a Poisson likelihood with $I\{y_{ij} = r\}$ as the response and $U_{ij,m} \exp(\mathbf{Z}_{ij}^T \mathbf{b}_{ir,m})$ as the offset term. As a consequence, we can obtain $\boldsymbol{\beta}_r^{(k+1)}$ ($\forall r$) from a weighted Poisson regression with weight $w_{i,m}^{(k)}$. For each Poisson regression, the original data are replicated M times.

Second, we estimate $\boldsymbol{\Sigma}^{(k+1)}$ with $\boldsymbol{\beta}$ fixed at $\boldsymbol{\beta}^{(k+1)}$. This step only relates to the second term in $\tilde{Q}(\phi | \phi^{(k)})$, which is simply the likelihood contribution of a linear combination of multivariate normal samples with weight $w_{i,m}^{(k)}$.

Suppose we have p exploratory variables. Then, we have $(p+1) \times (R-1)$ coefficient parameters to estimate in total. The M-step of our method estimates $p+1$ parameters at a time, and repeats the process $R-1$ times. In contrast, without introducing the artificial variables, one has to estimate all $(p+1) \times (R-1)$ coefficient parameters in one step. In this sense, our method results in greater simplicity in the M-step. This advantage is demonstrated by a simulation study in Section 6.

The process iterates between E-step and M-step until a convergence criterion is met. In this paper, the process will stop when the relative change of parameter estimates does not exceed a given tolerance, for instance, $1e-8$.

3.4 Escalation of the MC sample size

Convergence may not be obtained when using a constant MC sample size, M , due to a persistent MC error (Brian S. Caffo (2005)). On the other hand, it's not efficient to use large M at an early stage of the EM iterations. Sometimes, it's critical to increase M automatically over iterations. Booth and Hobert (1999) proposed to use normal approximation to MC error and construct a $100(1 - \alpha)\%$ confidence ellipsoid for $\phi^{(k+1)}$ and hence determine to increase M or not, by comparing whether $\phi^{(k)}$ lies in the confidence ellipsoid. Brian S. Caffo (2005) developed a data-driven strategy for increasing M which preserves the ascent property of the likelihood function over iterations with large probabilities. They constructed an approximate confidence interval for the change of $Q(\phi^{(k+1)} | \phi^{(k)})$ over two consecutive iterations and hence determine whether to increase M .

Methods proposed in the aforementioned articles all require a large amount of computation such as approximating MC error, second-order derivatives of $Q(\phi | \phi^{(k)})$ and so on. There does not exist standard software to implement these methods. It's obviously tedious to derive these formulas for those who want only to apply these methods to real problems. In this paper, we suggest a simpler strategy for the escalation of M , which is similar to a special case as in Brian S. Caffo (2005) with $\alpha=0.5$. We increase M by a given factor whenever observed likelihood function at $\phi^{(k+1)}$ is less than that evaluated at $\phi^{(k)}$. The observed likelihood function can be easily approximated using MC method.

4 Variance Estimation

Louis (1982) described a variance estimation method based on the observed information matrix

for EM algorithm. Let S be the complete data score vector. $S = \frac{\partial \log L_{CD}}{\partial \phi}$, where L_{CD} is given by (3.3). The observed information matrix, I_{obs} , can be estimated through the formula: $I_{obs} =$

$E_{U, \mathbf{b}} \{I_{CD}(U, \mathbf{b}) - SS^T(U, W) | L_{CD}(U, \mathbf{b})\}$, where $I_{CD} = -\frac{\partial^2 \log L_{CD}}{\partial \phi \partial \phi^T}$ is the complete data information matrix. I_{CD} overestimates the sample information as it assumes that missing data are known. The expression for I_{obs} represents the so-called missing information principle that observed information is complete information minus the missing information. Inverting the observed information matrix, we get the estimated covariance matrix. The procedure is tedious as it requires the calculation of the first and second derivatives. Hence, we seek for an alternative method to estimate the covariance matrix.

We estimate the variance components based on an approximation to the likelihood curvature by a quadratic form fit by linear regression based on the sampled points on likelihood surface around MLE. Related idea was used in Neilsen et al. (1992).

Applying the Taylor expansion to the log-likelihood function $\ell(\phi)$ at ϕ_{MLE} up to the second order term, we get

$$\ell(\phi) \approx \ell(\phi_{MLE}) + (\phi - \phi_{MLE})^T G(\phi_{MLE}) + (\phi - \phi_{MLE})^T D(\phi_{MLE})(\phi - \phi_{MLE})/2.$$

where $G(\cdot)$ and $D(\cdot)$ are the gradient vector and the Hessian matrix, respectively.

Since $G(\phi_{MLE}) \approx 0$,

$$\ell(\phi) - \ell(\phi_{MLE}) \approx (\phi - \phi_{MLE})^T D(\phi_{MLE})(\phi - \phi_{MLE})/2.$$

Denote $\ell(\phi) - \ell(\phi_{MLE})$ as $\Delta\ell$ and $\phi - \phi_{MLE}$ as $\Delta\phi$. Let $\{D(\cdot)\}_{ij} = d_{ij}$. If we randomly sample K points for ϕ within the close neighborhood of ϕ_{MLE} , we then have

$$\Delta\ell_k = \frac{1}{2} \sum_{i,j} d_{ij} \Delta\phi_{ki} \Delta\phi_{kj} + \epsilon_k, \forall k \in 1, \dots, K, \quad (4.1)$$

where $\Delta\phi_{ki}$ is the i^{th} element of $\Delta\phi$ of the k^{th} sampled point on the log-likelihood surface. The distribution of ϵ_k is Gaussian with mean zero.

We can estimate d_{ij} in the above equation by minimizing the following sum of squares:

$$SS = \sum_k \left\{ \Delta \ell_k - \frac{1}{2} \sum_{i,j} d_{ij} \Delta \phi_{ki} \Delta \phi_{kj} \right\}^2 = \sum_k \left\{ \Delta \ell_k - \frac{1}{2} \sum_i d_{ij} \Delta \phi_{ki}^2 - \sum_{i < j} d_{ij} \Delta \phi_{ki} \Delta \phi_{kj} \right\}^2.$$

Observe that the above sum of squares has the form of objective function in linear regression with response being $\Delta \ell_k$, independent variables $\frac{1}{2} \Delta (\phi_i^2 (i=1, \dots, m))$ and $\Delta \phi_i \Delta \phi_j$ for any $i < j$, and the $m(m+1)/2$ unknown elements in matrix D being the corresponding regression coefficients. Or we can directly recognize (4.1) as a linear regression model without intercept term. Denote the unknown elements in the matrix D as a vector, γ . Therefore, γ can be estimated, when K is large enough such that the design matrix of the linear regression model is non-singular. In other words, the Hessian matrix can be estimated through a multiple linear regression through the origin. Hence, approximations of the standard errors of the parameter estimates can be obtained easily.

Elements of the $K \Delta \phi$ points are sampled independently from a symmetric distribution with mean 0 on the log-likelihood surface. They need to be close enough to ϕ_{MLE} but not too close

causing numerical concerns. They are chosen to be in the $\frac{1}{\sqrt{n}}$ -range around ϕ_{MLE} .

The remaining part of the method is to solve (4.1). Simulation study in Section 6 suggests that the variance approximation method works well.

5 Algorithm

In this section, we outline the MCQEM algorithm in details.

1. Set initial values for β , Σ and MC sample size M . We may set $\beta^{(0)}$ to be the coefficient estimates by fitting a standard multinomial logit model, ignoring the within-cluster correlation. It is a good set of starting values especially when the within-cluster correlation is small. $\Sigma^{(0)}$ can be identity matrix. $M^{(0)}=200$. Let $k=0$.
2. Calculate the observed log-likelihood function at $\phi^{(k)}$, denoted as $\ell_{obs}(\phi^{(k)})$.
3. Randomly draw $M^{(k)}$ independent samples from $MVN(\mathbf{0}, \Sigma^{(k)})$ for $b_{i,m}$ and from Exp (1) for $U_{ij,m}$.
4. Calculate the weight $w_{i,m}^{(k)}$ and offset $U_{ij,m} \exp(Z_{ij,m}^T b_{ir})$. Estimate β (denoted as $\beta^{(new)}$) through $R-1$ Poisson regressions with the above weight and offset and $\Sigma^{(new)}$ from a set of multivariate random samples with the weight.
5. Evaluate the observed log-likelihood function at $\phi^{(new)}$, denoted as $\ell_{obs}(\phi^{(new)})$ and compare it with $\ell_{obs}(\phi^{(k)})$. If $\ell_{obs}(\phi^{(k)}) < \ell_{obs}(\phi^{(new)})$, accept $\phi^{(new)}$ as $\phi^{(k+1)}$, update k , $\ell_{obs}(\phi^{(k)})$ and $M^{(k+1)} = M^{(k)}$. Go back to step 3. Otherwise, increase $M^{(k)}$ by a give factor, for instance, 5/4, without updating other quantities and go back to Step 3.
6. Iterate over these steps until convergence.
7. Estimate the Hessian matrix as described in Section 4 and obtain the standard errors.

6 Simulation studies

In this section, two simulation studies are performed. First, a simulation study is conducted to illustrate the benefit of introducing artificial variables to simplify the M-step. Second, we do

a simulation study to illustrate the MCQEM algorithm using a random intercept multinomial logit model.

6.1 Simulation study 1

The introduction of artificial variables makes MCQEM different from MCEM in estimating the parameters at the M-step. In the usual MCEM algorithm, a standard multinomial logit model is fitted for the M-fold duplicated dataset, treating simulated $\mathbf{b}_{i,m}$ as observed. All $(p+1) \times (R-1)$ coefficient parameters are estimated jointly. In contrast, in MCQEM algorithm, we fit $(R-1)$ independent Poisson regressions, each having $(p+1)$ parameters as if $\mathbf{b}_{i,m}$ and $\mathbf{U}_{i,m}$ were observed. This simulation study is designed to show the advantage of doing so.

We simulate three independent random variables of size 1000 from standard normal distribution, and dichotomize the third variable using zero as a cutpoint. We vary the number of categories in the $R = 3, \dots, 15$ range to highlight the computational benefit achieved at the M-step. The set of true parameter values for $R = 15$ is a 4 by 14 matrix. For $R < 15$ we use the corresponding $R-1$ column submatrix.

For a given R , we apply two methods to twenty simulated datasets. The first method is to estimate all the parameters jointly in a multinomial logit model. The second is to fit $R-1$ Poisson regressions with prespecified artificial variables, which are drawn from standard exponential distribution. The convergence criteria for both methods are set to be the same. The setting mimics what we have at the M-steps of MCEM and MCQEM. Hence, it can be used to compare the M-steps of these two methods.

Figure 1 shows the relationship between the number of response categories, R , and total time used in parameter estimation for all twenty simulated datasets.

As shown in the figure 1, fitting $R-1$ Poisson regressions always take less time than estimating all parameters jointly in the multinomial logit model. The advantage of reduced dimension increases rapidly with the number of categories R .

6.2 Simulation study 2

We explore the performance of the MCQEM algorithm and the variance approximation method for correlated multinomial data using a random intercept model. Four exploratory variables are included along with random intercepts with the following parameterization.

$$\log \left(\frac{p_{ijr}}{p_{ijR}} \right) = \mathbf{X}_{ij}^T \boldsymbol{\beta}_r + b_{i\alpha_r}, r=1, \dots, R-1, \quad (6.1)$$

where the random intercept b_{ir} is re-parameterized as $b_{i\alpha_r}$. This is a special case of the general model (2.1) with random intercepts only and full dependency between $b_{i1}, \dots, b_{i(R-1)}$.

We consider a categorical response with three categories, regressed on the covariates of the real data described in Section 7. In this data set, there are 36 clusters of size seven to ten. The true parameters used to simulate the responses are taken to be the parameter estimates of the real dataset obtained through the proposed MCQEM algorithm. It is compared with adaptive Gauss-Hermite quadrature methods using NLMIXED in SAS.

Table 1 summarizes results of 500 simulations. The second column shows the true parameter values used to simulate data. The third column lists the empirical mean of parameter estimates using MCQEM algorithm proposed in this paper. The first number in the parenthesis is the standard error of the parameter estimate using variance approximation method described in

Section 4. The second number in the parenthesis indicates the percentage of the estimated 95% confidence intervals covering the true parameter value. The fourth column contains those from NLMIXED in SAS, using adaptive Gauss-Hermite quadrature method to approximation the intractable integrals.

The empirical mean of parameter estimates of 500 simulations are very close to the true parameter values, taking the small sample size (36 clusters in total) into consideration. This supports the proposed methodology for multinomial logit model with random effects.

As seen in Table 1, among all the 95% confidence intervals constructed based on our standard error estimates, about 95% contain the true parameter values. The empirical means of the standard errors are also very close to the empirical standard errors of the parameter estimates. These evidences suggest that the variance approximation method works well and is stable. Numerical errors with evaluation of standard errors lead to problems less than 0.6% of all simulations.

Comparing the third and fourth column in Table 1, our method is competitive to adaptive Gauss-Hermite quadrature method for low-dimensional problems. It is worth pointing out again that quadrature methods are no longer feasible for high-dimensional integrals due to computation cost.

7 Application to a real data set

Kuss and McLerran (2007) analyzed a data set on physician's recommendations and preferences in traumatic brain injury (TBI) rehabilitation. They kindly made their data available for this paper. For each of multiple TBI disease histories, 36 physicians were asked to choose an optimal rehabilitation setting from the following: in-patient, day-clinic and out-patient. Four binary covariates were measured as well. The four covariates are answers to the following 4 questions: (1) Is the physician a neurologist? (2) Is the physician a specialist? (3) Is time since last event longer than three months? (4) Is the patient severely handicapped after TBI? The four covariates are named neuro, special, time and severity respectively. Recommendations within the same physician are expected to be correlated. Treating the rehabilitation setting as nominal, this forms a typical multinomial problem with repeated measures.

To identify factors that influence setting recommendations, we apply our method to this data set. Random intercepts only model with unstructured covariance matrix and a model in the form of (6.1) are compared by likelihood ratio test with a p-value 0.32, which shows no evidence against model (6.1) for this real dataset. Hence, we fit model (6.1) for the real dataset.

Results from our method are compared with (1) independence model, namely, standard multinomial logit model, ignoring the correlation within cluster (shown in Column 2 in Table 2); (2) model (6.1), implemented in SAS using PROC NLMIXED with Gauss-Hermite quadrature method for integral approximation (shown in Column 3 in Table 2). Estimates of the independence model are obtained through PROC LOGISTIC in SAS and results from PROC NLMIXED are based on SAS codes modified from those provided in Kuss and McLerran (2007). Results of our method are shown in MCQEM column in Table 2. In order to be consistent with Kuss and McLerran (2007), we choose in-patient as the reference category.

From Table 2, comparing the likelihood values, we see that the independence model does not provide as good a fit as the correlated model. For the MCQEM algorithm and PROC NLMIXED with Gaussian quadrature method for multi-dimensional integration, the parameter estimates are all very close. These two methods also provide almost the same maximum log-likelihood. The estimated standard errors from PROC NLMIXED and the quadratic variance estimation method proposed in this paper are also very close. The benefit of our method is that

it simplifies the M steps in estimating the coefficient parameters and avoids derivatives for variance estimation.

8 Discussion

The MCQEM algorithm proposed in this paper is a MCEM-type algorithm with artificial variables introduced to break the dimension of the M-step into a series of Poisson regression sub-problems. Artificial variables slightly increase the dimension of the E-step requiring larger MC sample size to achieve the same accuracy. Therefore, the MCQEM algorithm is best suited for problems with smaller cluster size. When the number of categories becomes larger, quadrature methods become prohibitively slow. While our method as well as MCEM would require a larger number of MC sample size M in this case, it still remains feasible with larger number of categories. The MCQEM algorithm converges even when the model is non-identifiable due to empty/sparse categories.

Similar to EM algorithm, the MCQEM algorithm does not provide variance estimates automatically. We propose the variance approximation method based on the idea of Neilsen et al. (1992), which only requires knowledge of the log-likelihood function and MLE. This allows us to get variance estimates avoiding taking derivatives. It works well when the log-likelihood function is approximately quadratic around the MLE. When the condition is not met, the variance estimates are not reliable. The traditional variance estimates are no good in this case either. We have examined a few extreme cases, where the parameters are near the boundary of the parameter space, and the log-likelihood function is far from quadratic. Neither our method nor numerical differentiation (SAS) is reliable in those extreme cases.

When missing observations are present, if the mechanism is missing at random (MAR) or missing completely at random (MCAR), the method of this paper could be extended to deal with missing observations by incorporating them into the EM framework. If missing is not at random, delicate care may be needed. Methods proposed for this purpose in the literature such as sensitivity analysis and multiple imputation (Garrett Fitzmaurice (2008) and Roderick J.A. Little (2002)) should apply to our setting.

Acknowledgments

The authors thank Oliver Kuss to allow us to use their dataset analyzed in Kuss and McLerran (2007). The authors are also grateful to the reviewers for important and helpful comments and suggestions.

References

- Agresti, A. Categorical Data Analysis. John Wiley and Sons; New Jersey: 2002.
- Booth J, Hobert J. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 1999;61:265–285.
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993;88:9–25.
- Caffo, Brian S.; Wolfgang Jank, GLJ. Ascent-based mceM. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 2005;67:235–251.
- Chen J, Zhang D, Davidian M. A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* 2002;3(3):347–360. [PubMed: 12933602]
- Chen Z, Kuo L. A note on the estimation of multinomial logit models with random effects. *The American Statistician* 2001;55:89–95.
- Clarkson DB, Zhan Y. Using spherical-radial quadrature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics* 2002;11:639–659.

- Daniels M, Gatsonis C. Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine* 1997;16:2311–2325. [PubMed: 9351167]
- Fitzmaurice, Garrett; Marie Davidian, GVGM. *Longitudinal Data Analysis*. Chapman and Hall/CRC; Boca Raton,FL: 2008.
- Hartzel J, Agresti A, Caffo B. Multinomial logit random effects models. *Statistical Modelling* 2001;1:81–102.
- Hedeker D. A mixed-effects multinomial logistic regression model. *Statistics in Medicine* 2003;22:1433–1446. [PubMed: 12704607]
- Kuss O, McLerran D. A note on the estimation of the multinomial logistic model with correlated responses in sas. *Computer Methods and Programs in Biomedicine* 2007;87:262–269. [PubMed: 17686544]
- Lang J. On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society, Ser. B* 1996;58:253–266.
- Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 1982;4(2):226–233.
- Mccullagh, P.; Nelder, J. *Generalized linear models*. Chapman and Hall; London: 1989.
- McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997;92:162–170.
- Neilsen G, Gill R, Andersen P, TI S. A counting process approach to maximum likelihood estimation in frailty models. *Scand J Statist* 1992;19:25–43.
- Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* 2002;2:1C2.
- Roderick, JA.; Little, DBR. *statistical analysis with missing data*. Wiley-interscience; Hoboken, NJ: 2002.
- Tsodikov A, Chefo S. Generalized self-consistency:multinomial logit model and poisson likelihood. *Journal of Statistical Planning and Inference* 2008;138:2380–2397.
- Tsodikov A. Semiparametric models: A generalized self-consistency approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 2003;65(3):759–774.

Figure 1. Number of response categories vs. time used per twenty estimation procedures (dashed line: multinomial logit model; solid line: $R - 1$ Poisson models with prespecified artificial variables)

Table 1

Summary statistics of parameter estimates obtained by QEM algorithm and standard errors by the variance approximation method described in Section 4, compared with adaptive Gauss-Hermite quadrature method through NLMIXED in SAS. Results are based on 500 simulations

Parameters	True parameter values	QEM(se, coverage rate*100)	NLMIXED(se, coverage rate*100)
α_1	1.652	1.609 (0.387, 95.2)	1.566 (0.365, 90.8)
α_2	1.293	1.253 (0.326, 94.8)	1.224 (0.307, 92.6)
β_{10}	-2.954	-3.049 (0.848, 93.8)	-3.057 (0.835, 96.4)
β_{11}	1.333	1.368 (1.007, 92.4)	1.370 (0.917, 92.4)
β_{12}	0.316	0.323 (0.903, 91.8)	0.320 (0.852, 92.0)
β_{13}	4.073	4.219 (0.611, 94.0)	4.220 (0.607, 94.8)
β_{14}	-2.567	-2.660 (0.556, 95.2)	-2.660 (0.543, 95.2)
β_{20}	-1.097	-1.133 (0.634, 94.4)	-1.138 (0.611, 95.0)
β_{21}	0.283	0.293 (0.812, 92.0)	0.293 (0.723, 90.6)
β_{22}	-0.428	-0.469 (0.712, 97.0)	-0.471 (0.706, 94.8)
β_{23}	2.390	2.452 (0.429, 95.0)	2.450 (0.424, 95.8)
β_{24}	-1.612	-1.642 (0.496, 94.8)	-1.642 (0.491, 95.2)

Table 2

Comparison of Parameter Estimates (IM: Independence Model; NLMIXED: Proc NLMIXED in SAS)

Parameters	IM	NLMIXED	MCQEM
α_{OP}	-	1.611(0.370)	1.652(0.442)
α_{DC}	-	1.246(0.311)	1.293(0.337)
(Out-patient)			
Intercept	-2.429(0.566)	-2.948(0.810)	-2.954(0.756)
Neuro	1.073(0.481)	1.319(0.910)	1.333(0.898)
Special	0.296(0.426)	0.280(0.851)	0.316(0.855)
Time	3.150(0.456)	4.088(0.573)	4.073(0.551)
Severity	-2.022(0.441)	-2.571(0.526)	-2.567(0.522)
(Day-clinic)			
Intercept	-0.879(0.430)	-1.073(0.595)	-1.097(0.580)
Neuro	0.088(0.395)	0.269(0.713)	0.283(0.753)
Special	-0.446(0.403)	-0.453(0.697)	-0.428(0.744)
Time	1.723(0.324)	2.383(0.403)	2.390(0.401)
Severity	-1.204(0.414)	-1.609(0.471)	-1.612(0.461)
-2ℓ	492.9	462.2	462.2
No. of parameters	10	12	12