Scandinavian Journal of Statistics Theory and Applications

doi: 10.1111/j.1467-9469.2011.00777.x.

Generalized Linear Latent Variable Models with Flexible Distribution of Latent Variables

IRINA IRINCHEEVA and EVA CANTONI

University of Geneva

MARC G. GENTON

Texas A&M University

ABSTRACT. We consider a semi-nonparametric specification for the density of latent variables in Generalized Linear Latent Variable Models (GLLVM). This specification is flexible enough to allow for an asymmetric, multi-modal, heavy or light tailed smooth density. The degree of flexibility required by many applications of GLLVM can be achieved through this semi-nonparametric specification with a finite number of parameters estimated by maximum likelihood. Even with this additional flexibility, we obtain an explicit expression of the likelihood for conditionally normal manifest variables. We show by simulations that the estimated density of latent variables capture the true one with good degree of accuracy and is easy to visualize. By analysing two real data sets we show that a flexible distribution of latent variables is a useful tool for exploring the adequacy of the GLLVM in practice.

Key words: factor analysis, latent variable, non-Gaussian distribution, semi-nonparametric distribution, visualization

1. Introduction

Latent variables, as hypothetical constructs, are omnipresent in almost all sciences and in daily life. Indeed, constructs such as quality of life, physical health or disease are widespread in research and applications but cannot be measured directly. Usually scientists make and validate inference on those constructs with help of latent variable models using observable variables as proxies. In the aforementioned examples, we can imagine quality of life to be modelled through economic wealth and access to drinking water; physical health can be assessed through cholesterol and haemoglobin rates, body mass index, eyesight, hearing and presence of chronic diseases; virus infection or other diseases can be revealed by fever, level of some particular antibodies, erythrocyte sedimentation rate, level of C-reactive protein. The principal aim of Generalized Linear Latent Variable Models (GLLVM, concept by Bartholomew, 1980 and 1984) is to explain most of the variability of p observed (manifest) variables $X_1, \ldots, X_p$ by constructing $q < p$ latent variables $Z_1, \ldots, Z_q$. To this aim, GLLVM assumes that ΓZ, with unknown parameter matrix $\Gamma \in \mathbb{R}^{p \times q}$ and $Z = (Z_1, \ldots, Z_q)^T$, explains all the systematic variability of the manifest variables via the conditional probability density, or mass, function

$$g(x \mid z) = \prod_{j=1}^{p} g_j(x_j \mid \mu_j + \gamma_j^T z), \tag{1}$$

where $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, $z = (z_1, \ldots, z_q)^T \in \mathbb{R}^q$, μ_j is the location parameter of x_j, $\gamma_j \in \mathbb{R}^q$ is the jth row of the $p \times q$ parameter matrix Γ, $g_j(\cdot)$ is a probability density or mass function of a distribution from the exponential family. The marginal probability density or mass function of the manifest variables is

$$f(x \mid \mu, \Gamma, \psi) = \int_{\mathbb{R}^q} g(x \mid z) h(z)\, \mathrm{d}z$$
$$= \int_{\mathbb{R}^q} \left[\prod_{j=1}^{p} \exp\left\{ \frac{x_j(\mu_j + \gamma_j^T z) - b_j(\mu_j + \gamma_j^T z)}{\psi_j} + c_j(x_j, \psi_j) \right\} \right] h(z)\, \mathrm{d}z \qquad (2)$$

with functions $b_j(\cdot), c_j(\cdot,\cdot)$ and, in some cases, additional scale parameter $\psi_j \in \mathbb{R}$ with $\psi = (\psi_1, \ldots, \psi_p)^T$. The linear combination $\mu_j + \gamma_j^T z$ is related to the expected values of $X_j \mid z$ through the link function denoted here as $\theta_j^{-1}(\cdot)$: $\mathrm{E}(X_j \mid z) = \theta_j^{-1}(\mu_j + \gamma_j^T z)$. If needed, observable covariates $Y_1, \ldots, Y_m$ possibly explaining the manifest variables $X_1, \ldots, X_p$ can be introduced by setting $\mathrm{E}(X_j \mid z, y) = \theta_j^{-1}(\mu_j + \beta_j^T y + \gamma_j^T z)$, where $y, \beta_j \in \mathbb{R}^m$ and $\mu = (\mu_1, \ldots, \mu_p)^T, \beta = (\beta_1, \ldots, \beta_p)^T$ and Γ are parameters to be estimated. Then model (2) becomes a generalization (for responses with distribution from the exponential family) of the responses equation in a structural equation model as in Rabe-Hesketh & Skrondal (2004, page 78). Although the structural relation among latent variables is not the focus of GLLVM, its modelling can be done straightforwardly through an additional structural equation as in Liu *et al*. (2005). Model (2) with observable covariates can be generalized to two or more levels by adding additional subscript(s) to manifest and latent variables as in Rabe-Hesketh & Skrondal (2004, page 99), which renders possible modelling of, for example, multivariate longitudinal data as in Cagnone *et al*. (2009) or Dunson (2003). Thus GLLVM with observable covariates can be seen as an approach to a multivariate generalized mixed effects model.

Traditionally it is assumed that the density $h(z)$ of the latent variables is multivariate normal. Bartholomew (1988) advocated the adequacy of the normal distribution for two principal reasons. The first reason is the 'arbitrariness about the direction of measurement of a latent scale', for example, the convention that high customer satisfaction is given a high score on the corresponding latent variable. Bartholomew (1988) suggested that only symmetric distributions of latent variables can overcome this arbitrariness. This statement is refuted by Montanari & Viroli (2010b) who showed that any distribution of latent variables would overcome this arbitrariness. The second reason in which Bartholomew (1988) believed is that an incorrect specification of the latent variables distribution would not affect the estimates. To the contrary, Ma & Genton (2010) described settings of GLLVM where an inappropriate specification of the asymmetric latent variables distribution biases the estimates.

Using alternatives to normality for the latent variables is not new in the statistical literature on GLLVM and its submodels. For instance in factor analysis Montanari & Viroli (2010b) introduced skew-normal latent variables alluding to the frequent asymmetry of appreciations; Yung (1997) and Montanari & Viroli (2010a) modelled latent variables via mixture of normals in order to handle heterogeneity of clusters; Wedel & Kamakura (2001) assumed the latent variables to have a continuous distribution from the exponential family in order to construct test statistics. To date, the explicit expression of the integral in (2) exists only when $g(x \mid z)$ is multivariate normal and the distribution of latent variables $h(z)$ is multivariate normal or mixture of normals as in Yung (1997) and Montanari & Viroli (2010a). For the other cited cases, a numerical approximation of the integral is required. The model we propose approximates arbitrarily closely a wider class of latent variable distributions (and then of manifest variables too) than models proposed by Montanari & Viroli (2010b) and Wedel & Kamakura (2001) yet yields an explicit expression of the integral in (2) in case of conditionally normal manifest variables.

In another submodel of GLLVM, the latent trait model with binary manifest, Knott & Tzamourani (2007) estimated the latent variables distribution by bootstrap combined with non-parametric maximum likelihood and concluded that the usual normality assumption of

the latent variables 'is not always justified'. The semi-nonparametric (SNP) approach being different from the non-parametric maximum likelihood estimation (Laird, 1978) has an appealing smoothness property. The smooth density of the latent variable is easier to grasp: we do not have to take into account the possibility of other differently defined supports as in the case of the discrete mass-point distribution.

In the case of univariate responses the appropriateness of the normal distribution for other latent variables (or random effects) models have been studied: for structural measurement error models in Huang *et al*. (2006), for generalized mixed effects models in Rabe-Hesketh *et al*. (2003), who estimated the latent variable distribution with non-parametric likelihood, and Chen *et al*. (2002), who used the SNP approach as in the present article.

In addition to the conclusion about the inappropriateness of the normality assumption for the latent variables Rabe-Hesketh *et al*. (2003) highlighted the importance of the correct distributional assumptions for the prediction of latent scores. The estimated density of latent scores is simply the estimated density of latent variables. Its inappropriate specification and visualization lead to overlooking clusters, outliers and misinterpretation of the estimation results.

In some cases, the inadequacy of the normally distributed latent variables can be caused by the nonlinear dependence on latent variables as explored for structural equation models by Wall & Amemiya (2000) and generalized latent variable models by Rizopoulos & Moustaki (2008).

In this article we consider GLLVM with both discrete and continuous manifest variables and propose $h(z)$ in (2) to have the SNP specification introduced by Gallant & Nychka (1987)

$$h(z) = P_L^2(z)\phi(z), \qquad P_L(z) = \sum_{0 \le i_1 + \cdots + i_q \le L} a_{i_1 \ldots i_q} z_1^{i_1} \cdots z_q^{i_q}, \tag{3}$$

where $a_{i_1 \ldots i_q}$ are real coefficients of the polynomial $P_L(z)$ with a tuple $i_1, \ldots, i_q$ such that $i_1, \ldots, i_q \ge 0$ and $\phi(z)$ is the q-variate standard normal $N_q(0, I)$ density. It is straightforward to see that $L = 0$ corresponds to the case $Z \sim N_q(0, I)$. Further in the article we discuss the parametrization of $a_{i_1 \ldots i_q}$ (section 2.1) and how the flexibility and number of modes of the SNP density increase with the degree L of the polynomial P_L (section 2.3).

Combining the GLLVM settings (2) and the SNP specification (3) results in what we call an SNP-GLLVM, where the marginal probability density or mass function of the manifest variables X is

$$f(x \mid \mu, \Gamma, \psi, P_L) = \frac{1}{(2\pi)^{q/2}} \int_{\mathbb{R}^q} g(x \mid z) P_L^2(z) \exp\left\{-\frac{1}{2} z^T z\right\} \mathrm{d}z, \tag{4}$$

where the expression for $g(x \mid z)$ is given by (1).

In what follows we propose a necessary and sufficient condition for the identifiability of (4) and define estimators $\hat{\mu}, \hat{\Gamma}, \hat{\psi}$ via maximum likelihood. For conditionally normal manifest variables, the integral in (4) can be computed explicitly. One of our main results is the demonstration by simulations that in some GLLVM settings the incorrect specification of the latent variables distribution biases the estimators $\hat{\mu}$ and $\hat{\Gamma}$.

The estimated SNP density of the latent variables (or latent scores distribution) is easy to visualize which is an advantage when compared to the semiparametric GLLVM estimator proposed recently by Ma & Genton (2010). An obvious non-normality (multi-modality and/or skewness) of latent scores distribution can indicate the presence of outliers, possible nonlinearity in dependence on latent variables, non-homogeneity of population or simply the inadequacy of the normal latent density to the particular data.

2. Semi-nonparametric GLLVM

2.1. Parametrization of $P_L(z)$

Restrictions must be imposed on the coefficients of $P_L(z)$ in order for $h(z)$ in (3) to be a density. This can be done as in Gallant & Tauchen (1989) by introducing a proportionality constant $1/\int P_L^2(z)\phi(z)\,\mathrm{d}z$ and setting the constant term of the polynomial equal to 1. Here we choose another parametrization of $P_L(z)$ that avoids difficulties of constrained optimization. This parametrization is proposed by Zhang & Davidian (2001) and consists in rewriting the validity condition on $h(z)$ as

$$1=\int_{\mathbb{R}^q} P_L^2(z)\phi(z)\,\mathrm{d}z=\ \mathrm{E}\{P_L^2(W)\}=a^T\mathrm{E}(\tilde{W}\tilde{W}^T)a=a^TAa, \tag{5}$$

with $W\sim N_q(0,I)$, $P_L(W)=a^T\tilde{W}$, $\tilde{W}=(1,W_1,\ldots,W_q,W_1^2,W_1W_2,W_2^2,\ldots,W_q^L)^T$, so that A is a positive definite matrix by definition. Therefore, there exists a positive definite matrix B such that $A=B^TB$. Defining $c=Ba$, (5) becomes $c^Tc=1$. Hence, $c=(c_1,\ldots,c_d)^T$ can be represented in polar coordinates: $c_1=\sin\varphi_1$, $c_2=\cos\varphi_1\sin\varphi_2,\ldots$, $c_{d-1}=\cos\varphi_1\cdots\cos\varphi_{d-2}\sin\varphi_{d-1}$, $c_d=\cos\varphi_1\cos\varphi_2\cdots\cos\varphi_{d-2}\cos\varphi_{d-1}$, with angles $-\pi/2<\varphi_t\le\pi/2$, $t=1,\ldots,d-1$ in order for c to take values only on a half of the unit sphere in $\mathbb{R}^d$. More details on the polar coordinates transformation can be found in Scott (1992). Note that $d=\sum_{k=0}^{L}\binom{q+k-1}{k}$ according to Stetter (2004, page 228).

Thus, the density (3) can be rewritten as

$$h(z\mid\varphi,L)=P_L^2(z)\phi(z)=\left(a^T\tilde{z}\right)^2\phi(z), \tag{6}$$

where $a=B^{-1}c$, $\tilde{z}=(1,z_1,\ldots,z_q,z_1^2,z_1z_2,z_2^2,\ldots,z_q^L)^T$ and $\varphi=(\varphi_1,\ldots,\varphi_{d-1})^T$. For example, when $q=1$ (one latent variable), $L=2$ and $P_L(z)=a_0+a_1z+a_2z^2$, we obtain $a_0=\sin\varphi_1-\frac{1}{\sqrt{2}}\cos\varphi_1\cos\varphi_2$, $a_1=\cos\varphi_1\sin\varphi_2$ and $a_2=\frac{1}{\sqrt{2}}\cos\varphi_1\cos\varphi_2$.

2.2. Identifiability and constraints

As noted by many researchers, for example Rabe-Hesketh & Skrondal (2001) and Rabe-Hesketh & Skrondal (2004), the major difficulty of all the models with latent variables is identifiability. According to Hastie *et al.* (2001, page 494): 'this aspect has left many analysts skeptical of factor analysis, and may account for its lack of popularity in contemporary statistics'.

A parametric statistical model is said to be identified if distinct values of parameters correspond to distinct probability density or mass functions of the response variables. With this definition we investigate how any affine transformation of the SNP latent variables Z affects the resulting probability density or mass function (4) of the random vector X.

Proposition 1. *For any $P_L^2(z)\neq 0$, the orthogonal transformation $Z_1=CZ$, $(CC^T=C^TC=I_q)$ is the one and only one affine transformation of the random vector Z leaving the probability density or mass function* (4) *of the random vector X unchanged.*

Corollary 1. *The loadings matrix Γ is defined up to to the orthogonal transformation $Z_1=CZ$, $(CC^T=C^TC=I_q)$.*

The proofs of these results can be found in the supplementary material.

Huber *et al.* (2004) demonstrated that if the loadings matrix is defined up to an orthogonal transformation then a sufficient condition for identifiability of a GLLVM with q normal latent variables is that $q(q-1)/2$ elements of the matrix Γ are set to zero. In other words,

after permutations the elements of the upper triangle of Γ should be constrained. The same authors proved that if, in addition, at least one of the elements of the loadings matrix is constrained to be either smaller or larger than zero, then the loadings matrix is necessarily identified. We can find similar conclusions in Ma & Genton (2010) for semiparametric GLLVM. The same number of $q(q-1)/2$ constraints is used by Jöreskog (1967) to obtain a single solution in factor analysis.

2.3. Multi-modality and flexibility of SNP

Similar to distributions considered by Ma & Genton (2004), the number of modes of the semi-nonparametric density increases with the degree L of the polynomial $P_L(z)$ and the number q of latent variables. Indeed, a necessary condition for a mode (local extremum) at the point z_0 is a null gradient:

$$\frac{\partial}{\partial z} P_L^2(z)\phi(z)\Big|_{z=z_0} = \left[\left\{2\frac{\partial P_L(z)}{\partial z} - zP_L(z)\right\} P_L(z)\phi(z)\right]\Big|_{z=z_0} = 0, \text{ i.e.,} \quad \text{either } P_L(z)\Big|_{z=z_0} = 0 \tag{7}$$

$$\text{or } \left\{2\frac{\partial P_L(z)}{\partial z} - zP_L(z)\right\}\Big|_{z=z_0} = 0. \tag{8}$$

The set of real solutions of (7) can contain from 0 to L distinct manifolds of dimension $q-1$ or less. It is easy to see that the solution of (7), if it exists, always corresponds to a local minimum of the density $P_L^2(z)\phi(z)$. Thus, if (7) has up to L different solutions, $P_L^2(z)\phi(z)$ has up to $L-1$ different modes. Independently of the fitted data, if L is odd, the SNP density is equal to zero on a manifold of dimension $q-1$ (i.e. if $q=2$ and L is odd then the SNP density has to be equal to zero on a curve in $\mathbb{R}^2$). For this reason high odd degrees L should be avoided.

Equation (8) is a system of q polynomials where each polynomial is of degree $L+1$ and depends on q variables. In a regular case, that is, without assuming that some coefficients in (6) are null, the system (8) can have up to $(L+1)^q$ different isolated point solutions (i.e. solutions containing only one point, not manifolds such as curves) according to Stetter (2004, page 228).

Defining the number of points where both (7) and (8) hold is not trivial. But assuming that (7) has L different isolated point solutions in which $\partial P_L(z)/\partial z = 0$ and (8) has $(L+1)^q$ different isolated point solutions, we obtain that (7) and (8) together define at most $(L+1)^q$ isolated point solutions. This implies that an SNP density can have at most three modes when $L=2,\ q=1$; four modes when $L=2,\ q=2$; and 13 modes when $L=2,\ q=3$. Our practical experience when $L=2,\ q=1$ and $L=2,\ q=2$ confirms these conclusions.

The sufficient condition for a local maximum (minimum) at $z=z_0$ is that the Hessian matrix $\partial P_L^2(z)\phi(z)/\partial z\partial z^T$ at this point is negative definite (respectively positive definite):

$$\begin{aligned}\frac{\partial}{\partial z\partial z^T} P_L^2(z)\phi(z) = {} & 2\frac{\partial P_L(z)}{\partial z}\frac{\partial P_L(z)}{\partial z^T}\phi(z) + 2\frac{\partial^2 P_L(z)}{\partial z\partial z^T} P_L(z)\phi(z) \\ & + zz^T P_L(z)\phi(z) - 2\frac{\partial P_L(z)}{\partial z} z^T P_L(z)\phi(z) \\ & - 2z\frac{\partial P_L(z)}{\partial z^T} P_L(z)\phi(z) - I_q P_L^2(z)\phi(z).\end{aligned} \tag{9}$$

Once a density of latent variables is estimated, the solutions of (8) can be found numerically and the expression (10) estimated at these points. Hence, the number of modes can be established.

3. Inference in SNP-GLLVM

3.1. Conditionally normal manifest variables

Suppose the conditional density of manifest variables given the latent ones are multivariate normal with the structural scale parameter given by a diagonal positive definite matrix $\Psi = \operatorname{diag}(\psi) \in \mathbb{R}^{p\times p}$. Then the marginal density of x can be written as:

$$f(x;\mu,\Gamma,\Psi,P_L) = |2\pi(\Psi+\Gamma\Gamma^T)|^{-1/2} \exp\left\{-\frac{1}{2}x_0^T(\Psi+\Gamma\Gamma^T)^{-1}x_0\right\} \mathrm{E}_{z\,|\,x_0}\{P_L^2(z)\}, \tag{10}$$

where $x_0 = x-\mu$, $B = I_q + \Gamma^T\Psi^{-1}\Gamma$ and

$$\mathrm{E}_{z\,|\,x_0}\{P_L^2(z)\} = |2\pi B^{-1}|^{-1/2} \int_{\mathbb{R}^q} P_L^2(z) \times \exp\left\{-\frac{1}{2}(z-B^{-1}\Gamma^T\Psi^{-1}x_0)^T B(z-B^{-1}\Gamma^T\Psi^{-1}x_0)\right\} \mathrm{d}z. \tag{11}$$

As all the moments of the multivariate normal distribution are known and completely defined by the first two moments, $\mathrm{E}_{z\,|\,x_0}\{P_L^2(z)\}$ exists in explicit form and represents a $2L$-degree polynomial in x_0. Hence the marginal density of x exists in closed form. When $P_L(z)\equiv 1$ we obtain the classical factor analysis model for normally distributed manifest variables, as described, for example, by Mardia *et al.* (1979).

Using (10) we obtain the following log-likelihood function

$$\ell(\mu,\Gamma,\Psi,\varphi,L \mid x_1,\ldots,x_n) = -\frac{n}{2}\log|2\pi(\Psi+\Gamma\Gamma^T)| - \frac{1}{2}\sum_{i=1}^n x_{0i}^T(\Psi+\Gamma\Gamma^T)^{-1}x_{0i} + \sum_{i=1}^n \log\left[\mathrm{E}_{z\,|\,x_{0i}}\{P_L^2(z)\}\right]. \tag{12}$$

The parameters of interest are those inherited from factor analysis, namely μ,Γ,Ψ, with additional parameters φ and L responsible for the shape of the latent variables density. In practice L is fixed by the rule to be discussed in section 3.3. The final estimators are defined as

$$\hat{\mu} = \mu^* + \Gamma^*\,\tilde{\mathrm{E}}(Z), \qquad \hat{\Gamma} = \Gamma^*\ \ \tilde{\mathrm{cov}}^{1/2}(Z), \quad \hat{\Psi} = \Psi^*, \tag{13}$$

where $(\mu^*,\Gamma^*,\Psi^*,\varphi^*) = \operatorname{argmax}_{\mu,\Gamma,\Psi,\varphi}\ell(\mu,\Gamma,\Psi,\varphi\,|\,L,x_1,\ldots,x_n)$, $\tilde{\mathrm{E}}(Z)$ and $\tilde{\mathrm{cov}}^{1/2}(Z)$ are found given φ^* and the SNP density (6). Thus, $\hat{\mu}$ and $\hat{\Gamma}$ are the estimators corresponding to the uncorrelated latent variables with zero expectation and unit variance.

In the optimization of $\ell(\mu,\Gamma,\Psi,\varphi\,|\,L,x_1,\ldots,x_n)$ we use an analytically computed gradient and Hessian matrix (the gradient can be found in section S4 of the supplementary material, the Hessian expression is available upon request). It should be stressed that the Hessian is computed in a matrix form offering a considerable advantage in R implementation compared to existing Hessian matrix computations such as in Lawley (1967), Jennrich & Thayer (1973) and Ramsey (2010). The optimization is done in R with the nlminb function and is sensitive to the choice of initial values. We discuss how to cope with this problem at the end of the next section.

3.2. Mixture of discrete and continuous manifest variables

In practice the presence of both continuous and discrete responses is more frequent than exclusively continuous responses. Suppose, for example, that among p manifest variables the first p_1 are normal conditionally on the latent variables, and the last $p-p_1$ are conditionally Bernoulli, that is, the joint conditional probability mass function from (1) is

$$g(x \mid z)=\prod_{j=1}^{p_1}\left[\frac{1}{\sqrt{2\pi\psi_j}}\exp\left\{-\frac{1}{2\psi_j}(x_j-\mu_j-\gamma_j^T z)^2\right\}\right]\prod_{j=p_1+1}^{p}\left\{\frac{\exp(x_j\mu_j+x_j\gamma_j^T z)}{1+\exp(\mu_j+\gamma_j^T z)}\right\}, \qquad (14)$$

where the expression in the last brackets is obtained by setting $\mathrm{pr}(x_j=1)=p_j$ and choosing the logit link, that is, $\log\{p_j/(1-p_j)\}=\mu_j+\gamma_j^T z$. Then, the manifest variables marginal density for the SNP-GLLVM model is obtained straightforwardly by plugging (14) in (4). We approximate the corresponding log-likelihood function $\ell(\mu,\Gamma,\Psi,\varphi \mid L, x_1,\ldots,x_n)$ with one latent variable by computing the integral with the R command integrate. The latter uses multiple algorithms including different adaptive integration algorithms for which 'the evaluation points are clustered in the neighbourhood of difficult spot of each integrand' (Piessens *et al.*, 1983).

In a similar GLLVM setting but with normal distribution of latent variables, Huber *et al.* (2004) implemented a Laplace approximation of the integral. We highlight here that the Laplace approximation for integrals is designed for integrand with only one absolute maximum (De Bruijn, 1981, page 63) and cannot be used as approximation of the integrand in (4) which can have multiple local maxima. Similarly, Gaussian or Gauss–Hermite quadratures and adaptive Gaussian quadrature as in Schilling & Bock (2005) will perform poorly. Other alternatives for computing the integral would be to consider a Monte Carlo EM algorithm as implemented by Chen *et al.* (2002).

The estimators $\hat{\mu}$ and $\hat{\Gamma}$ are defined as in (13) with the approximation $\tilde{\ell}(\mu,\Gamma,\Psi,\varphi \mid L, x_1,\ldots, x_n)$ (due to the integral) of the log-likelihood function $\ell(\mu,\Gamma,\Psi,\varphi \mid L,\ x_1,\ldots,x_n)$. The optimization is achieved via nlminb with 10^{-4} as absolute and relative tolerance and an analytically computed gradient and Hessian (the gradient is available in section S5 of the supplementary material, analytical expression of the Hessian is available upon request). As previously, the optimized function has multiple local optimums. Thus, an appropriate set of initial values is essential for a reliable optimization. We use as initial values for the parameters μ^*, Γ^* and Ψ^* their estimations by maximum likelihood under the normality assumption of latent variables. Initial values for the φ parameters are taken through the grid of initial values constructed by the R command cover.design (Furrer *et al.*, 2009) in the space $[-\pi/2-\pi/10, \pi/2+\pi/10]^{d-1}$. The number of initial values depends on $d-1$ and is defined empirically (we stop to increase the number of initial values if the best value of the optimized function does not change after few successive increments). In our experience, this approach is faster, and more reliable than the genetic optimization algorithm (with a very large population size) implemented in the package rgenoud (Mebane & Sekhon, 2011) or any other optimization method implemented in the R command optim (R Development Core Team, 2011).

The SNP-GLLVM estimation is computationally expensive. The computational time depends strongly on the implementation of the objective function (implementation in C called from R, as for analysis in section 5.2 is faster than a pure R implementation), on the speed of the implemented optimizer, on the initial values and their 'quality', that is on how close they are to the local optimum (when this optimum does exist), on the number of assumed latent variables, on the sample size and, of course, on the computational resources. For example, using our R implementation on a relatively old Acer TravelMate 3270 laptop running under Windows Vista (with 1.66 GHz Intel Duo Core CPU T5500 and 2 GB RAM) the

computational time to estimate SNP2-GLLVM (i.e. $L=2$) with one latent variable on a simulated sample of size 500 with three conditionally normal and one conditionally Bernoulli manifest variable is on average 30 minutes when the true latent variable distribution is a mixture of normals and we use a set of seven arbitrary initial values. With all other things kept equal this computational time drops to 15 minutes when the true latent variable distribution is SNP2. Our R implementations of the SNP2-GLLVM estimations with one latent variable for the mixture of conditionally normal and Bernoulli manifest variables, and for two latent variables for the conditionally normal manifest variables are available in the supplementary material (SNP0-GLLVM and SNP1-GLLVM are particular cases of SNP2-GLLVM estimation).

Other types of manifest variables can be considered in addition to normal and Bernoulli. For example, in section 5.2 we assume one of the manifest variables in the real data set to be conditionally Poisson distributed. In this case, without loss of generality, we can assume that among p manifest variables the first p_1+p_2 variables x_j for $j=1,\ldots,p_1$ and $j=p_1,\ldots,p_2$ are respectively normally and Bernoulli distributed with probability density and probability mass function described in (14); and the last x_j for $j=p_2,\ldots,p$ are conditionally Poisson distributed manifest variables with probability mass functions

$$g_j(x_j \mid z)=\exp\{x_j(\mu_j+\gamma_j^T z)-\exp(\mu_j+\gamma_j^T z)-\log x_j!\}. \tag{15}$$

Plugging these probability density and probability mass functions in (1) and (2) we can easily obtain the corresponding GLLVM likelihood function.

3.3. Tuning the flexibility of the SNP density

The flexibility of the SNP density is controlled by the degree L of the polynomial $P_L(z)$ in (6). Different possibilities have been explored to choose L: the original work by Gallant & Nychka (1987) proposed to fix L by a deterministic rule $L=n^\alpha, 0<\alpha<1$. Davidian & Gallant (1993) and Fenton & Gallant (1996) explored under different settings whether an adaptive rule for the choice of L can be applied. Following these authors we select L on the basis of one of the information criteria taking the form $-\ell(\mu,\Gamma,\Psi,\varphi \mid L,x_1,\ldots,x_n)+C(n)k$, where k is the number of unconstrained parameters in the model with fixed L and $C(n)=1$ for the Akaike Information Criterion (AIC, Akaike, 1974), $C(n)=0.5\log n$ for the Schwarz Information Criterion (BIC, Schwarz, 1978), and $C(n)=\log\log n$ for the Hannan–Quinn criterion (HQ, Hannan, 1987).

For linear mixed effects models with focus on latent variables inference Vaida & Blanchard (2005) proposed a conditional Akaike information criterion (cAIC) based on the effective degrees of freedom introduced by Hodges & Sargent (2001). Lu *et al.* (2007) suggested to approximate the number of effective degrees of freedom in generalized linear hierarchical models with the Laplace approximation of the integral in the likelihood function. Multiple maxima under the integral in (14) makes the method of Lu *et al.* (2007) not applicable to SNP-GLLVM.

The likelihood ratio tests for testing the hypothesis $L=0$ (i.e. $\varphi_1=\pi/2$) and $L=1$ (i.e. $\varphi_2=\pi/2$) suffer from the irregularity conditions discussed by Drton (2009) amplified by the fact that $\varphi_1=\pi/2$, $\varphi_2=\pi/2$ are also boundary points. Resampling techniques for obtaining the likelihood ratio statistic distribution could help, but this is beyond the scope of this paper.

Given that the above alternatives are not applicable to our setting, we restrict ourselves to the use of AIC, BIC and HQ for choosing L. When the exact log-likelihood function $\ell(\mu,\Gamma,\Psi,\varphi \mid L,x_1,\ldots,x_n)$ cannot be computed, as for the case of the Bernoulli distribution, we approximate it by $\tilde{\ell}(\mu,\Gamma,\Psi,\varphi \mid L,x_1,\ldots,x_n)$ as described in section 3.2.

4. Monte Carlo simulations

We explore the performance of the proposed method on finite samples by simulating 600 samples of size 500 issued from the GLLVM with four manifest variables and one latent: three manifest conditionally normal and one conditionally Bernoulli. The univariate scores of the latent variable are issued from three different distributions: (i) symmetric unimodal normal $N(0,1)$; (ii) asymmetric trimodal SNP distribution $h(z)=(-\cos 0.7/\sqrt{2}+z\sin 0.7+z^2\cos 0.7/\sqrt{2})^2\phi(z)$; and (iii) asymmetric bimodal mixture of normals $0.9N(2,1)+0.1N(-2,0.25)$. For simulating from the SNP distribution we use the algorithm proposed by Gallant & Tauchen (1992).

For each simulated data set we estimate the coefficients of the SNP-GLLVM by the methodology of sections 2 and 3 for $L=2$ (SNP2), $L=1$ (SNP1) and $L=0$ (SNP0). The latter corresponds to the traditional maximum likelihood estimation under the normality assumption of the latent variable. Theoretically, as proved by Gallant & Nychka (1987), the parameters μ, Γ and Ψ are estimated consistently if L is sufficiently large and together with φ generate a density (6) close enough to the true one. In practice, as illustrated later in this section, an unduly large value of L can result in overfitting and bias because of the integral approximation, while $L=1$ or 2 are usually sufficient to detect the departure from normality of the considered latent variable densities.

To make the simulation results comparable we impose the latent variable to have variance equal to the true one, that is: (i) $\mathrm{var}(Z)=1$ for the normal density; (ii) $\mathrm{var}(Z)=2.228$ for the SNP density; and (iii) $\mathrm{var}(Z)=4.135$ for the mixture of normals density. These choices give a slight advantage to the SNP0 estimator. We use grids of six initial values for SNP1 optimization and of 13 initial values for SNP2 (constructed as discussed in section 3.2).

We compute the AIC, BIC and HQ information criteria discussed in section 3.3 for SNP2, SNP1 and SNP0 estimations on each data set. In Table 1 we report detailed simulation results for normal and SNP2 generated latent variables. For the normal latent the estimates are all nearly unbiased, despite three trimodal estimated SNP2 densities not selected by any information criterion (all other estimated SNP1 and SNP2 densities are unimodal and nearly symmetric). When the true latent variable distribution is SNP2 all parameters estimates corresponding to the conditionally normal manifest variables (i.e. $\mu_1, \mu_2, \mu_3, \gamma_1, \gamma_2$ and γ_3) are nearly unbiased. The fact that the conditionally normal observable variables are not sensitive to the wrong specification of the latent variable distribution has been already observed in the literature theoretically by Anderson & Amemiya (1988) and in simulations by Ma & Genton (2010). However, the SNP0 and SNP1 estimates of the parameters related to the loading of the Bernoulli manifest variable (γ_4) present biases though not large. The bias is clear for the SNP0 estimates, when the assumed latent variable density is far from the true one, and diminishes when the estimated density gets closer to the true one. It is surprising that the bias in the SNP2 estimates of μ_4 is larger than the SNP1 bias and almost equal to the SNP0 bias for the same parameter. A closer inspection of estimates shows that the medians of the μ_4 estimates for both SNP1 and SNP2 are exactly at the true value and the biases of the mean are both due to a few (fewer for SNP1) extreme values in μ_4 estimates. The median of the μ_4 estimates by SNP0 is equal to the mean despite the presence of a few extreme values. Similarly, a few extreme values are found when inspecting the SNP1 and SNP2 estimates of γ_4 (the median of γ_4 estimates by SNP1 is equal to the mean while the median of γ_4 estimates by SNP2 is equal to the true value). This phenomenon illustrates the bias induced by the approximate integration as discussed by Pinheiro & Chao (2006). It seems logical that this bias is larger for the integrand with greater amount of difficult spots. The integral approximation bias is an additional reason for a moderate use of high L values in our implementation.

Table 1. *Simulation results on* 600 *samples of size* 500. *The true latent variable distribution is normal and asymmetric trimodal semi-nonparametric (SNP2) distribution* $h(z)=(-\cos 0.7/\sqrt{2}+z\sin 0.7+z^2\cos 0.7/\sqrt{2})^2\phi(z)$. *Mean parameters are denoted by* $(\mu_1,\mu_2,\mu_3,\mu_4)^T$, *loadings by* $(\gamma_1,\gamma_2,\gamma_3,\gamma_4)^T$ *and uniquenesses by* $(\psi_1,\psi_2,\psi_3)^T$ *with true values in parentheses. AVE MC, average of estimates; SD, standard deviation; SE, average standard errors estimated by sandwich covariance matrix*

	SNP0			SNP1			SNP2		
	AVE MC	SD	SE	AVE MC	SD	SE	AVE MC	SD	SE
NORMAL LATENT									
$\mu_1(0)$	0.06	0.08	0.08	0.00	0.08	0.08	0.00	0.08	0.08
$\mu_2(0)$	0.07	0.09	0.08	0.00	0.09	0.08	0.00	0.09	0.08
$\mu_3(0)$	0.06	0.08	0.08	0.00	0.08	0.08	0.00	0.08	0.08
$\mu_4(0.7)$	0.70	0.16	0.16	0.70	0.16	0.17	0.70	0.16	0.16
$\gamma_1(1.4)$	1.40	0.07	0.07	1.40	0.07	0.07	1.40	0.07	0.07
$\gamma_2(1.6)$	1.60	0.07	0.07	1.60	0.07	0.07	1.60	0.07	0.07
$\gamma_3(1.4)$	1.40	0.07	0.07	1.40	0.07	0.07	1.40	0.07	0.07
$\gamma_4(2)$	2.02	0.23	0.24	2.02	0.23	0.25	2.01	0.23	0.24
$\psi_1(1)$	1.00	0.09	0.09	1.00	0.09	0.09	1.00	0.09	0.09
$\psi_2(1)$	1.00	0.11	0.10	1.00	0.11	0.10	1.00	0.11	0.10
$\psi_3(1)$	1.00	0.09	0.09	1.00	0.09	0.09	1.00	0.09	0.09
AIC preferred 79.2% of time SNP0; 9.8% of time SNP1 and 1% of time SNP2									
BIC preferred 99.7% of time SNP0; 0.15% of time SNP1 and 0.15% of time SNP2									
HQ preferred 95.5% of time SNP0; 2.5% of time SNP1 and 2% of time SNP2									
SNP2 LATENT									
μ_1 (1.95)	1.96	0.10	0.10	1.96	0.10	0.10	1.96	0.10	0.10
μ_2 (2.22)	2.23	0.12	0.12	2.23	0.12	0.12	2.23	0.12	0.12
μ_2 (1.95)	1.95	0.10	0.10	1.95	0.10	0.10	1.95	0.10	0.10
μ_4 (3.49)	3.62	0.43	0.46	3.56	0.45	0.44	3.59	0.48	0.48
$\gamma_1(1.4)$	1.40	0.08	0.08	1.39	0.06	0.09	1.40	0.06	0.09
$\gamma_2(1.6)$	1.60	0.08	0.09	1.59	0.07	0.10	1.60	0.07	0.10
$\gamma_3(1.4)$	1.39	0.08	0.08	1.39	0.06	0.09	1.39	0.06	0.09
$\gamma_4(2)$	2.15	0.36	0.42	1.90	0.25	0.35	2.04	0.30	0.41
$\psi_1(1)$	0.98	0.09	0.09	0.98	0.08	0.08	0.98	0.08	0.08
$\psi_2(1)$	1.00	0.11	0.11	0.99	0.10	0.09	0.99	0.10	0.09
$\psi_3(1)$	1.00	0.09	0.09	1.00	0.09	0.08	1.00	0.09	0.08
AIC, BIC and HQ preferred 100% of time SNP2									

Table 2 summarizes the simulation results for the mixture of normal latent variable. As previously, SNP0 estimates of parameters corresponding to the continuous manifest variables are all nearly unbiased while the SNP0 estimates related to the Bernoulli manifest variable present biases. We conclude that for GLLVM with discrete manifest variables the wrong specification of the latent variable distribution induces a bias in the estimation. Similar conclusions can be found in Ma & Genton (2010). The differences in μ_4 and γ_4 estimates confirm the integral approximation bias. The table with AIC, BIC and HQ selected estimation results for mixture of normal latent variable are available in Table S1 in the supplementary material. For normal and SNP2 latent variables a similar table is not of interest because AIC, BIC and HQ almost always choose the right model.

From the results of Tables 1 and 2 we infer that AIC prefers models with larger L, BIC with smaller L and HQ choices are intermediate. Our conclusions meet those by Zhang & Davidian (2001) and Chen *et al.* (2002).

The advantage of using the proposed method can be appreciated when looking at the shape of estimated densities in Fig. 1. The figure illustrates that the SNP1 and SNP2 specifications selected by HQ capture the main features of the true density of the latent variable.

Table 2. *Simulation results on* 600 *samples of size* 500. *The true latent variable distribution is asymmetric mixture of normals* $0.9N(2,1)+0.1N(-2,0.25)$. *Mean parameters are denoted by* $(\mu_1,\mu_2,\mu_3,\mu_4)^T$, *loadings by* $(\gamma_1,\gamma_2,\gamma_3,\gamma_4)^T$ *and uniquenesses by* $(\psi_1,\psi_2,\psi_3)^T$ *with true values in parentheses. AVE MC, average of estimates; SD, standard deviation; SE, average standard errors estimated by sandwich covariance matrix*

	SNP0			SNP1			SNP2		
	AVE MC	SD	SE	AVE MC	SD	SE	AVE MC	SD	SE
μ_1 (2.24)	2.25	0.10	0.11	2.25	0.10	0.11	2.25	0.10	0.11
μ_2 (2.57)	2.57	0.12	0.12	2.57	0.12	0.12	2.57	0.12	0.12
μ_3 (2.24)	2.25	0.11	0.11	2.25	0.11	0.11	2.25	0.11	0.11
μ_4 (3.90)	4.36	0.56	0.58	3.91	0.43	0.44	3.95	0.46	0.65
γ_1(1.4)	1.39	0.08	0.09	1.39	0.06	0.10	1.39	0.06	0.10
γ_2(1.6)	1.59	0.09	0.10	1.59	0.07	0.11	1.59	0.07	0.11
γ_3(1.4)	1.40	0.08	0.09	1.39	0.06	0.10	1.39	0.06	0.10
γ_4(2)	2.22	0.40	0.53	2.00	0.29	0.45	2.05	0.29	0.63
ψ_1(1)	0.99	0.09	0.09	1.00	0.08	0.09	1.00	0.09	0.09
ψ_2(1)	0.99	0.11	0.11	1.01	0.11	0.10	1.01	0.11	0.10
ψ_3(1)	0.99	0.09	0.09	1.00	0.09	0.09	1.00	0.09	0.09

AIC preferred 0% of time SNP0; 11.4% of time SNP1 and 88.6% of time SNP2
BIC preferred 0% of time SNP0; 69.7% of time SNP1 and 30.3% of time SNP2
HQ preferred 0% of time SNP0; 33.4% of time SNP1 and 66.6% of time SNP2

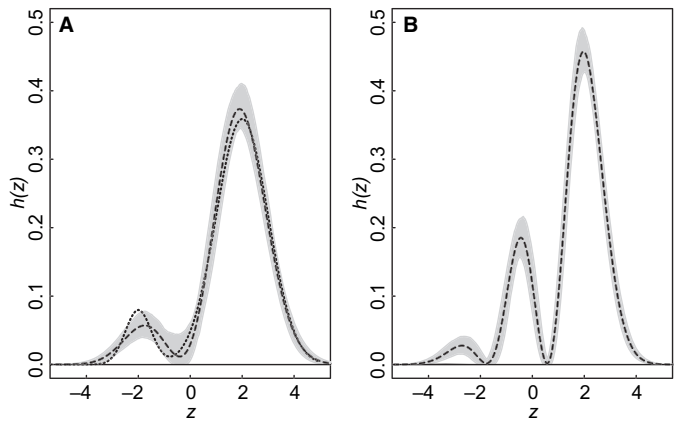

Fig. 1. The dashed line is the average of estimated densities for fits preferred by Hannan-Quinn Criterion, the shaded area is the pointwise estimated confidence envelope, the dotted line is the true density for (A) mixture of normals $0.9N(2,1)+0.1N(-2,0.25)$, (B) semi-nonparametric (SNP2) density $h(z)=(-\cos 0.7/\sqrt{2}+z\sin 0.7+z^2\cos 0.7/\sqrt{2})^2\phi(z)$ (dashed and dotted lines coincide for SNP2).

5. Data analyses

5.1. Swineford–Holzinger data analysis

This data set is a subset of the widely cited data set introduced by Holzinger & Swineford (1939) and contains scores of nine psychological tests for 145 individuals. These nine tests are: 'visual perception', 'cubes', 'lozenges', 'paragraph comprehension', 'sentence completion', 'word meaning', 'addition', 'counting group of dots' and 'straight-curved capitals'. The first three tests are usually supposed to measure the spatial ability, the second three the verbal

Table 3. *Information criteria divided by* $n = 145$ *(smaller values are preferred) for the Swineford–Holzinger data*

Model	AIC/145	BIC/145	HQ/145
$q = 2, L = 0$	11.46	11.82	11.61
$q = 2, L = 1$	11.43	11.81	11.58
$q = 2, L = 2$	11.78	12.19	11.95
$q = 3, L = 0$	11.34	11.80	11.53
$q = 3, L = 1$	11.31	11.77	11.50
$q = 3, L = 2$	11.35	11.87	11.56

ability and the last three mental speed. This data set is used in Jöreskog (1969), Jöreskog & Sörbom (1993) and as an illustration of non-normal data in LISREL 8.80.

We assume all manifest variables to be conditionally normal given the latent and fit the standardized data with the SNP0, SNP1 and SNP2 estimation methods for two ($q = 2$) and three ($q = 3$) latent variables. The loadings and uniquenesses estimates by SNP0 were used as initial values for SNP1 and SNP2 estimations. The estimated information criteria for different models are reported in Table 3.

All computed criteria chose the SNP1 model with three latent variables ($q = 3, L = 1$) confirming the conclusions of Jöreskog (1969) that three latent variables is a reasonable assumption for this data set. As expected the estimates of loadings and uniquenesses by SNP0, SNP1 and SNP2 are very close (the detailed estimation results are available in Table S2 in the supplementary material).

We assess the estimated latent variables density by plotting contours of its bivariate marginal densities in Fig. 2. The visualized estimated densities suggest the presence of two groups for the three latent variables solution and illustrate the inadequacy of a trivariate normal latent variable on this data.

5.2. Swiss consumption data analysis

This data set contains $n = 9960$ observations and is part of the Swiss consumption survey data collected in 1990. Continuous variables are food, clothing, leisure – natural logarithm of household expenses in food, clothing and leisure, respectively (assumed conditionally normal given the latent); three binary variables, dishwasher, car and motorcycle indicate the presence of a dishwasher, a car or a motorcycle in the household; finally the count variable bicycle (assumed conditionally Poisson) indicates the number of bicycles in the household. These variables are supposed to represent the latent variable 'financial wealth of the household' in its different realizations. Different subsets of this data set were analysed by Ma & Genton (2010), Huber *et al.* (2004) and Moustaki & Victoria-Feser (2006). The exploratory analysis of the data set in Fig. 3 reveals the asymmetry of the log-expenses distributions (despite the log transformation). This suggests the presence of particular groups of consumers in Switzerland, for example, consumers with extremely high food and leisure expenditures, possessing dishwasher, car and at least two bicycles.

We fitted an SNP-GLLVM with one ($q = 1$) and two ($q = 2$) latent variables to this data set. The integral involved in the log-likelihood function when $q = 2$ was approximated with the adaptive multidimensional integration algorithm based on Genz & Malik (1980) and implemented in the R-package 'cubature' by Steven G. Johnson (Johnson & Narasimhan, 2009). This package is suited for cases with at most $q = 6$ latent variables. The integral involved in the log-likelihood function when $q = 1$ is approximated as described in section 3.2 and implemented in section 4.

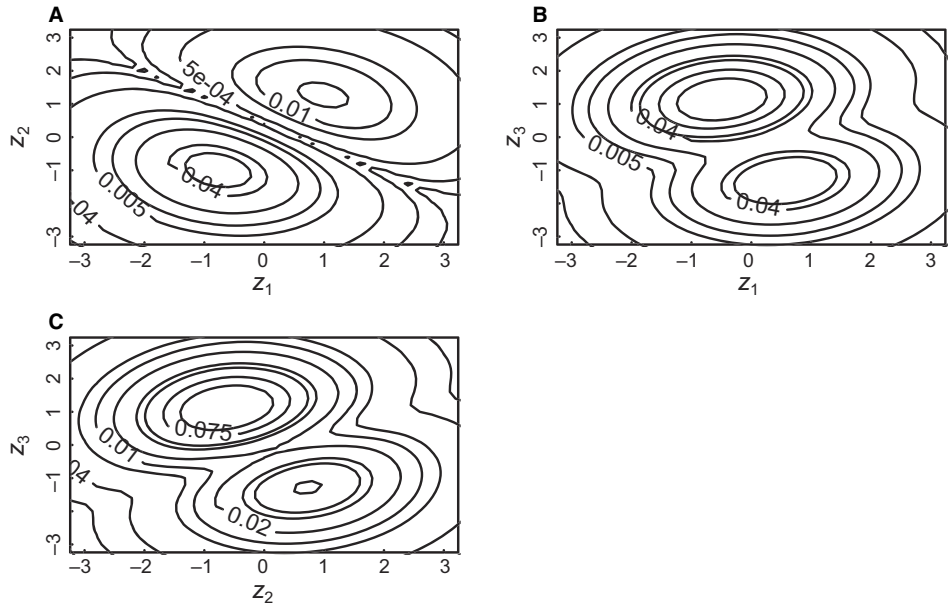

Fig. 2. Estimation of the trivariate semi-nonparametric (SNP1) latent variable on Swineford–Holzinger data: contours of its marginal bivariate densities.

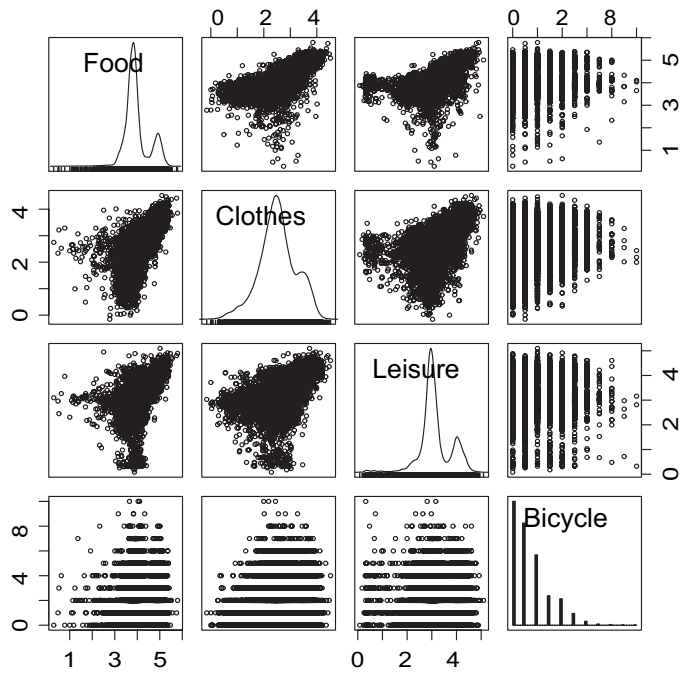

Fig. 3. Scatter plots of the Swiss consumption survey data set.

Table 4 presents the values of the estimated information criteria AIC, BIC and HQ for different models. All three information criteria choose the SNP2 model with $q=2$ as the best fit.

Table 4. *Information criteria divided by* $n=9960$ *(smaller values are preferred) and estimated coefficients with estimated standard errors for the SNP-GLLVM model with two latent variables fitted to Swiss consumption data. Mean parameters are denoted by* $(\mu_{\text{food}}, \mu_{\text{clothing}}, \mu_{\text{leisure}}, \mu_{\text{dishwasher}}, \mu_{\text{car}}, \mu_{\text{motorcycle}}, \mu_{\text{bicycle}})^T$*, loadings for ith latent variable by* $(\gamma_{i,\text{food}}, \gamma_{i,\text{clothing}}, \gamma_{i,\text{leisure}}, \gamma_{i,\text{dishwasher}}, \gamma_{i,\text{car}}, \gamma_{i,\text{motorcycle}}, \gamma_{i,\text{bicycle}})^T$ *and uniquenesses by* $(\psi_{\text{food}}, \psi_{\text{clothing}}, \psi_{\text{leisure}})^T$

Model	AIC/9960	BIC/9960	HQ/9960
$q=1, L=0$	5.60	5.61	5.60
$q=1, L=1$	5.41	5.42	5.42
$q=1, L=2$	5.38	5.39	5.38
$q=2, L=0$	5.52	5.53	5.53
$q=2, L=1$	5.36	5.36	5.36
$q=2, L=2$	5.28	5.29	5.28

	SNP0		SNP1		SNP2	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
μ_{food}	4.18	0.0039	4.15	0.0098	3.46	0.0043
μ_{clothing}	2.76	0.0051	2.76	0.0072	2.01	0.0052
μ_{leisure}	3.38	0.0056	3.22	0.0059	2.61	0.0053
$\mu_{\text{dishwasher}}$	−0.48	0.0028	−0.56	0.0105	−0.59	0.0201
μ_{car}	1.76	0.0075	3.18	0.0101	1.84	0.0162
$\mu_{\text{motorcycle}}$	−1.24	0.0071	−0.41	0.0209	−1.02	0.0501
μ_{bicycle}	−0.05	0.0002	0.96	0.0014	0.33	0.0071
$\gamma_{1,\text{food}}$	0.33	0.0016	0.43	0.0125	0.86	0.0018
$\gamma_{1,\text{clothing}}$	0.42	0.0033	0.55	0.0244	0.60	0.0261
$\gamma_{1,\text{leisure}}$	0.17	0.0020	0.25	0.0053	0.23	0.0071
$\gamma_{1,\text{dishwasher}}$	1.58	0.0041	2.57	0.0080	5.02	0.0103
$\gamma_{1,\text{car}}$	2.10	0.0010	3.00	0.0024	5.23	0.0136
$\gamma_{1,\text{motorcycle}}$	1.21	0.0088	1.60	0.0091	1.83	0.0066
$\gamma_{1,\text{bicycle}}$	1.59	0.0081	1.67	0.0460	3.03	0.0085
$\gamma_{2,\text{food}}$	−0.95	0.0019	−1.03	0.0072	−1.99	0.0043
$\gamma_{2,\text{clothing}}$	−1.06	0.0090	−1.18	0.0390	−2.22	0.0514
$\gamma_{2,\text{leisure}}$	−1.00	0.0037	−1.11	0.0076	−1.87	0.0052
$\gamma_{2,\text{dishwasher}}$	−0.43	0.0095	−0.31	0.0121	−2.54	0.0287
$\gamma_{2,\text{car}}$	−0.86	0.0018	−0.62	0.0027	−2.17	0.0042
$\gamma_{2,\text{motorcycle}}$	0.26	0.0072	0.20	0.0201	−0.66	0.0281
$\gamma_{2,\text{bicycle}}$	0.16	0.0009	−0.04	0.0007	−1.08	0.0007
ψ_{food}	0.07	0.0006	0.07	0.0015	0.06	0.0016
ψ_{clothing}	0.20	0.0008	0.21	0.0035	0.22	0.0038
ψ_{leisure}	0.19	0.0013	0.18	0.0013	0.18	0.0018

Estimated uniquenesses and loadings for the model with two latent variables with covariance matrix standardized to be the identity matrix can be seen in Table 4 (the corresponding results for $q=1$ is available in Table S3 in the supplementary material). An orthogonal rotation of the loadings could suggest a different interpretation but the pattern from unrotated loadings is the most intuitive. Also the SNP2 estimation is very dissimilar with SNP0, and this dissimilarity is increased by orthogonal rotation (varimax). For interpretation, as suggested by Moustaki & Knott (2000), loadings should be standardized because normal, binary and Poisson manifest variables are measured on different scales. Taking this into account for all three estimators, the first latent variable measures mostly what the household possesses already (with car being the most important measure for SNP0 and SNP1, for SNP2 this importance is shared between car and dishwasher), while the second latent variable measures the current expenses of the household with opposite sign. The clothing expense is the most important component in the second latent for all three estimators.

We present the estimated bivariate SNP2 density on Swiss consumption data in Fig. 4 by showing marginal densities and contours of the estimated bivariate density (a figure of

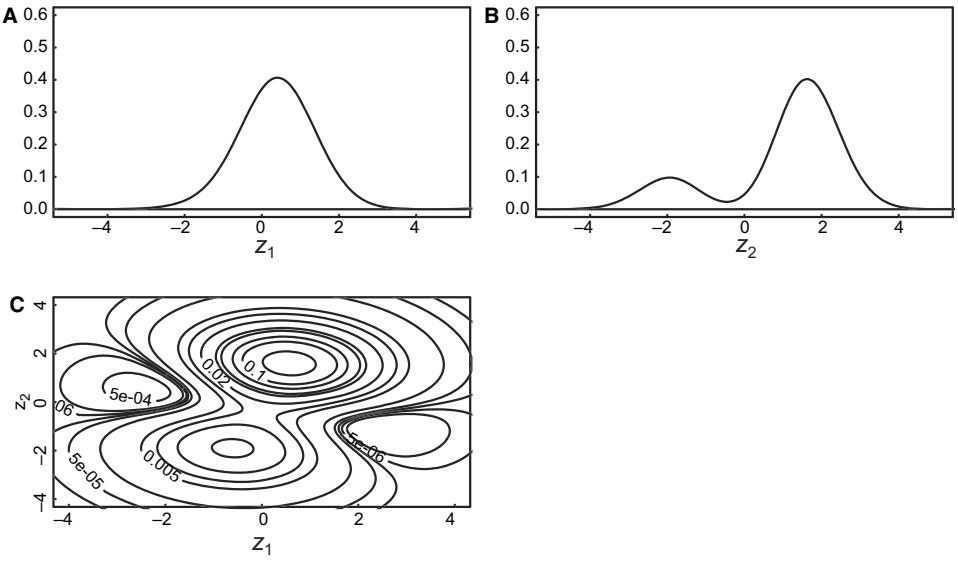

Fig. 4. Estimated semi-nonparametric (SNP2) density of two latent variables for the SNP-GLLVM fitted to Swiss Consumption data. (A) and (B) marginal densities of the latent variables; (C) contours of the joint bivariate density.

the estimated density when $q = 1$ can be found in the supplementary material, Fig. S1). The SNP2 estimated density clearly detects the presence of four unequal groups of households: an important group of households with low expenditures, a less important group with high expenditures (we interpreted by taking into account the opposite sign of loadings) but not differentiated clearly by belongings and two tiny groups of households with almost similar expenses but opposite in belongings. The latent variable related with belongings is slightly skewed to the right, while the latent variable related with expenses is clearly bimodal.

6. Discussion

We considered GLLVM with flexible distribution of latent variables specified by smooth densities from the SNP approach of Gallant & Nychka (1987). The proposed method was shown to estimate the true density of latent variables with a good degree of accuracy. The estimated density is easy to visualize. We established by simulations that the GLVVM with a combination of binary and normal observable variables is sensitive to the wrong specification of the true latent variables distribution. Our experience makes us believe that this holds with other discrete distributions for the manifest variables. One may argue that our SNP-GLLVM does not gain much over normal modelling when there is no discrete manifest variable, however exactly in this case the estimation and visualization easiness give an important insight into the behaviour of the assumed latent variables. This is why we expect the proposed method to be useful to the community of latent variables users and researchers.

Supporting Information

Additional Supporting Information for this article may be found in the online version of this article.

S1. Simulation results selected with AIC, BIC and HQ criteria for the mixture of normal latent variable as mentioned in section 4.

S2. Detailed estimation results for Swineford–Holzinger data in section 5.1.
S3. SNP-GLLVM with one latent variable fitted to Swiss consumption data from section 5.2: coefficients with standard errors and estimated densities of latent variable.
S4. Gradient of the log-likelihood function for SNP-GLLVM with conditionally normal manifest variables in section 3.1.
S5. Gradient of the log-likelihood function for SNP-GLLVM with mixed scale manifest variables in section 3.2.
S6. Proof of Proposition 2.1 and Corollary 2.2 from section 2.2.
S7. R code for estimating SNP-GLLVM with mixed scale manifest variables and one latent variable as mentioned in section 3.2.
S8. R code for estimating SNP-GLLVM with conditionally normal manifest variables and two latent variables as mentioned in section 3.2.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.
Anderson, T. & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16**, 759–771
Bartholomew, D. J. (1980). Factor analysis for categorical data. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **42**, 293–321.
Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika* **71**, 221–232.
Bartholomew, D. J. (1988). The sensitivity of latent trait analysis to choice of prior distribution. *British J. Math. Statist. Psych.* **41**, 101–107.
Cagnone, S., Moustaki, I. & Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal respondes. *British J. Math. Statist. Psych.* **62**, 401–415.
Chen, J., Zhang, D. & Davidian, M. (2002). A Monte-Carlo EM algorithm for generalized mixed models with flexible random effects distribution. *Biostatistics* **3**, 347–360.
Davidian, M. & Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–488.
De Bruijn, N. G. (1981). *Asymptotic methods in analysis*. Dover, New York.
Drton, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37**, 979–1012.
Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *J. Amer. Statist. Assoc.* **98**, 555–563.
Fenton, V. M. & Gallant, A. R. (1996). Qualitative and asymptotic performance of the SNP density estimators. *J. Econometrics.* **74**, 77–118.
Furrer, R., Nychka, D. & Sain, S. (2009). Package fields [Computer software manual]. Vienna. Austria. Available on http://www.R-project.org
Gallant, A. R. & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**, 363–390.
Gallant, A. R. & Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications. *Econometrica* **57**, 1091–1120.
Gallant, A. R. & Tauchen, G. (1992). A nonparametric approach to nonlinear time series analysis: estimation and simulation, part II. In *New directions in time series analysis* (eds D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt & M. S. Taqqu), Vol. **45**, 71–92. Springer-Verlag, New York.
Genz, A. & Malik, A. (1980). An adaptive algorithm for numerical integration over an *n*-dimensional rectangular region. *Int. J. Comput. Appl. Math.* **6**, 295–302.
Hannan, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2**, 135–151.
Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*. Springer.
Hodges, J. S. & Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parametrised models. *Biometrika* **88**, 367–379.

Holzinger, K. & Swineford, F. (1939). A study in factor analysis: the stability of a bi-factor solution. *Supplementary Educational Monographs* **48**, xi–91.

Huang, X., Stefanski, L. A. & Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93**, 53–64.

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

Jennrich, R. I. & Thayer, D. T. (1973). A note on Lawley's formulas for standard errors in maximum likelihood factor analysis. *Psychometrika* **38**, 571–580.

Johnson, S. G. & Narasimhan, B. (2009). Cubature: adaptive multivariate integration over hypercubes [Computer software manual].

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

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202.

Jöreskog, K. & Sörbom, D. (1993). *LISREL 8: structural equation modeling with the SIMPLIS command language*. Scientific Software International, Lincolnwood.

Knott, M. & Tzamourani, P. (2007). Bootstrapping the estimated latent distribution of the two-parameter latent trait model. *British J. Math. Statist. Psych.* **60**, 175–191.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73**, 805–811.

Lawley, D. (1967). Some new results in maximum likelihood lactor analysis. *Proceedings of the Royal Society of Edinburgh* **A67**, 256–264.

Liu, X., Wall, M. M. & Hodges, J. S. (2005). Generalized spatial structural equation models. *Biostatistics* **6**, 539–557.

Lu, H., Hodges, J. S. & Carlin, B. P. (2007). Measuring the complexity of generalized hierarchical models. *Canad. J. Statist.* **35**, 69–87.

Ma, Y. & Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scand. J. Statist.* **31**, 459–468.

Ma, Y. & Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **72**, 475–495.

Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate analysis*. Academic Press, London.

Mebane, W. R. J. & Sekhon, J. S. (2011). Genetic optimization using derivatives: the rgenoud Package for R. *J. Stat. Software* **42**, 1–26.

Montanari, A. & Viroli, C. (2010a). Heteroscedastic factor mixture analysis. *Statist. Model.* **10**, 441–460.

Montanari, A. & Viroli, C. (2010b). A skew-normal factor model for the analysis of student satisfaction towards university courses. *J. Appl. Stat.* **43**, 473–487.

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

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

Piessens, R., Doncker-Kapenga, E. de & Überhuber, C. W. (1983). *Quadpack a subroutine package for automatic integration*. Springer Series in Computational Mathematics, Berlin.

Pinheiro, J. C. & Chao, E. C. (2006). Efficient Laplacian and Gaussian quadrature algorithms for multilevel generalized mixed models. *J. Comput. Graph. Statist.* **15**, 58–81.

R Development Core Team (2011). R: a language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available on http://www.R-project.org

Rabe-Hesketh, S. & Skrondal, A. (2001). Parametrization of multivariate random effects models for categorical data. *Biometrics* **57**, 1256–1264.

Rabe-Hesketh, S. & Skrondal, A. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. CRC Press, Boca Raton.

Rabe-Hesketh, S., Pickles, A. & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statist. Model.* **3**, 215–232.

Ramsey, J. (2010, May). Maximum likelihood estimation for the unconstrained factor analysis model: second order optimization using tensor analysis. In *Banff international research station for mathematical innovation and discovery, functional data analysis: future directions*.

Rizopoulos, D. & Moustaki, I. (2008). Generalized latent variable models with non-linear effect. *British J. Math. Statist. Psych.* **61**, 415–438.

Schilling, S. & Bock, R. D. (2005). High-dimensional maximum likelihood item factor analysis by adaptive quadrature. *Psychometrika* **70**, 533–555.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
Scott, D. W. (1992). *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics, New York.
Stetter, H. J. (2004). Numerical polynomial algebra. SIAM, Philadelphia.
Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.
Wall, M. M. & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *J. Amer. Statist. Assoc.* **95**, 929–940.
Wedel, M. & Kamakura, W. A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika* **66**, 515–530.
Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika* **62**, 297–330.
Zhang, D. & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**, 795–802.

Received October 2010, in final form September 2011

Irina Irincheeva, Department of Economics, 40 Bd du Pont d'Arve, CH-1211 Genève 4, Switzerland.
E-mail: iirincheeva@mail.ru