Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 4, *pp.* 475–495

Explicit estimating equations for semiparametric generalized linear latent variable models

Yanyuan Ma and Marc G. Genton

Texas A&M University, College Station, USA

[Received January 2009. Final revision January 2010]

Summary. We study generalized linear latent variable models without requiring a distributional assumption of the latent variables. Using a geometric approach, we derive consistent semiparametric estimators. We demonstrate that these models have a property which is similar to that of a sufficient complete statistic, which enables us to simplify the estimating procedure and explicitly to formulate the semiparametric estimating equations. We further show that the explicit estimators have the usual root n consistency and asymptotic normality. We explain the computational implementation of our method and illustrate the numerical performance of the estimators in finite sample situations via extensive simulation studies. The advantage of our estimators over the existing likelihood approach is also shown via numerical comparison. We employ the method to analyse a real data example from economics.

Keywords: Complete statistic; Estimation efficiency; Latent variable; Maximum likelihood estimator; Quadrature points; Robustness; Score function; Sufficient statistic

1. Introduction

Models with latent variables are used extensively in social, economic and health sciences for analysing relationships between observed (manifest) and unobserved (latent) variables. For instance, in psychology, theoretical concepts such as intelligence are widely recognized to be important but cannot be measured directly. Instead, researchers have devised experiments that provide indirect measures of intelligence, e.g. intelligence quotient scores, vocabulary and reading speed. In economics, welfare and poverty cannot be measured directly; hence income, expenditure and various other indicators on households are used as substitutes. Similarly, in health studies, certain personal traits such as one's appetite or lifestyle are difficult to quantify or measure and only related answers on questionnaires allow researchers to infer on these traits. Factor analysis (Spearman, 1904) is probably the most well-known latent variable model, based on the assumption of multivariate normality for the distribution of the manifest and latent variables. It has been studied and extended by numerous researchers, such as Jöreskog (1967), Bartholomew (1980, 1984a, b), Moustaki (1996) and Sammel *et al.* (1997). Recently, Bartholomew and Knott (1999) and Moustaki and Knott (2000) proposed a generalized linear latent variable model (GLLVM) framework that allows the distribution of the manifest variables to belong to the exponential family, i.e. either continuous or discrete variables. It is closely related to structural equation models or random-effects (multilevel) models; see also Skrondal and Rabe-Hesketh (2004) for a comprehensive overview.

As a motivating example for GLLVMs, we consider the survey on Swiss consumption in

Address for correspondence: Marc G. Genton, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.
E-mail: genton@stat.tamu.edu

 1369–7412/10/72475

1990 that has been provided by the Swiss Federal Statistical Office. The data set contains both continuous and discrete (binary) manifest variable components. The continuous components are household expenses on food, clothing and leisure, whereas the binary components are the indicator of purchasing a dishwasher and a car in that year. It is perceptible that these elements are inherently related to the wealthiness of a household, which can be considered as the latent variable. Our interest is in quantifying how the household wealthiness affects these expenditures. Conditionally on this latent variable, we model the continuous components with a normal distribution and the binary components with a binomial distribution. However, we do not further model the household wealthiness to be normal, or any other distribution, because wealthiness is not observed. This example gives rise to a GLLVM without normal assumption on the latent variable distribution.

A GLLVM contains a p-dimensional latent variable $\tilde{Z}$ and a $(p+q)$-dimensional manifest variable X, where $p, q \geqslant 0$. In one of its most general forms, the model can be summarized as follows. We observe independent identically distributed data X_i, $i = 1, \ldots, n$, where $X_i = (X_i^{(1)}, \ldots, X_i^{(p+q)})^{\mathrm{T}}$ such that the probability density function of X, omitting the subindex i, is

$$f_X(x) = \int f_{X|\tilde{Z}}(x|\tilde{z})\, f_{\tilde{Z}}(\tilde{z})\, \mathrm{d}\mu(\tilde{z}) = \int \prod_{j=1}^{p+q} f_{X^{(j)}|\tilde{Z}}(x^{(j)}|\tilde{z})\, f_{\tilde{Z}}(\tilde{z})\, \mathrm{d}\mu(\tilde{z})$$
$$= \int \prod_{j=1}^{p+q} \exp\left\{ \frac{x^{(j)}\eta_j - b_j(\eta_j)}{\phi_j} + c_j(x^{(j)}, \phi_j) \right\} f_{\tilde{Z}}(\tilde{z})\, \mathrm{d}\mu(\tilde{z}), \tag{1}$$

where $\eta_j = \tilde{\alpha}_{\mathrm{con},j} + \tilde{\alpha}_j^{\mathrm{T}}\tilde{z}$ and $\tilde{\alpha}_j = (\tilde{\alpha}_{1j}, \ldots, \tilde{\alpha}_{pj})^{\mathrm{T}}$, and the subindex 'con' indicates the constant terms. Here $f_{X|\tilde{Z}}$ denotes the conditional probability density function of X on $\tilde{Z}$. The random vector $\tilde{Z} = (\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(p)})^{\mathrm{T}}$ is unobservable and hence represents latent random variables. Its marginal probability density function is denoted $f_{\tilde{Z}}$. We use μ to denote the σ-finite measure with respect to which $\tilde{Z}$ has probability density $f_{\tilde{Z}}(\cdot)$. Typically μ is the Lebesgue measure for continuous variables and the counting measure for discrete variables. The primary interest is usually in estimating $\tilde{\alpha}_{\mathrm{con},j}$ and $\tilde{\alpha}_j$, which form the main part of the parameter of interest. The scale parameter ϕ_j is of interest in the case of continuous observed components; hence we treat it as a part of the parameter of interest as well. The relationship between the manifest variable and the latent variables is completely captured by these parameters. The functions b_j and c_j have known forms. For example in the Swiss consumption data, for continuous $X^{(j)}$, $b_j(\eta_j) = \eta_j^2/2$ and

$$c_j(x^{(j)}, \phi_j) = -\frac{1}{2}\left\{ \frac{x^{(j)2}}{\phi_j} + \log(2\pi\phi_j) \right\},$$

with ϕ_j being the conditional variance of $X^{(j)}$, whereas, for binary $X^{(j)}$, $b(\eta_j) = \log\{1 + \exp(\eta_j)\}$ and $c(x^{(j)}, \phi_j) = 0$ with $\phi_j = 1$. Since the latent variable $\tilde{Z}$ is never observed, the density $f_{\tilde{Z}}(\cdot)$ is unknown and is considered as a nuisance parameter.

GLLVMs are designed as a flexible modelling approach; consequently they are rather complex models. This has led to a perfunctory use of the standard likelihood approach for inference in GLLVMs. In fact, it is customary to assume that $f_{\tilde{Z}}(\cdot)$ is multivariate normal, and to proceed with maximum likelihood to estimate $\tilde{\alpha}_{\mathrm{con},j}$s, $\tilde{\alpha}_j$s and ϕ_js via various numerical treatments in approximating the integrals in the log-likelihood. Those treatments include adaptive quadrature approaches (Skrondal and Rabe-Hesketh (2004), page 165), Laplace approximations (Huber *et al.*, 2004) and Monte Carlo methods (Yau and McGilchrist, 1996). Concerns regarding outliers or model deviations from the exponential family in the manifest variable distributions

have given rise to robust estimation procedures via indirect inference methods (Moustaki and Victoria-Feser, 2006).

In our view, a vulnerable assumption in the standard likelihood approach is the normal distributional assumption on the latent variable $\tilde{Z}$. Being latent, $\tilde{Z}$ is certainly unobservable; hence any model assumption on $f_{\tilde{Z}}(\cdot)$ is rather subjective and prone to errors. In fact, we believe that the normality assumption is taken only to facilitate the computation. Usually, the most straightforward approach to handle an unknown distribution is to estimate it, and this leads to the standard non-parametric maximum likelihood estimators; see for example, Skrondal and Rabe-Hesketh (2004), page 182, for its use in GLLVMs. Nevertheless, non-parametric estimation itself has difficulty both in terms of statistical property analysis and computational efficiency. It is even less attractive here since the target is a distribution of unavailable variables; hence it essentially solves a deconvolution-type problem. Consequently, the rate of the estimation is very slow and, to the best of our knowledge, no asymptotic theory for non-parametric maximum likelihood estimators is available for GLLVMs. Since $f_{\tilde{Z}}(\cdot)$ itself does not contain information on the relationship between the two sets of random variables and hence is not of main interest, our goal in this paper is to bypass estimating the distribution $f_{\tilde{Z}}(\cdot)$, and to estimate the parameters of interest and to make inference directly. For this, we treat the model as a semiparametric problem and approach the inference issue via constructing influence functions that do not rely on the correct specification of $f_{\tilde{Z}}(\cdot)$. The existence of such influence functions is rooted in the structure of the GLLVM, which is convex in terms of its nuisance parameter (Bickel *et al.* (1998), section 7.2). We obtain such influence functions through projecting a score function to a subspace which contains all influence functions. This projection will result in an estimator that is consistent even if the nuisance parameter $f_{\tilde{Z}}(\cdot)$ is misspecified or badly estimated. Moreover, taking advantage of an additional property of the GLLVM similar to that of a sufficient and complete statistic (Lindsay, 1982, 1983), we can obtain significant simplifications in constructing the estimators, deriving the theoretical aspects of the asymptotic properties, and carrying out the computation procedure. Furthermore, the class of estimators that we propose includes the optimal class which achieves the semiparametric efficiency bound.

The rest of the paper is organized as follows. We first address the identifiability issue in GLLVMs in Section 2. The identifiability issue is somewhat recognized in the application community, e.g. in psychology, but we have not been able to find a thorough and clear analysis of this important matter. In this section, we also lay out the unique parameter space that we shall work in. We then derive the class of semiparametric estimators in Section 3. The asymptotic properties are established in Section 4 and computational issues are addressed in Section 5. We implement the proposed estimator on the Swiss consumption data in Section 6. The satisfactory real data analysis results are followed by simulation studies in Section 7, where we demonstrate the finite sample properties of the estimators proposed, and their advantage over the classical likelihood approach. We conclude with a discussion in Section 8. Technical details are collected in Appendix A.

The program that was used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Identifiability of the model

A rather prominent identifiability issue calls for attention in GLLVMs. For convenience, we rewrite the model in matrix form as

$$f_X(x) = \int f_{X|\tilde{Z}}(x|\tilde{z})\, f_{\tilde{Z}}(\tilde{z})\, \mathrm{d}\mu(\tilde{z})$$
$$= \int \exp\left[(\tilde{\theta}_{\mathrm{con}}^{\mathrm{T}} + \tilde{z}^{\mathrm{T}}\tilde{A}_1)\Phi^{-1}x - \sum_{j=1}^{p+q} \phi_j^{-1} b_j \{(\tilde{\theta}_{\mathrm{con}}^{\mathrm{T}} + \tilde{z}^{\mathrm{T}}\tilde{A}_1)e_j\} + \sum_{j=1}^{p+q} c_j(x^{(j)}, \phi_j)\right] f_{\tilde{Z}}(\tilde{z})\, \mathrm{d}\mu(\tilde{z}),$$

where e_j $(j = 1, \ldots, p+q)$ is the jth unit $(p+q)$-vector, $\tilde{\theta}_{\mathrm{con}} = (\tilde{\alpha}_{\mathrm{con},1}, \ldots, \tilde{\alpha}_{\mathrm{con},p+q})^{\mathrm{T}}$, $\tilde{A}_1 = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{p+q})$ and $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_{p+q})$. Note that $\tilde{A}_1$ is a $p \times (p+q)$ matrix. For models that are parsimonious, the matrix $\tilde{A}_1$ is also of full rank (with rank p); see Appendix A.1.

For any $p \times p$ non-singular matrix G and any p-vector b, $\{\tilde{\theta}_{\mathrm{con}}, \tilde{A}_1, \Phi, f_{\tilde{Z}}(\tilde{z})\}$ and $\{\tilde{\theta}_{\mathrm{con}} - (G^{-1}\tilde{A}_1)^{\mathrm{T}}b, G^{-1}\tilde{A}_1, \Phi, f_Z(z)\}$ yield the same $f_X(x)$ for $Z = G\tilde{Z} + b$. This indicates that the GLLVM does not uniquely decide the parameters and does not distinguish $\tilde{Z}$ from its linear transformation, even if we have an explicit meaning for the latent variable $\tilde{Z}$. This unidentifiability cannot be eased by simply imposing distributional assumptions on $\tilde{Z}$. For example, the classical normality assumption on $f_{\tilde{Z}}(\cdot)$ has exactly the same unidentifiability. The identifiability issue is inherent in the structure of GLLVMs and is not merely a side effect from relaxing the distributional assumption on the latent variables. Considering that distributional assumptions on $\tilde{Z}$ are generally without foundation and prone to error, we propose to solve for a representative solution $(\theta_{\mathrm{con}}, A_1, \Phi)$ in the solution family and to provide the whole family $\{(\theta_{\mathrm{con}} + A_1^{\mathrm{T}}b, GA_1, \Phi), \forall G \text{ invertible}, \forall b\}$ as the solution set.

To select a simple representative solution, we require $A_1 = (U_0\, e_1\, U_1\, e_2\, U_2 \ldots e_p\, U_p) = (I_p A^*)P$. Here e_k $(k = 1, \ldots, p)$ is the kth unit p-vector, U_k $(k = 0, \ldots, p)$ are $p \times r_k$ $(r_k \geqslant 0)$ matrices with each column being the linear combination of $e_t, t \leqslant k$ (hence U_0 is a zero matrix; if $r_k = 0$, then U_k does not appear), $A^* = (U_0\, U_1 \ldots U_p)$ and P is the corresponding permutation matrix. In Appendix A.2, we show the existence and uniqueness of A_1. We further require $\theta_{\mathrm{con}}^{\mathrm{T}} = (0_p^{\mathrm{T}}\, \theta_r^{\mathrm{T}})P$, where P is the same matrix as in the expression for A_1, 0_p is the zero p-vector and θ_r is a q-vector. It is easy to see that setting $b = -(I_p\ 0)P\theta_{\mathrm{con}}$ will satisfy this requirement, and this is the only choice of b. Because of the structure that we impose on A_1, we can augment A_1 with A_2 so that $A = (A_1^{\mathrm{T}}\, A_2^{\mathrm{T}})^{\mathrm{T}}$ is invertible. Specifically, we set $A_2 = (0\ I_q)P$.

To identify A_1 is to identify A^* and P. Under our requirement for A_1 and θ_{con}, writing $P\Phi^{-1} = P\Phi^{-1}P^{-1}P = \tilde{\Phi}^{-1}P$, the conditional probability density function $f_{X|Z}$ can thus be alternatively written as

$$f_{X|Z}(x|z) = \exp(\{(0_p^{\mathrm{T}}\, \theta_r^{\mathrm{T}}) + z^{\mathrm{T}}(I_p\ A^*)\}\tilde{\Phi}^{-1}Px - \sum_{j=1}^{p+q} \phi_j^{-1} b_j [\{(0_p^{\mathrm{T}}\, \theta_r^{\mathrm{T}}) + z^{\mathrm{T}}(I_p\ A^*)\}Pe_j]$$
$$+ \sum_{j=1}^{p+q} c_j(x^{(j)}, \phi_j)),$$

where Z is the linear transformation of $\tilde{Z}$ that yields the current form of $(\theta_{\mathrm{con}}, A_1, \Phi)$. Note that Px is only a permutation of x; it does not change the relationship between X and Z, so we can always order X to have $P = I_{p+q}$. Thus, in what follows, we assume that $P = I_{p+q}$ and hence $\tilde{\Phi} = \Phi$, $A_1 = (I_p\ A^*)$ and $\theta_{\mathrm{con}}^{\mathrm{T}} = (0_p^{\mathrm{T}}\, \theta_r^{\mathrm{T}})$.

Within the solution set $\{(\theta_{\mathrm{con}} + A_1^{\mathrm{T}}b, GA_1, \Phi), \forall G \text{ invertible}, \forall b\}$, one is the true solution $(\tilde{\theta}_{\mathrm{con}}, \tilde{A}_1, \Phi)$ that has the untransformed $\tilde{Z}$ as the latent variable. Assume that $\tilde{\theta}_{\mathrm{con}} = \theta_{\mathrm{con}} + A_1^{\mathrm{T}}b_0$, $\tilde{A}_1 = G_0A_1$ and correspondingly $Z = G_0^{\mathrm{T}}\tilde{Z} + b_0$. Here and throughout the text, we use the subindex '0' to indicate the truth. The form of θ_{con} and A_1 allows us to write a simpler version of model (1) as

$$f_X(x) = \int \prod_{j=1}^{p} \exp\left\{ \frac{x^{(j)} z_j - b_j(z_j)}{\phi_j} + c_j(x^{(j)}, \phi_j) \right\} \times \prod_{j=p+1}^{p+q} \exp\left\{ \frac{x^{(j)} \eta_{j-p} - b_j(\eta_{j-p})}{\phi_j} + c_j(x^{(j)}, \phi_j) \right\} f_Z(z)\, \mathrm{d}\mu(z), \tag{2}$$

where $\eta_j = \theta_{rj} + (A_j^*)^{\mathrm{T}} z$ and A_j^* is the jth column of A^*, $j = 1, \ldots, q$. This model is identical to the GLLVM (1), and it has the benefit of being identifiable. Once we estimate $(\theta_{\mathrm{con}}, A_1, \Phi)$, then we can obtain the solution set $\{(\theta_{\mathrm{con}} + A_1^{\mathrm{T}} b, GA_1, \Phi)\}$. It is worth emphasizing that the reason we obtain a solution set instead of a single solution is the underlying unidentifiability. It is a consequence of the structure of the GLLVM, not a consequence of relaxing the normality assumption on the latent variables. If additional information allows us to determine G_0 and b_0, then we can recover $\tilde{\theta}_{\mathrm{con}}$ and $\tilde{A}_1$.

In GLLVMs, the major interest is in understanding the dependence of the manifest variable on the latent variables, which is captured in GA_1. Even without fixing a unique G, we can interpret the solution set as a relative description of such dependence. For example, assume that X_1 and X_2 are respectively expenditure on food and clothing, Z is household wealthiness and the estimated $\hat{A}_1 = (1, 2)$. It can then be interpreted as the clothing expenditure being related to household wealthiness twice as heavily as the food expenditure, regardless of which G we pick in forming a specific solution in the solution set. For a multivariate latent variable ($p > 1$), the relative dependence is with respect to the whole latent variable set. For example, if Z contains an additional component 'material desire', and we have $\hat{A}_1 = ((1, 2)^{\mathrm{T}}(2, 2)^{\mathrm{T}})$, then obviously $A_1 e_2 = DA_1 e_1$, where D is a diagonal matrix with 2 and 1 on the diagonal. This does not mean that expenditure on clothing is twice as heavily related to wealthiness as expenditure on food, and their dependence on material desire is the same. Considering an arbitrary invertible G, we obtain

$$GA_1 e_2 = GDG^{-1}(GA_1 e_1).$$

We can see that GDG^{-1} has the same eigenvalues as D, and the eigenvectors are orthogonal to each other. This means that we can always find two linear combinations of wealthiness and material desire that are orthogonal to each other and conclude that expenditure on clothing is twice as heavily dependent as expenditure on food on one component, and the dependence is the same on the other component. Thus, the relative dependence on the latent variable set, instead of on each latent variable, is what can be learned from GLLVMs. This is precisely caused by the fact that a GLLVM does not define a unique separation between the latent variables in its model structure.

We now briefly mention the possibilities in deciding G_0 and b_0; hence the identifiability and the recovery of $\tilde{\theta}_{\mathrm{con}}$ and $\tilde{A}_1$. As we have pointed out, assumptions other than the distribution family of the latent variables must be made to do so. A common assumption is about the mean and variance–covariance of the latent variable, $E(\tilde{Z}) = 0$ and $\mathrm{var}(\tilde{Z}) = \tilde{V}$. Using the generalized linear structure in model (2), we immediately know

$$\begin{aligned} E(X^{(j)}) &= E\{b'(Z^{(j)})\}, \\ \mathrm{var}(X^{(j)}) &= E\{b''(Z^{(j)})\}\phi_j + \mathrm{var}\{b'(Z^{(j)})\}, \\ \mathrm{cov}(X^{(j)}, X^{(k)}) &= \mathrm{cov}\{b'(Z^{(j)}), b'(Z^{(k)})\}, \end{aligned}$$

for all $1 \leqslant j, k \leqslant p$. Since the quantities on the left-hand side can be estimated from the observations, we can obtain $E(Z)$ and $\mathrm{var}(Z)$ from the above relationships by using a delta method.

Because $Z = G_0\tilde{Z} + b_0$, we obtain $b_0 = E(Z)$ and $\mathrm{var}(Z) = G_0\tilde{V}G_0^{\mathrm{T}}$. This, however, does not uniquely decide G_0. A whole family of G_0, caused by an arbitrary orthogonal matrix choice in solving G_0 from $\mathrm{var}(Z) = G_0\tilde{V}G_0^{\mathrm{T}}$, will qualify. To obtain a unique G_0, we must impose structures on G_0, e.g. that G_0 is upper triangular.

Although the mean 0 assumption on $\tilde{Z}$ is reasonable, a variance–covariance matrix on the unobservable $\tilde{Z}$ is not very convincing. Assuming a matrix structure on G_0 is even less appealing. These are truly awkward constraints just to hand-pick one solution from the solution family. It does not add to the understanding of the dependence between manifest and latent variables. Since $(\theta_{\mathrm{con}}, A_1 \text{ (or } A^*), \Phi)$ uniquely decides the whole solution set, and a solution set is what a GLLVM allows us to identify, we shall in what follows focus on $(\theta_{\mathrm{con}}, A_1 \text{ (or } A^*), \Phi)$ and the corresponding Z.

3. A class of semiparametric estimators

3.1. Semiparametric results

We use β to denote the collection of all the parameters of interest, i.e. $\beta = (\alpha_{\mathrm{con},p+1}, \ldots, \alpha_{\mathrm{con},p+q}, \alpha_{p+1}^{\mathrm{T}}, \ldots, \alpha_{p+q}^{\mathrm{T}}, \phi_1, \ldots, \phi_{p+q})^{\mathrm{T}}$. The vector β has length $m = pq + p + 2q$. Treating the density $f_Z(\cdot)$ as an unknown nuisance parameter, thus using a semiparametric approach, we derive the nuisance tangent space Λ and the nuisance tangent space orthogonal complement $\Lambda^{\perp}$ to be

$$\Lambda = [E\{h(Z)|X\} : E\{h(Z)\} = 0, E\{h^{\mathrm{T}}(Z)\,h(Z)\} < \infty],$$
$$\Lambda^{\perp} = [g(X) : E\{g(X)|Z\} = 0, E\{g^{\mathrm{T}}(X)\,g(X)\} < \infty],$$

where both h and g are dimension m vector functions. The semiparametric approach is explained nicely in Tsiatis (2006). We describe the concepts of Λ and $\Lambda^{\perp}$ and the details of the derivation in Appendix A.3. The score vector, which is defined as $\partial \log\{f_X(x)\}/\partial\beta$, can be easily verified to be $S_\beta(X) = E\{S_\beta(X|Z)|X\}$, where

$$S_\beta(X|Z) = \frac{\partial \log\{f_{X|Z}(X|Z;\beta)\}}{\partial\beta}.$$

Projecting the score vector $S_\beta(x)$ onto the space $\Lambda^{\perp}$, which is denoted $\Pi(S_\beta|\Lambda^{\perp})$, we obtain the efficient score vector function

$$S_{\mathrm{eff}}(X) \equiv \Pi(S_\beta|\Lambda^{\perp}) = S_\beta(X) - E\{a(Z)|X\},$$

where $E\{a(Z)\} = 0$ and

$$E[S_\beta(X) - E\{a(Z)|X\}|Z] = 0.$$

The validity of the expression for S_{eff} can be easily verified by noting that $S_{\mathrm{eff}} \in \Lambda^{\perp}$ and $S_\beta - S_{\mathrm{eff}} = E\{a(Z)|X\} \in \Lambda$. The estimator could be obtained through first solving for $a(Z)$ from the above equality, and then calculating S_{eff} to form an estimating equation $\Sigma_{i=1}^{n} S_{\mathrm{eff}}(X_i;\beta) = 0$. However, to perform this calculation, we need to have the distribution $f_Z(\cdot)$, which is unknown. Here, we propose to conjecture a possibly misspecified $f_Z(\cdot)$ and go through all the above calculations. We claim that the construction under a misspecified $f_Z(\cdot)$, which is denoted $f_Z^*(\cdot)$, will still yield a consistent estimator. To make the construction of the estimation procedure under $f_Z^*(\cdot)$ explicit, we use an asterisk to denote all the calculations affected. Specifically, we calculate S_{eff}^* through $S_{\mathrm{eff}}^*(X) = S_\beta^*(X) - E^*\{a(Z)|X\}$, where $S_\beta^*(X) = E^*\{S_\beta(X|Z)|X\}$ and

$$E\{S_\beta^*(X)|Z\} = E[E^*\{a(Z)|X\}|Z], \tag{3}$$

and obtain the estimating equation

$$\sum_{i=1}^{n} S^*_{\text{eff}}(X_i;\beta)=0.$$

Because S^*_{eff} still has mean 0 owing to

$$E\{S^*_{\text{eff}}(X_i;\beta_0)|Z\}=E\{S^*_{\beta}(X)|Z\}-E[E^*\{a(Z)|X\}|Z]=0,$$

hence the estimator is indeed consistent even if the model $f^*_Z(\cdot)$ is incorrect. An intuitive understanding of the consistency under $f^*_Z(\cdot)$ is that, when solving for $a(Z)$ in equation (3), we ensured that $E(S^*_{\text{eff}}|Z)=0$, where the conditional expectation itself does not rely on f_Z, and it subsequently guarantees that $E(S^*_{\text{eff}})=0$; hence the consistency of the estimator.

Solving for $a(Z)$ from equation (3) is certainly not an easy task. A similar equation has been employed in Tsiatis and Ma (2004) and has been noted to be an ill-posed problem and to be numerically unstable sometimes. Fortunately, the form of GLLVM facilitates a further simplification that avoids solving for equation (3), which we shall explore next.

3.2. Explicit form estimators

Denoting $W=A_1\Phi^{-1}X$ and $Y=A_2\Phi^{-1}X$, where A_1 and A_2 are defined in Section 2, we have

$$\begin{aligned} f_{W,Y|Z}(w,y|z) &= f_{X|Z}\left\{\Phi A^{-1}\begin{pmatrix} w \\ y \end{pmatrix}\middle| z\right\} J^{-1} \\ &= \exp\left[\theta_{\text{con}}^{\mathrm{T}} A^{-1}\begin{pmatrix} w \\ y \end{pmatrix} + z^{\mathrm{T}} w - \sum_{j=1}^{p+q} \phi_j^{-1} b_j(\alpha_{\text{con},j}+\alpha_j^{\mathrm{T}} z) \right. \\ &\quad \left. + \sum_{j=1}^{p+q} c_j\left\{e_j^{\mathrm{T}}\Phi A^{-1}\begin{pmatrix} w \\ y \end{pmatrix}, \phi_j\right\}\right] J^{-1}, \end{aligned}$$

where $J=|\det(A\Phi^{-1})|$. We show next that the random variable W has a similar property to that of a sufficient and complete statistic.

Theorem 1. In GLLVMs, the random variables W, Y and Z satisfy the relationships

$$\begin{aligned} f_{Y|W,Z}(y|w,z) &= f_{Y|W}(y|w), \\ f_{Z|W,Y}(z|w,y) &= f_{Z|W}(z|w). \end{aligned} \tag{4}$$

In addition, for any function $h(\cdot)$, we have

$$E\{h(W)|z\}=0 \Rightarrow h(W)=0. \tag{5}$$

Proof. We first establish the sufficiency property of W in expression (4). Direct calculation shows that

$$\begin{aligned} f_{Y|W,Z}(y|w,z) &= \frac{f_{W,Y|Z}(w,y|z)}{\int f_{W,Y|Z}(w,y|z)\,\mathrm{d}\mu(y)} \\ &= \frac{\exp\left[\theta_{\text{con}}^{\mathrm{T}} A^{-1}\begin{pmatrix} w \\ y \end{pmatrix} + \sum_{j=1}^{p+q} c_j\left\{e_j^{\mathrm{T}}\Phi A^{-1}\begin{pmatrix} w \\ y \end{pmatrix}, \phi_j\right\}\right]}{\int \exp\left[\theta_{\text{con}}^{\mathrm{T}} A^{-1}\begin{pmatrix} w \\ y \end{pmatrix} + \sum_{j=1}^{p+q} c_j\left\{e_j^{\mathrm{T}}\Phi A^{-1}\begin{pmatrix} w \\ y \end{pmatrix}, \phi_j\right\}\right]\mathrm{d}\mu(y)}. \end{aligned}$$

The last term does not involve Z; hence $f_{Y|W,Z}(y|w,z) = f_{Y|W}(y|w)$. This means that Y is independent of Z conditional on W; hence we also have $f_{Z|W,Y}(z|w,y) = f_{Z|W}(z|w)$.

We now show the completeness property of W in result (5). Assume that a certain $h(w)$ satisfies $E\{h(W)|z\} = 0$. This expands to

$$\begin{aligned} 0 &= \int h(w)\, f_{W,Y|Z}(w,y|z)\, \mathrm{d}\mu(y)\, \mathrm{d}\mu(w) \\ &= J^{-1} \int \int \exp\left[\theta_{\mathrm{con}}^{\mathrm{T}} A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} + \sum_{j=1}^{p+q} c_j \left\{ e_j^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_j \right\}\right] \mathrm{d}\mu(y) \\ &\quad \times h(w) \exp(z^{\mathrm{T}} w)\, \mathrm{d}\mu(w) \exp\left\{ - \sum_{j=1}^{p+q} \phi_j^{-1} b_j(\alpha_{jc} + \alpha_j^{\mathrm{T}} z) \right\}. \end{aligned}$$

Denote

$$g(w) = \int \exp\left[\theta_{\mathrm{con}}^{\mathrm{T}} A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} + \sum_{j=1}^{p+q} c_j \left\{ e_j^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_j \right\}\right] \mathrm{d}\mu(y),$$

a positive function of w; then we have

$$0 = \int h(w)\, g(w) \exp(z^{\mathrm{T}} w)\, \mathrm{d}\mu(w).$$

Hence, $h(w)\, g(w) = 0$. Because $g(w) > 0$ for any w, we obtain $h(w) = 0$. □

The relationships in theorem 1 turn out to be of vital importance for computational simplification. Because

$$E\{S_\beta^*(X)|Z\} = E[E\{S_\beta^*(X)|W,Z\}|Z]$$

and

$$E[E^*\{a(Z)|X\}|Z] = E[E^*\{a(Z)|W,Y\}|Z],$$

we obtain

$$E[E\{S_\beta^*(X)|W,Z\}|Z] = E[E^*\{a(Z)|W,Y\}|Z].$$

Using expression (4), we have

$$E\{S_\beta^*(X)|W,Z\} = E\left[S_\beta^* \left\{ \Phi A^{-1} \begin{pmatrix} W \\ Y \end{pmatrix} \right\} \Big| W, Z \right] = E\{S_\beta^*(X)|W\},$$
$$E^*\{a(Z)|W,Y\} = E^*\{a(Z)|W\}.$$

Thus, we have

$$E[E\{S_\beta^*(X)|W\}|Z] = E[E^*\{a(Z)|W\}|Z],$$

or equivalently

$$E[E\{S_\beta^*(X)|W\} - E^*\{a(Z)|W\}|Z] = 0.$$

From result (5), we have

$$E\{S_\beta^*(X)|W\} - E^*\{a(Z)|W\} = 0;$$

hence, $E^*\{a(Z)|X\} = E^*\{a(Z)|W,Y\} = E^*\{a(Z)|W\} = E\{S_\beta^*(X)|W\}$. Recalling the form of S_{eff}, we have in fact obtained

$$S^*_{\text{eff}}(X) = S^*_\beta(X) - E^*\{a(Z)|X\}$$
$$= S^*_\beta(X) - E\{S^*_\beta(X)|W\},$$

which is an explicit form of S^*_{eff}. Thus, the procedure of solving for a is bypassed. We proceed to form the estimating equation by using the expression for S^*_{eff} that is given above.

Specific calculation of the partial derivatives of $\log\{f_{X|Z}(x|z;\beta)\}$ with respect to $\alpha_{\text{con},k}$, α_k and ϕ_k yields

$$S_{\alpha_{\text{con},k}}(X|Z) = \phi_k^{-1} e_k^{\mathrm{T}} X - \phi_k^{-1} b_k'(\alpha_{\text{con},k} + \alpha_k^{\mathrm{T}} Z),$$

$$S_{\alpha_{lk}}(X|Z) = \phi_k^{-1} Z^{\mathrm{T}} O_{lk} X - \phi_k^{-1} b_k'(\alpha_{\text{con},k} + \alpha_k^{\mathrm{T}} Z) Z_l,$$

$$S_{\phi_k}(X|Z) = -\phi_k^{-2}\alpha_{\text{con},k} e_k^{\mathrm{T}} X - \phi_k^{-2} Z^{\mathrm{T}}(0\ \alpha_k\ 0)X + \phi_k^{-2} b_k(\alpha_{\text{con},k} + \alpha_k^{\mathrm{T}} Z) + c'_{k2}(X^{(k)}, \phi_k),$$

where $l = 1, \ldots, p$, $k = p+1, \ldots, p+q$ for the first two equations and $k = 1, \ldots, p+q$ for the last equation. Here O_{lk} denotes a $p \times (p+q)$ matrix where only the lkth element is 1; all others are 0. Making use of the relationship $S^*_{\text{eff}}(X) = S^*_\beta(X) - E^*\{S_\beta(X)|W\} = E^*\{S_\beta(X|Z)|W, Y\} - E^*\{S_\beta(X|Z)|W\}$, we further obtain

$$S^*_{\text{eff}}|_{\alpha_{\text{con},k}} = \phi_k^{-1} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E(Y|w) \end{pmatrix},$$

$$S^*_{\text{eff}}|_{\alpha_{lk}} = \phi_k^{-1} E^*(Z|w)^{\mathrm{T}} O_{lk} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E(Y|w) \end{pmatrix},$$

$$S^*_{\text{eff}}|_{\phi_k} = -\phi_k^{-2}\alpha_{\text{con},k} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E(Y|w) \end{pmatrix} - \phi_k^{-2} E^*(Z|w)^{\mathrm{T}}(0\,\alpha_k\,0)\Phi A^{-1} \begin{pmatrix} 0 \\ y - E(Y|w) \end{pmatrix}$$
$$+ c'_{k2}\left(e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_k \right) - E\left\{ c'_{k2}\left(e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_k \right) \Big| w \right\}.$$

Here, the only component involving an asterisk, and hence depending on the unknown distribution $f_Z(\cdot)$, is $E^*(Z|w)$. The detail of the derivation of these projected score functions is in Appendix A.4. A further observation is that, to obtain a consistent estimator, we do not really need to posit a model for $f_Z(\cdot)$ and then to calculate $E^*(Z|w)$. Instead, we can use an arbitrary p-dimension function $h(w)$ to replace $E^*(Z|w)$. As long as the system of equations does not degenerate, we are still guaranteed to obtain a consistent estimator. This is because the underlying mechanism that drives the consistency is $E(S^*_{\text{eff}}|w) = 0$ when calculated at the true parameter values. This property continues to hold when we replace $E^*(Z|w)$ by an arbitrary function of w.

Taking into account the special form of

$$A = \begin{pmatrix} I_p & A^* \\ 0 & I_q \end{pmatrix},$$
$$A^{-1} = \begin{pmatrix} I_p & -A^* \\ 0 & I_q \end{pmatrix},$$

we obtain the system of estimating equations that is equivalent to the system that is formed by S^*_{eff} while replacing $E^*(Z|w)$ by $h(w)$:

$$\sum_{i=1}^{n} \{y_i - E(Y|w_i)\} = 0, \tag{6}$$

$$\sum_{i=1}^{n} [h(w_i) \otimes \{y_i - E(Y|w_i)\}] = 0, \tag{7}$$

$$\sum_{i=1}^{n}\left[c'_{k2}\left(e_k^{\mathrm{T}}\Phi A^{-1}\begin{pmatrix}w_i\\ y_i\end{pmatrix},\phi_k\right)-E\left\{c'_{k2}\left(e_k^{\mathrm{T}}\Phi A^{-1}\begin{pmatrix}w_i\\ Y\end{pmatrix},\phi_k\right)\Big|w_i\right\}\right]=0, \tag{8}$$

where '$\otimes$' is the Kronecker product, and $k=1,\ldots,p+q$ for the last equation. The set of estimating equations (6)–(8) forms a class of root n consistent estimators that are indexed by the function $h(w)$, and this is the class of estimators that we propose for the GLLVM. When we use a specific $h(w)=E(Z|w)$, the corresponding estimator is efficient, in that it has the smallest possible estimation variance, i.e. it reaches the semiparametric efficiency bound. This is because, when $h(w)=E(Z|w)$, the estimating equation is equivalent to the equation that is formed by the true efficient score functions; hence it inherits the properties of the efficient score estimator. In particular, it automatically achieves the optimal efficiency. In the literature, similar properties have been observed. For example, in generalized estimating equations, the true variance–covariance structure will yield efficiency, whereas a misspecification still warranties consistency. The simplification and the resulting class of explicit estimators for the GLLVM in Section 3.2 also have a similar flavour to those in mixture models (Lindsay, 1982, 1983), measurement error models (Stefanski and Carroll, 1987) and generalized linear mixed models (Sartori and Severini, 2004), although the derivation is completely different.

4. Asymptotic properties

Because the construction of the estimator is through projection of the score vector onto the nuisance tangent space orthogonal complement $\Lambda^{\perp}$, the resulting estimator is guaranteed to be efficient if we had used a correct $f_Z(\cdot)$. In addition, as a function of the nuisance parameter $f_Z(\cdot)$, the likelihood of a single observation given in model (1) satisfies the convexity property (Bickel *et al.*, 1998; Newey, 1990)

$$f_X\{x;\lambda f_{Z1}+(1-\lambda)f_{Z2}\}=\lambda f_X(x;f_{Z1})+(1-\lambda)f_X(x;f_{Z2});$$

hence the resulting estimator remains consistent if the projection is performed under a postulated $f_Z^*(\cdot)$ as well. We summarize these observations in a somewhat stronger sense in theorems 2 and 3.

Theorem 2. Denote equations (6)–(8) as $\Sigma_{i=1}^{n}\Psi(X_i,\beta)=0$. Assume that $\hat{\beta}$ solves the equations. Then $n^{1/2}(\hat{\beta}-\beta_0)\to N(0,B^{-1}VB^{-\mathrm{T}})$ in distribution when $n\to\infty$. Here, $V=\mathrm{var}\{\Psi(X_i;\beta_0)\}$ and $B=E\{\partial\Psi(X_i;\beta_0)/\partial\beta^{\mathrm{T}}\}$.

Proof. Obviously, the explicit form of equations (6)–(8) ensures that $E\{\Psi(X_i;\beta_0)\}=0$. Expanding around β_0, we obtain

$$\begin{aligned}0&=\sum_{i=1}^{n}\Psi(X_i;\hat{\beta})\\&=\sum_{i=1}^{n}\Psi(X_i;\beta_0)+\sum_{i=1}^{n}\frac{\partial\Psi(X_i;\beta^*)}{\partial\beta^{\mathrm{T}}}(\hat{\beta}-\beta_0),\end{aligned}$$

where β^* lies on the line that connects β_0 and $\hat{\beta}$. Therefore we have

$$\begin{aligned}n^{1/2}(\hat{\beta}-\beta_0)&=-\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial\Psi(X_i;\beta^*)}{\partial\beta^{\mathrm{T}}}\right\}^{-1}n^{-1/2}\sum_{i=1}^{n}\Psi(X_i;\beta_0)\\&=-E\left\{\frac{\partial\Psi(X_i;\beta_0)}{\partial\beta^{\mathrm{T}}}\right\}^{-1}n^{-1/2}\sum_{i=1}^{n}\Psi(X_i;\beta_0)+o_p(1);\end{aligned}$$

hence the result follows. □

Theorem 3. If the function $h(w)$ in the estimating equation satisfies $h(w) = E^*(Z|w;\hat{\gamma})$, where $\hat{\gamma}$ is a root n consistent estimator of γ_0 and $f_Z^*(\cdot;\gamma_0) = f_Z(\cdot)$, then $\hat{\beta}$ is semiparametric efficient and $\mathrm{var}(\hat{\beta}) = \mathrm{var}(S_{\mathrm{eff}})^{-1}$.

Proof. The proof is similar to the proof of theorem 2; noting that $\Psi(X_i;\beta_0,\gamma_0) = S_{\mathrm{eff}}(X_i)$, we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Psi(X_i;\hat{\beta},\hat{\gamma}) \\ &= \sum_{i=1}^{n} S_{\mathrm{eff}}(X_i) + \sum_{i=1}^{n} \frac{\partial \Psi(X_i;\beta^*,\gamma^*)}{\partial \beta^{\mathrm{T}}}(\hat{\beta}-\beta_0) + \sum_{i=1}^{n} \frac{\partial \Psi(X_i;\beta^*,\gamma^*)}{\partial \gamma^{\mathrm{T}}}(\hat{\gamma}-\gamma_0) \end{aligned}$$

where (β^*,γ^*) lies on the line that connects (β_0,γ_0) and $(\hat{\beta},\hat{\gamma})$. Note that

$$n^{-1}\sum_{i=1}^{n} \frac{\partial \Psi(X_i;\beta^*,\gamma^*)}{\partial \gamma^{\mathrm{T}}} = E\left\{ \frac{\partial \Psi(X_i;\beta_0,\gamma_0)}{\partial \gamma^{\mathrm{T}}} \right\} + o_p(1)$$

and

$$E\left\{ \frac{\partial \Psi(X_i;\beta_0,\gamma_0)}{\partial \gamma^{\mathrm{T}}} \right\} = -E\left[S_{\mathrm{eff}}(X_i) \frac{\partial \log\{f_X(X_i;\beta_0,\gamma_0)\}}{\partial \gamma^{\mathrm{T}}} \right] = 0$$

because of the orthogonality between the projected score vector S_{eff} and any element in the nuisance tangent space Λ. In addition, we also have $n^{1/2}(\hat{\gamma}-\gamma_0) = O_p(1)$; therefore we have

$$\begin{aligned} n^{1/2}(\hat{\beta}-\beta_0) &= -\left\{ n^{-1}\sum_{i=1}^{n} \frac{\partial \Psi(X_i;\beta^*,\gamma^*)}{\partial \beta^{\mathrm{T}}} \right\}^{-1} n^{-1/2}\sum_{i=1}^{n} S_{\mathrm{eff}}(X_i) + o_p(1) \\ &= -\left\{ E\frac{\partial S_{\mathrm{eff}}(X_i)}{\partial \beta^{\mathrm{T}}} \right\}^{-1} n^{-1/2}\sum_{i=1}^{n} S_{\mathrm{eff}}(X_i) + o_p(1). \end{aligned}$$

The result follows by noting that

$$-E\left(\frac{\partial S_{\mathrm{eff}}}{\partial \beta^{\mathrm{T}}} \right) = E\left\{ S_{\mathrm{eff}} \frac{\partial f_X(X)}{\partial \beta^{\mathrm{T}}} \right\} = E(S_{\mathrm{eff}} S_{\beta}^{\mathrm{T}}) = \mathrm{var}(S_{\mathrm{eff}}). \qquad \square$$

Although theorem 3 lays out a theoretical condition for the efficiency, it is obviously not easy to achieve in practice because we usually do not have good knowledge of $f_Z(\cdot)$ to postulate a reasonable model for it and to carry out the parameter estimation that is involved in such a model. For this reason, in practice, we recommend the use of simple functions $h(w)$ to facilitate an easy computation and to be content with the root n consistency as guaranteed by theorem 2.

It is worth pointing out an implication of theorem 2. As long as a correct model is chosen for $f_Z(\cdot)$, the estimation of any parameters that are involved in such a model does not cause any loss of efficiency, provided that they are estimated at root n rate. In fact, much more is true. In the case when a model is chosen for $f_Z(\cdot)$, regardless of whether the model is correct or misspecified, the estimation of the parameters that are involved in the chosen model does not have any effect on the first-order property of the estimation of β. Thus, in theory, we could opt for a rather complicated model for $f_Z(\cdot)$ to minimize the effect of model misspecification at no asymptotic cost.

5. Computational treatment

The estimating equations (6)–(8) can be solved by using a standard Newton–Raphson algorithm.

The only computational issue worth pointing out is the calculation of $E(Y|w)$ and $E(c'_{k2}|w)$.

The computation of $E(Y|w)$ is relatively straightforward. For the models where Y is discrete, we typically can use a sum or a truncated sum to obtain an approximation. When Y is continuous, $E(Y|w)$ can be calculated with one's favourite numerical integration method, e.g. quadrature methods, Laplace approximations or Monte Carlo methods. Because Laplace approximations have recently been advocated by Huber *et al.* (2004), we provide a detailed description of the method in this context.

If we denote

$$t = \theta_{\mathrm{con}}^{\mathrm{T}} A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} + \sum_{j=1}^{p+q} c_j \left\{ e_j^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_j \right\},$$

then $f_{Y|W}(y|w) = \exp(t)/\int \exp(t)\,\mathrm{d}\mu(y)$. Suppose that the maximum of $t(y)$ is obtained at y_0; then

$$\begin{aligned} E(Y|w) &= \frac{\int \exp\{t(y)\} y \,\mathrm{d}\mu(y)}{\int \exp\{t(y)\}\,\mathrm{d}\mu(y)} \\ &\approx \frac{\int \exp\{t(y_0)\} \exp\{-(y-y_0)^{\mathrm{T}} t''(y_0)(y-y_0)/2\}(y_0 + y - y_0)\,\mathrm{d}\mu(y)}{\int \exp\{t(y_0)\} \exp\{-(y-y_0)^{\mathrm{T}} t''(y_0)(y-y_0)/2\}\,\mathrm{d}\mu(y)} = y_0. \end{aligned}$$

We thus obtain y_{i0} for $i = 1, \ldots, n$ from maximizing $t(y)$ with respect to y at each observation. In Appendix A.5, we establish that one can typically solve the maximization problem through solving

$$(-A^{*\mathrm{T}} I_q) \left\{ \theta_{\mathrm{con}} + \sum_{j=1}^{p+q} c'_{j1} \left(e_j^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w_i \\ y_{i0} \end{pmatrix}, \phi_j \right) \Phi e_j \right\} = 0, \qquad i = 1, \ldots, n.$$

The calculation of $E(c'_{k2}|w)$ mostly follows the same pattern as for $E(Y|w)$ and depends strongly on the function c'_{k2}. For some model settings, e.g. Poisson or binomial models, no unknown ϕ is involved, so equation (8) does not appear at all. For the normal model, it is a matter of calculating $E(Y^2|w)$, which can be handled similarly to $E(Y|w)$. For gamma models with unknown shape parameter or for inverse Gaussian models, it involves calculating $E\{\log(Y)|w\}$ or $E(1/Y|w)$. These are not suitable with Laplace approximations, so we propose to use other numerical methods, e.g. Monte Carlo methods or quadrature methods. Various other models generate different forms for the function c'_{k2} and one will need to analyse them case by case and subsequently to approximate the integration by using an appropriate numerical approach.

6. Swiss consumption data

We return to the Swiss consumption data that were mentioned in Section 1. This quite familiar data set in the GLLVM literature (Moustaki and Victoria-Feser, 2006) contains 1963 observations. In traditional GLLVM modelling of these data, the distribution of wealthiness is assumed to be standard normal. However, similarly to the distribution of salary, the normal distribution assumption on wealthiness is rather questionable, especially for Switzerland, where there is a visible proportion of extremely wealthy individuals. On the basis of such concerns, we leave the wealthiness distribution unspecified and implement the semiparametric method proposed. For

Table 1. GLLVM analysis on the Swiss consumption data

	Results from semiparametric method		*Results from MLE method*	
	Estimate	*Standard error*	*Estimate*	*Standard error*
$\alpha_{\text{con,food}}$	0	—	0	—
$\alpha_{\text{con,clothing}}$	1.322	0.039	1.322	0.068
$\alpha_{\text{con,leisure}}$	1.443	0.039	1.443	0.061
$\alpha_{\text{con,dishwasher}}$	−0.143	0.106	−0.287	0.080
$\boldsymbol{\alpha_{\text{c,car}}}$	**2.589**	**0.274**	**1.826**	**0.129**
α_{food}	1	—	1	—
$\boldsymbol{\alpha_{\text{clothing}}}$	**1.597**	**0.593**	**3.055**	**0.690**
$\boldsymbol{\alpha_{\text{leisure}}}$	**1.400**	**0.509**	**2.505**	**0.463**
$\alpha_{\text{dishwasher}}$	4.448	0.597	3.326	0.434
$\boldsymbol{\alpha_{\text{car}}}$	**6.983**	**1.438**	**3.845**	**0.640**
ϕ_{food}	0.853	0.064	0.938	0.042
ϕ_{clothing}	0.625	0.132	0.423	0.086
ϕ_{leisure}	0.712	0.140	0.612	0.070

comparison, we also implemented the classical normal-based maximum likelihood estimator (MLE) on this data set. The resulting estimates and standard errors are summarized in Table 1, with coefficients in bold for significant differences between the two estimation methods at the 5% level. (We calculate the 95% confidence intervals for both estimators by using theorem 2 and standard results of the MLE. If neither of the two intervals covers the other estimator, then we call the two estimators significantly different at the 5% level.) As we can see, compared with the MLE method, the semiparametric estimation of the baseline α_{con} is quite similar to that of the MLE for the normal variables, whereas it is different for the binomial variables. This is an indication that the normal distribution assumption on the distribution of wealthiness in this sample is likely to be false. We further performed a goodness-of-fit test on the normality of the latent variable by inspecting the continuous manifest variable components. If the normality holds, then the food, clothing and leisure expenditures will have a trivariate normal distribution; hence the squared norm of their Mahalanobis transform would have a χ^2_3-distribution. The Kolmogorov–Smirnov test rejects the χ^2_3-distribution at the 0.01 level; hence the normality assumption is highly unlikely to hold. The semiparametric method also estimates a relatively larger influence of wealthiness on the expenditure on dishwashers and cars (especially cars), whereas the influence on more frequent routine expenses such as clothing and leisure is estimated to be relatively smaller. This agrees better with our general observation that, in the modern era for developed countries, food and clothing are no longer a major indicator of wealthiness. The effect of household wealthiness is mainly reflected in the more expensive expenditures.

In this example, we have opted for a simple form $h(w) = w$ to facilitate the implementation. The validity of this choice will be demonstrated through a satisfactory result from the first set of simulations, which is designed to mimic this data example. The guideline in selecting h that we would like to recommend, in the absence of knowledge of $f_Z(\cdot)$, is to choose a simple h (such as linear), and gradually to increase its complexity (such as higher order polynomials), until the estimation variance is sufficiently small. If for various reasons a feasible conjecture or rough estimation of $f_Z(\cdot)$ is available, then we should use $E^*(Z|w)$, calculated under the

corresponding approximated $f_Z(\cdot)$, as h. More involved approaches are possible at the cost of more computation; see Lindsay (1985). The generalized method of moments can also be implemented in the case of an excessive number of candidates for h, but these are certainly out of the scope of the current paper.

We would like to point out that, in the example that we present here, we have modelled the effect of the latent variable to be linear, which is naturally the first and simplest modelling approach. The nature of GLLVMs is that the latent variable is unobservable and often even only exists in concept. Hence, such linearity is very difficult to verify. Extending the linearity to non-linear or even non-parametric relationships is interesting and is a topic of on-going research. A statistically sound procedure to test and validate the GLLVM is important and can be very challenging.

7. Simulation study

The difference in the parameter estimates in the example in Section 6 indicates that the normal distribution assumption on the latent variable might be false. It also reflects that a misspecification of the latent variable distribution could have an effect on the parameter estimation. To verify that the difference is caused by the normality assumption of the MLE, we compare the results of the estimator proposed and the MLE on simulated data sets when the latent variable distribution is truly normal. To investigate the effect of the misspecification of the latent variable distribution further, we also design the simulation study with different departures from normality to reveal its consequence for both the estimator proposed and the MLE. The simulation will also study the finite sample properties of the estimators proposed. We use sample size $n = 500$ and 1000 simulation replicates throughout the simulations. We consider univariate and bivariate latent variables.

7.1. Univariate latent variables

We generate the latent variable Z from four different distributions. They are

(a) a normal distribution with mean -1 and variance 1.5,
(b) a gamma distribution with shape parameter 1.4 and scale parameter 1,
(c) a mixture of normal distributions with mean 3 and variance 1 for 90% of the data and mean -3 and variance 0.25 for 10% of the data, and
(d) a Student t-distribution with mean -1 and 3 degrees of freedom.

The first case is provided as a benchmark, whereas the remaining three distributions provide different scenarios where the latent variable distribution is skewed, bimodal and heavy tailed. The manifest variable X consists of three normal components and two binomial components, given the latent variable Z.

We carry out the proposed semiparametric estimation in all the four situations, choosing $h(w) = w$. For comparison, we also compute the MLE under the normality assumption of the latent variable. As far as the semiparametric model is concerned, where the latent variable distribution is left unspecified, a more suitable name for such an estimator is pseudo-MLE. However, here we follow the GLLVM literature and term it the MLE throughout the paper. In implementing the classical normal-based MLE, we decide *a priori* the mean and variance of the latent variable. In our experiment, we simply use the true underlying mean and variance, which represent the optimal condition for the classical MLE and should consequently perform better than in reality, when these quantities are approximated or estimated from external knowledge or extra data. To compare the simulation results from the two estimators, we align the MLE

results into the same format as the proposed estimator result, i.e. we obtain G and b from the first p constraints of

$$\hat{\theta}_{\text{con}}^{\text{T}} - bG^{-1}\hat{A}_1 = (0_p^{\text{T}}\ \theta_r^{\text{T}}),$$
$$G^{-1}\hat{A}_1 = (I_p\ A^*),$$

and then calculate (θ_r, A^*) by using the corresponding b and G^{-1}.

The results of the simulations are given in Tables 2 and 3. As we expected, both estimators are consistent when the true distribution of the latent variable is indeed normal. We had expected the MLE to be more efficient than the estimator proposed, which makes much weaker assumptions. But the difference is nearly negligible, even under this favourable condition for the MLE where the mean and variance of the latent variable are assumed to be known. When the underlying distribution deviates from normality, the MLE is biased. This bias exhibits mainly in the parameters corresponding to the binomial manifest variable components. The insensitivity of the normal manifest random variable parameters to the latent variable distribution

Table 2. Simulation study†

β	β_0	*Results from semiparametric method*				*Results from MLE method*			
		$\hat{\beta}$	$\text{var}(\hat{\beta})$	$\widehat{\text{var}}(\hat{\beta})$	*95%cov (%)*	$\hat{\beta}$	$\text{var}(\hat{\beta})$	$\widehat{\text{var}}(\hat{\beta})$	*95%cov (%)*
Normal distribution									
θ_{r1}	−0.3	−0.314	0.063	0.071	96.5	−0.314	0.063	0.071	96.5
θ_{r2}	−0.2	−0.218	0.082	0.091	94.6	−0.218	0.082	0.091	94.6
θ_{r3}	−1.8	−1.857	0.179	0.181	95.0	−1.855	0.177	0.179	95.3
θ_{r4}	−3.2	−3.281	0.312	0.340	96.1	−3.276	0.299	0.325	95.9
A_{11}^*	1.4	1.409	0.026	0.029	96.0	1.409	0.026	0.029	95.9
A_{12}^*	1.6	1.610	0.033	0.037	95.3	1.610	0.033	0.037	95.3
A_{13}^*	1.4	1.440	0.078	0.078	95.3	1.439	0.077	0.077	95.8
A_{14}^*	2	2.047	0.130	0.137	95.9	2.044	0.123	0.131	96.6
ϕ_1	1	0.993	0.005	0.006	94.8	0.993	0.005	0.006	95.3
ϕ_2	1	0.997	0.008	0.009	95.4	0.998	0.008	0.009	95.3
ϕ_3	1	0.996	0.011	0.012	96.5	0.996	0.011	0.012	96.4
Gamma distribution									
θ_{r1}	−0.3	−0.349	0.275	0.312	96.9	−0.347	0.244	0.2816	96.1
θ_{r2}	−0.2	−0.254	0.377	0.432	96.0	−0.255	0.328	0.3833	96.0
θ_{r3}	−1.8	−1.985	2.092	2.634	94.9	−1.061	0.692	0.7692	81.5
θ_{r4}	−3.2	−3.524	5.886	5.684	93.9	−1.772	0.831	0.9543	59.8
A_{11}^*	1.4	1.418	0.037	0.042	96.5	1.418	0.032	0.038	96.7
A_{12}^*	1.6	1.621	0.051	0.058	96.2	1.621	0.044	0.052	96.2
A_{13}^*	1.4	1.480	0.340	0.433	95.1	1.109	0.106	0.119	80.9
A_{14}^*	2	2.143	0.988	0.971	94.1	1.428	0.135	0.152	61.1
ϕ_1	1	0.994	0.007	0.006	93.7	0.995	0.006	0.006	93.6
ϕ_2	1	0.990	0.013	0.013	94.9	0.990	0.012	0.011	93.7
ϕ_3	1	0.996	0.018	0.019	95.9	0.994	0.015	0.016	95.5

†The true latent variable distribution is a normal distribution (upper part of the table) and gamma distribution (lower part of the table). Parameter β consists of four components of θ_r, four components of A^* and three components of ϕ (corresponding to three normal manifest variables), in that order. In θ_r and A^*, the parameters of the normal manifest variables precede those of the binomial variables. The true value β_0, the average estimates $\hat{\beta}$, the sample variance $\text{var}(\hat{\beta})$, the average estimated variance $\widehat{\text{var}}(\hat{\beta})$ and the 95% confidence interval coverage 95%cov are presented. Results are based on 1000 simulations and sample size 500.

Table 3. Simulation study†

β	β_0	*Results from semiparametric method*				*Results from MLE method*			
		$\hat{\beta}$	$var(\hat{\beta})$	$\widehat{var}(\hat{\beta})$	*95%cov (%)*	$\hat{\beta}$	$var(\hat{\beta})$	$\widehat{var}(\hat{\beta})$	*95%cov (%)*
Mixture distribution									
θ_{r1}	−0.3	−0.310	0.077	0.080	95.1	−0.314	0.081	0.085	95.3
θ_{r2}	−0.2	−0.212	0.097	0.100	94.8	−0.217	0.106	0.109	94.9
θ_{r3}	−1.8	−1.845	0.219	0.223	95.5	−2.426	0.334	0.335	86.7
θ_{r4}	−3.2	−3.345	0.501	0.537	96.8	−4.479	0.798	0.840	80.5
A^*_{11}	1.4	1.405	0.007	0.007	95.4	1.406	0.007	0.008	95.6
A^*_{12}	1.6	1.605	0.009	0.009	95.3	1.607	0.010	0.010	95.5
A^*_{13}	1.4	1.419	0.028	0.028	95.5	1.655	0.050	0.049	85.7
A^*_{14}	2	2.062	0.065	0.068	96.5	2.570	0.140	0.144	75.8
ϕ_1	1	0.992	0.005	0.006	95.1	0.993	0.006	0.006	95.1
ϕ_2	1	0.997	0.008	0.008	95.0	0.995	0.010	0.010	94.9
ϕ_3	1	1.003	0.011	0.011	95.9	0.100	0.015	0.015	95.8
Student t-distribution									
θ_{r1}	−0.3	−0.316	0.034	0.037	95.7	−0.316	0.033	0.036	95.6
θ_{r2}	−0.2	−0.215	0.044	0.047	95.9	−0.213	0.042	0.045	96.4
θ_{r3}	−1.8	−1.855	0.166	0.159	94.8	−1.603	0.110	0.105	86.0
θ_{r4}	−3.2	−3.276	0.324	0.333	96.2	−2.801	0.189	0.188	78.1
A^*_{11}	1.4	1.411	0.013	0.014	95.9	1.411	0.012	0.013	95.8
A^*_{12}	1.6	1.609	0.017	0.018	95.4	1.607	0.016	0.017	95.3
A^*_{13}	1.4	1.438	0.071	0.068	94.8	1.269	0.046	0.044	84.3
A^*_{14}	2	2.052	0.133	0.134	95.9	1.737	0.075	0.072	74.9
ϕ_1	1	0.998	0.006	0.006	95.4	0.997	0.006	0.006	94.8
ϕ_2	1	0.995	0.011	0.010	94.6	0.995	0.010	0.009	94.1
ϕ_3	1	0.999	0.014	0.015	94.8	1.001	0.012	0.013	95.5

†The true latent variable distribution is a mixture of normal distributions (upper part of the table) and a Student t-distribution (lower part of the table). Parameter β consists of four components of θ_r, four components of A^* and three components of ϕ (corresponding to three normal manifest variables), in that order. In θ_r and A^*, the parameters of the normal manifest variables precede those of the binomial variables. The true value β_0, the average estimates $\hat{\beta}$, the sample variance $\mathrm{var}(\hat{\beta})$, the average estimated variance $\widehat{\mathrm{var}}(\hat{\beta})$ and the 95% confidence interval coverage 95%cov are presented. Results are based on 1000 simulations and sample size 500.

has been noted in the literature on mixed effects models (Butler and Louis, 1992; Verbeke and Lesaffre, 1997). There, the correct specification of the latent variable distribution contributes to the gain in efficiency of the final estimator. However, our estimator is semiparametric in nature and does not require a correct specification of such a distribution. As a result, such a gain should not be expected here. In contrast, the special flexibility of GLLVMs comes mainly through allowing both continuous and discrete manifest variable components; hence, even with the apparent robustness of the MLE for the continuous manifest variable components, the bias on the discrete manifest variable components cannot be dismissed.

7.2. Bivariate latent variables

To explore extreme outliers and the multivariate situation in the latent variable, we also conducted a simulation where the latent variable is of dimension 2. In the first experiment, the true latent variable follows a bivariate normal distribution with mean $(-0.5, 0.5)^{\mathrm{T}}$ and variance–covariance matrix $(1, 0; 0, 1.5)$. Because the normality of the latent variable holds, we

Table 4. Simulation study†

β	β_0	*Results from semiparametric method*				*Results from MLE method*			
		$\hat{\beta}$	$var(\hat{\beta})$	$\widehat{var}(\hat{\beta})$	95%cov (%)	$\hat{\beta}$	$var(\hat{\beta})$	$\widehat{var}(\hat{\beta})$	95%cov (%)
Bivariate normal distribution									
θ_{r1}	−0.5	−0.526	0.273	0.286	90.8	−0.501	0.199	0.355	94.4
θ_{r2}	−1.5	−1.578	0.294	0.284	92.1	−1.555	0.237	0.311	94.7
θ_{r3}	−1	−1.055	0.394	0.367	91.8	−1.048	0.315	0.396	93.9
A^*_{11}	2.5	2.488	0.323	0.375	91.5	2.515	0.257	0.480	94.1
A^*_{12}	2	2.035	0.238	0.242	92.8	2.066	0.223	0.257	94.9
A^*_{13}	1.5	1.517	0.217	0.218	90.3	1.537	0.175	0.253	93.0
A^*_{21}	1.8	1.823	0.187	0.191	89.8	1.802	0.131	0.242	94.1
A^*_{22}	1.6	1.682	0.253	0.251	91.5	1.665	0.212	0.282	95.0
A^*_{23}	2.1	2.171	0.476	0.492	94.5	2.175	0.457	0.494	94.9
ϕ_1	1	0.927	0.086	0.104	93.0	0.955	0.058	0.089	95.3
ϕ_2	1	0.940	0.145	0.246	90.9	0.947	0.100	0.240	94.4
ϕ_3	1	0.998	0.206	0.285	96.5	0.997	0.204	0.313	95.4
Mixture distribution with extreme outliers									
θ_{r1}	−0.5	−0.500	0.025	0.023	95.8	−0.558	2.075×10^3	4.397×10^7	55.9
θ_{r2}	−1.5	−1.534	0.055	0.060	96.5	−14.163	1.854×10^4	1.019×10^8	62.6
θ_{r3}	−1	−1.125	0.401	0.256	96.0	−21.776	4.305×10^4	1.399×10^8	64.8
A^*_{11}	2.5	2.512	0.016	0.016	96.0	1.7296	3.678×10^2	6.433×10^7	58.8
A^*_{12}	2	2.062	0.065	0.074	96.1	10.5	9.621×10^3	8.438×10^7	65.3
A^*_{13}	1.5	1.609	0.531	0.172	95.9	18.449	7.402×10^4	2.987×10^8	63.6
A^*_{21}	1.8	1.815	0.020	0.019	94.9	0.2577	6.09×10^2	1.789×10^8	49.4
A^*_{22}	1.6	1.652	0.063	0.070	96.3	−4.3972	6.343×10^3	2.083×10^8	59.7
A^*_{23}	2.1	2.362	1.464	0.438	95.5	1.7987	7.889×10^3	4.534×10^7	60.6
ϕ_1	1	0.990	0.092	0.079	94.7	46.532	9.550×10^3	6.833×10^4	21.2
ϕ_2	1	1.028	0.450	0.174	95.6	58.758	9.278×10^3	2.062×10^4	18.0
ϕ_3	1	1.024	0.204	0.225	96.2	34.135	8.542×10^3	6.009×10^3	43.5

†The true latent variable is a bivariate normal distribution (upper part of the table) and a mixture of bivariate normal distribution with extreme outliers (lower part of the table). Parameter β consists of θ_r, A^* and ϕ (corresponding to the normal manifest variables). In θ_r and A^*, the parameters of the normal manifest variables precede those of the binomial variables. The true value β_0, the average estimates $\hat{\beta}$, the sample variance var($\hat{\beta}$), the average estimated variance $\widehat{\text{var}}(\hat{\beta})$ and the 95% confidence interval coverage 95%cov are presented. Results are based on 1000 simulations and sample size 500.

expect consistent estimation from both the semiparametric estimator and the MLE, and it is indeed so in the upper half of Table 4. In the second experiment, 90% of the latent variables follow the bivariate normal distribution that was described above, whereas the other 10% of them are generated from a bivariate normal distribution with mean $(10, -10)^{\mathrm{T}}$ and variance–covariance matrix $(0.25, 0; 0, 1.125)$. The centre of the two distributions is 14.8 away; hence this setting represents a case of extreme outliers in the latent variable distribution, and inference in such situations poses a severe challenge. The result that is associated with the semiparametric estimator is presented in the lower part of Table 4. As we can see, the consistency of the estimator holds even under such an extreme outlier situation. On the contrary, when we computed the MLE in the same simulation setting, the result is not satisfying at all, with huge biases. In general, the variance estimation and 95% coverage are less precise in the bivariate latent variable cases than in the univariate case. It is intuitively clear that, since we have more parameters and more complex models, larger sample sizes are needed to obtain more precise results. Indeed,

in a simulation that is not reported here, when we increased the sample size to $n = 1000$, both the variance estimation and the coverage results improved. In contrast, the variance and 95% coverage are totally off for the MLE when the normal assumption is violated. In fact, it is very difficult to obtain numerical convergence for the MLE, and the computation is much more time consuming as well.

8. Discussion

In this paper, we have relaxed the usual normality assumption on the latent variable distribution of the GLLVMs. Such relaxation induces a much more flexible semiparametric model. In this semiparametric context, we have provided a class of root n consistent estimators which includes the optimal estimator in terms of efficiency of estimation. The estimator is robust in the sense that its consistency holds regardless of the distributional assumption on the latent variables, although the operation does require us to conjecture such a distribution. We developed the estimator in the context of GLLVMs because of the widely spread usage of such models. However, the implicit estimator that was proposed in Section 3.1 is certainly not restricted to a GLLVM. Specifically, if the distribution of the manifest variable deviates from the generalized linear structure, the computational simplification in Section 3.2 will no longer be applicable, but the estimator that was proposed in Section 3.1 is still valid. We believe that the methodology that is proposed here will be of interest to the GLLVM community.

As a special and simpler subclass of GLLVMs, the random-effect model has received much attention. It has been observed (Ma *et al.*, 2004) that, for random-effect models, using a normal distribution assumption on the random effect, the MLE *often, but not always*, produces nearly consistent estimators for the parameters. The consequence of the possible misspecification of the random-effect distribution mainly exhibits in the loss of efficiency. Although GLLVMs bear much similarity to random-effect models, our simulation results have shown that the *often* empirically observed resistance of the MLE method to the distributional assumption on the random effect does not generalize to the more complex and more general GLLVM situations. In contrast, since random-effect models are a special case of GLLVMs, the method that is proposed here is also suitable for random-effect models.

We would like to mention that, in the generalized linear mixed effect model framework, a more flexible model for the random effect, such as the t-distribution (Lee and Nelder, 2006), has been used to increase robustness against the misspecification of the distribution (Noh *et al.*, 2005). Conceptually, the same could be implemented in GLLVMs. However, owing to the complex structure of GLLVMs, the computation is extremely difficult. In addition, the above robustness is only reflected in limited empirical studies; the scope of the robustness and its theoretical justification are still lacking even in the mixed effect model framework. On the contrary, the method that is proposed here is robust (consistent) against any latent variable distribution, and the robustness property is established both theoretically and numerically. When the latent variable model is indeed t, our method provides less efficient estimation than if we assume a t-model (Kang *et al.*, 2005). However, such a loss of efficiency is mandatory whenever we adopt a more flexible model. The results in the upper part of Table 2 also indicate that the loss of efficiency in practice could be negligible.

A different, possibly more classical aspect of robustness concerns the departure of a small proportion of the observed data from the proposed conditional distribution assumption (Huber, 1981; Hampel *et al.*, 1986). There, a bounded influence robust estimator made consistent through an indirect inference procedure has been proposed by Moustaki and Victoria-Feser (2006). This concerns the manifest variable distribution and the robustness is achieved through controlling

the influence of the 'outlying observations'. In contrast, our concern is focused on the distribution of the latent variables and the semiparametric approach yields robustness through 'not modelling' the distribution of the latent variables. It would be of great interest to develop methodologies that are robust in both respects. One possibility is to establish the indirect inference method in a semiparametric framework, which will be challenging and fundamental from both the semiparametric and the robustness inferential point of view.

Acknowledgements

We thank Maria-Pia Victoria-Feser for providing the Swiss consumption data. Ma's research was partially supported by National Science Foundation grant DMS-0906341. Genton's research was partially supported by National Science Foundation grants DMS-0504896 and CMG ATM-0620624, and award KUSC1-016-04 made by King Abdullah University of Science and Technology.

Appendix A

A.1. Proof that $\tilde{A}_1$ has full rank

If $\tilde{A}_1$ is not full rank, then $\tilde{A}_1 = G(B^{\mathrm{T}}0)^{\mathrm{T}}$, where G is a $p \times p$ invertible matrix, and B is a $p_1 \times (p+q)$ matrix with $p_1 < p$. Treat $G^{\mathrm{T}}\tilde{Z}$ as the new set of latent variables; then the last component of $G^{\mathrm{T}}\tilde{Z}$ never appears in the model and hence can be suppressed. This is contradictory to the parsimony of the model.

A.2. Existence and uniqueness of A_1

The parsimony of the model ensures that A_1 has full rank. Let G be the matrix that is formed by the first p linearly independent columns of A_1. Then $G^{-1}A_1$ automatically has the required form.

To see the uniqueness, we consider the opposite. If A_1 and A_2 both satisfy the requirement and $A_1 \neq A_2$, then we have $A_1 = GA_2$, or $(U_0\, e_1\, U_1\, e_2 \ldots e_p\, U_p) = G(\tilde{U}_0\, e_1\, \tilde{U}_1\, e_2 \ldots e_p\, \tilde{U}_p)$. Consider the first block corresponding to $\tilde{U}_0$. We have $0 = G\tilde{U}_0$; hence we obtain that U_0 has the same size as $\tilde{U}_0$. Consequently, we have $e_1 = Ge_1$; hence the first column of G_1 is e_1. Now consider the block corresponding to $\tilde{U}_1$. It is easy to see that $G\tilde{U}_1 = \tilde{U}_1$; hence $U_1 = \tilde{U}_1$ and the second column of G is e_2. Similar arguments can be repeatedly used to establish that $U_i = \tilde{U}_i$ for all $i = 0, \ldots, p$ and that G is the identity. Hence the uniqueness is shown.

A.3. Description and derivation of Λ and $\Lambda^{\perp}$

Consider the Hilbert space $\mathcal{H}$ consisting of all the mean 0, finite variances, length m vector functions of X, where the inner product between two functions h and g is defined as $E(h^{\mathrm{T}}g)$. Here and in the following definitions, all the expectations are calculated under the true distribution. The nuisance tangent space Λ is a subspace of $\mathcal{H}$ defined as the mean-squared closure of all the elements of the form BS, where S is an arbitrary nuisance score vector function, and B is any conformable matrix with m rows. Here the nuisance score vector functions are calculated conventionally in every possible valid parameterization of the infinite dimensional nuisance parameter, where a 'valid parameterization' means that there is one parameter value which yields the truth. Furthermore, $\Lambda^{\perp}$ is defined to be the orthogonal complement of Λ in $\mathcal{H}$.

A semiparametric approach mainly consists of several steps. The first step is to identify the two subspaces Λ and $\Lambda^{\perp}$. The second step is to calculate the usual score function with respect to the parameter of interest. The third step is to project this score function onto $\Lambda^{\perp}$. The projection is computed using the operation in $\mathcal{H}$, and the result is used to build estimating equations for the parameter of interest.

Following the above description, we now derive Λ and $\Lambda^{\perp}$ in our model. In the expression of $f_X(\cdot)$ in model (1), the only nuisance parameter is $f_Z(\cdot)$, which is not subject to any constraints except that it is a valid density function. Thus any mean 0 function $h(\cdot)$ could be the result of $\partial \log\{f_Z(\cdot;\gamma)\}/\partial\gamma|_{\gamma=\gamma_0}$. Specifically, we could have

$$f_Z(z;\gamma) = \frac{f_{Z0}(z)[1+\exp\{-2\gamma\, h(z)\}]^{-1}}{\int f_{Z0}(z)[1+\exp\{-2\gamma\, h(z)\}]^{-1}\,\mathrm{d}\mu(z)}$$

with $\gamma_0 = 0$, and it is straightforward to verify that $h(\cdot) = \partial \log\{f_Z(\cdot;\gamma)\}/\partial\gamma|_{\gamma=\gamma_0}$. Thus the nuisance score vector of $f_X(x;\beta,\gamma)$ can be calculated as

$$\frac{\partial \log\left\{\int f_{X|Z}(x|z)\, f_Z(z;\gamma)\,\mathrm{d}\mu(z)\right\}}{\partial\gamma} = \frac{\int f_{X|Z}(x|z)\,\partial\log\{f_Z(z;\gamma)\}/\partial\gamma\, f_Z(z)\,\mathrm{d}\mu(z)}{\int f_{X|Z}(x|z)\, f_Z(z;\gamma)\,\mathrm{d}\mu(z)} = E\{h(Z)|x\}.$$

We therefore obtain the expression for Λ. It is easily verified that, for any $g(x)$ that satisfies $E\{g(X)|z\} = 0$, we have $E[g(X)^{\mathrm{T}} E\{h(Z)|X\}] = E[E\{g(X)^{\mathrm{T}}|Z\}\, h(Z)] = 0$; hence $g \perp \Lambda$, or $g \in \Lambda^{\perp}$. However, if $g \in \Lambda^{\perp}$, then $g(\cdot)$ is such that $E[g^{\mathrm{T}}(X)\{h(Z)|X\}] = 0$ holds for any mean 0 function $h(Z)$, i.e. $0 = E[g^{\mathrm{T}}(X)\, E\{h(Z)|X\}] = E[E\{g^{\mathrm{T}}(X)|Z\}\, h(Z)]$ for any mean 0 $h(Z)$. Thus $E\{g(X)|Z\}$ itself must be 0. Therefore we have shown the validity of the form of $\Lambda^{\perp}$ that is given in the main text.

A.4. Derivation of the projected score functions

Straightforward calculation shows that

$$\begin{aligned}
S^*_{\mathrm{eff}}|_{\alpha_{\mathrm{con},k}} &= E^*\{\phi_k^{-1} e_k^{\mathrm{T}} X - \phi_k^{-1} b_k'(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z)|w, y\} - E^*\{\phi_k^{-1} e_k^{\mathrm{T}} X - \phi_k^{-1} b_k'(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z)|w\} \\
&= \phi_k^{-1} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} - E^*(\phi_k^{-1} e_k^{\mathrm{T}} X|w) = \phi_k^{-1} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} - \phi_k^{-1} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ E^*(Y|w) \end{pmatrix} \\
&= \phi_k^{-1} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E^*(Y|w) \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
S^*_{\mathrm{eff}}|_{\alpha_{lk}} &= \phi_k^{-1}[E^*\{Z^{\mathrm{T}} O_{lk} X - b_k'(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z) Z_l|w, y\} - E^*\{Z^{\mathrm{T}} O_{lk} X - b_k'(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z) Z_l|w\}] \\
&= \phi_k^{-1} E^*\{Z^{\mathrm{T}} O_{lk} X|w, y\} - \phi_k^{-1} E^*(Z^{\mathrm{T}} O_{lk} X|w) \\
&= \phi_k^{-1} E^*(Z|w)^{\mathrm{T}} O_{lk} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} - \phi_k^{-1} E^*(Z|w)^{\mathrm{T}} O_{lk} \Phi A^{-1} \begin{pmatrix} w \\ E^*(Y|w) \end{pmatrix} \\
&= \phi_k^{-1} E^*(Z|w)^{\mathrm{T}} O_{lk} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E^*(Y|w) \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
S^*_{\mathrm{eff}}|_{\phi_k} &= E^*\{-\phi_k^{-2}\alpha_{\mathrm{con},k} e_k^{\mathrm{T}} X - \phi_k^{-2} Z^{\mathrm{T}}(0\ \alpha_k\ 0) X + \phi_k^{-2} b_k(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z) + c_{k2}'(x^{(k)}, \phi_k)|w, y\} \\
&\quad - E^*\{-\phi_k^{-2}\alpha_{\mathrm{con},k} e_k^{\mathrm{T}} X - \phi_k^{-2} Z^{\mathrm{T}}(0\ \alpha_k\ 0) X + \phi_k^{-2} b_k(\alpha_{\mathrm{con},k} + \alpha_k^{\mathrm{T}} Z) + c_{k2}'(x^{(k)}, \phi_k)|w\} \\
&= -\phi_k^{-2}\alpha_{\mathrm{con},k} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} - \phi_k^{-2} E^*(Z|w)^{\mathrm{T}}(0\ \alpha_k\ 0) \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix} + E^*\{c_{k2}'(x^{(k)}, \phi_k)|w, y\} \\
&\quad + \phi_k^{-2}\alpha_{\mathrm{con},k} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ E^*(Y|w) \end{pmatrix} + \phi_k^{-2} E^*(Z|w)^{\mathrm{T}}(0\ \alpha_k\ 0) \Phi A^{-1} \begin{pmatrix} w \\ E^*(Y|w) \end{pmatrix} - E^*\{c_{k2}'(x^{(k)}, \phi_k)|w\} \\
&= -\phi_k^{-2}\alpha_{\mathrm{con},k} e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} 0 \\ y - E^*(Y|w) \end{pmatrix} - \phi_k^{-2} E^*(Z|w)^{\mathrm{T}}(0\ \alpha_k\ 0) \Phi A^{-1} \begin{pmatrix} 0 \\ y - E^*(Y|w) \end{pmatrix} \\
&\quad + c_{k2}'\left(e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_k\right) - E^*\left\{c_{k2}'\left(e_k^{\mathrm{T}} \Phi A^{-1} \begin{pmatrix} w \\ y \end{pmatrix}, \phi_k\right)\middle| w\right\}.
\end{aligned}$$

A.5. Calculation of maximizing t(y)

We have the equivalence

$$\max_y\{t(y)\} \Leftrightarrow \max_y\left\{\theta_{\text{con}}^{\text{T}}A^{-1}\begin{pmatrix}w\\y\end{pmatrix}+\sum_{j=1}^{p+q}c_j\left(e_j^{\text{T}}\Phi A^{-1}\begin{pmatrix}w\\y\end{pmatrix},\phi_j\right)\right\}$$

$$\Leftrightarrow (0\ I_q)A^{-\text{T}}\theta_{\text{con}}+\sum_{j=1}^{p+q}c'_{j1}\left(e_j^{\text{T}}\Phi A^{-1}\begin{pmatrix}w\\y_0\end{pmatrix},\phi_j\right)(0\ I_q)A^{-\text{T}}\Phi e_j=0$$

$$\Leftrightarrow (-A^{*\text{T}}I_q)\left\{\theta_{\text{con}}+\sum_{j=1}^{p+q}c'_{j1}\left(e_j^{\text{T}}\Phi A^{-1}\begin{pmatrix}w\\y_0\end{pmatrix},\phi_j\right)\Phi e_j\right\}=0,$$

where c'_{j1} is the derivative of c_j with respect to the first argument. Thus the result follows.

References

Bartholomew, D. J. (1980) Factor analysis for categorical data (with discussion). *J. R. Statist. Soc.* B, **42**, 293–321.

Bartholomew, D. J. (1984a) The foundations of factor analysis. *Biometrika*, **71**, 221–232.

Bartholomew, D. J. (1984b) Scaling binary data using a factor model. *J. R. Statist. Soc.* B, **46**, 120–123.

Bartholomew, D. J. and Knott, M. (1999) *Latent Variable Models and Factor Analysis*. London: Hodder Arnold.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Butler, S. M. and Louis, T. A. (1992) Random effects models with nonparametric priors. *Statist. Med.*, **11**, 1981–2000.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: the Approach based on Influence Functions*. New York: Wiley.

Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.

Huber, P., Ronchetti, E. and Victoria-Feser, M.-P. (2004) Estimation of generalized linear latent variable models. *J. R. Statist. Soc.* B, **66**, 893–908.

Jöreskog, K. (1967) Some contributions to maximum likelihood factor analysis. *Psychometrika*, **32**, 443–482.

Kang, W., Lee, M. S. and Lee, Y. (2005) HGLM versus conditional estimators for the analysis of clustered binary data. *Statist. Med.*, **24**, 741–752.

Lee, Y. and Nelder, J. A. (2006) Double hierarchical generalized linear models (with discussion). *Appl. Statist.*, **55**, 139–185.

Lindsay, B. C. (1982) Conditional score functions: some optimality results. *Biometrika*, **69**, 503–512.

Lindsay, B. C. (1983) Efficiency of the conditional score in a mixture setting. *Ann. Statist.*, **11**, 486–497.

Lindsay, B. C. (1985) Using empirical partially Bayes inference for increasing efficiency. *Ann. Statist.*, **13**, 914–931.

Ma, Y., Genton, M. G. and Davidian, M. (2004) Linear mixed effects models with flexible generalized skew-elliptical random effects. In *Skew-elliptical Distributions and Their Applications: a Journey beyond Normality* (ed. M. G. Genton), pp. 339–358. Boca Raton: Chapman and Hall–CRC.

Moustaki, I. (1996) A latent trait and a latent class model for mixed observed variables. *Br. J. Math. Statist. Psychol.*, **49**, 313–334.

Moustaki, I. and Knott, M. (2000) Generalized latent trait models. *Psychometrika*, **65**, 391–411.

Moustaki, I. and Victoria-Feser, M.-P. (2006) Bounded-influence robust estimation in generalized linear latent variable models. *J. Am. Statist. Ass.*, **101**, 644–653.

Newey, W. K. (1990) Semiparametric efficiency bounds. *J. Appl. Econometr.*, **5**, 99–135.

Noh, M., Lee, Y. and Pawitan, Y. (2005) Robust ascertainment-adjusted parameter estimation. *Genet. Epidem.*, **29**, 68–75.

Sammel, M. D., Ryan, L. M. and Legler, J. M. (1997) Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc.* B, **59**, 667–678.

Sartori, N. and Severini, T. A. (2004) Conditional likelihood inference in generalized linear mixed models. *Statist. Sin.*, **14**, 349–360.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. London: Chapman and Hall.

Spearman, C. (1904) General intelligence objectively determined and measured. *Am. J. Psychol.*, **15**, 201–293.

Stefanski, L. A. and Carroll, R. J. (1987) Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**, 703–716.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

Tsiatis, A. A. and Ma, Y. (2004) Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91**, 835–848.

Verbeke, G. and Lesaffre, E. (1997) The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computnl Statist. Data Anal.*, **23**, 541–556.

Yau, K. and McGilchrist, C. (1996) Simulation study of the GLLVM method applied to the analysis of clustered survival data. *J. Statist. Computn Simuln*, **55**, 189–200.