Advances in Latent Variables - Methods, Models and Applications

Brescia - Department of Economics and Management

June 19, 2013 – June 21, 2013

Electronic Book "Advances in Latent Variables"

The full version of the Conference Proceedings, collected in the **Electronic Book "Advances in Latent Variables"**, Eds Brentari E., Carpita M., Vita e Pensiero, Milan, Italy, ISBN 978 88 343 2556 8, are available here.

SOLI-A1.5 - Advances in longitudinal data analysis

Sparse Nonparametric Graphical Models for Random Effect Distribution in GLMMs

S. Viviani, M. Alfó and P. Brutti

Abstract A generalized linear mixed model with a nonparametric distribution for the random effect is proposed. The normality assumption for the random effects may be too restrictive to represent the between-subject distribution, especially when the longitudinal response is non-Gaussian. Starting from nonparametric graphical models, we take advantage of the nonparanormal approach to build a flexible latent, individual-specific structure for the longitudinal profiles. The nonparanormal method is particularly appealing since it acts on transformations of multivariate non-Gaussian random variables, and assumes that these transformations are multivariate Gaussian. Moreover, it is particularly convenient to handle the joint distribution for high-dimensional variables.

Key words: Generalized linear mixed models, Graphical models, Random effect distribution, Non-parametric approach

1 Introduction

In longitudinal studies, the pattern of change with respect to time of a non-Gaussian outcome of interest is often accounted for through generalized linear mixed model, see for instance [9] and [1]. This model postulates a linear relationship between a given link function of the response expected value and some covariates with associated fixed and random effects. A natural heterogeneity, deriving either from un-

S. Viviani
Department of Statistics, Sapienza University of Rome, e-mail: sara.viviani@uniroma1.it

M. Alfó
Department of Statistics, Sapienza University of Rome, e-mail: marco.alfo@uniroma1.it

P. Brutti
Department of Statistics, Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

observed characteristics or varying effects of measured covariates among observed subjects, is considered through the introduction of individual-specific latent effects, and may includes genetic or environmental factors. Standard theory assume that the random effects are normally distributed.

While inference on the fixed effects has been found to be robust to misspecification of the random effect distribution, especially when the number of measurement per individual is high enough, see for instance [2] and [17], the choice of an appropriate random effect density seems to be relevant for what concerns efficiency and (unbiased) standard error estimation, see [18].

We propose a class of generalized linear mixed models with nonparametric random effects, to allow for more flexible distributional assumptions on between-subjects variability. In literature, relevant contributions in this field are, among others, [8], where the nonparametric maximum likelihood estimate (NPMLE) is defined through a discrete random effect distribution, [12] with smoothed nonparametric ML estimator, [19], where a semiparametric method is proposed, and [4] with P-spline based random effect distribution. Our approach is different and it is based on one family of nonparametric graphical model, referred to as the *nonparanormal* distributions. The nonparanormal can be seen as an extension of additive models for regression to graphical modeling. Flexibility is introduced by working on the multivariate Gaussian transformation $f(Y)$ of the non-Gaussian random variable $Y = (Y_1, \dots, Y_d)$. This approach can be linked to Gaussian copulas, see [14], when the marginal distributions are fully nonparametric. A detailed overview on graphical models is [7], where the nonparanormal and the forest density families approaches are compared. Essentially, these families are two different ways of representing a graphical model: the nonparanormal is distribution based, while the forest density forces the graphical structure to be a tree or a forest.

In this paper, we focus only on the nonparanormal distribution, applying this concept to generalized linear mixed models. Effectively, the nonparanormal distribution has been considered for *observed* random variables, while, at our knowledge, no attempt has been done to extend it to latent random variables.

We compare the nonparanormal latent approach to the approach based on Gaussian random effects in generalized linear mixed models in different settings, highlighting the situations where the proposed approach is more convenient.

The rest of the paper is as follows. In Section 2 we introduce the class of generalized linear mixed models and discuss the role of the random effect distribution. In Section 3, the nonparanormal distribution and its application to generalized linear mixed models are reviewed in details.

2 Generalized linear mixed model

In Section 1, we stated that our aim is at proposing a flexible random effect distribution for generalized linear mixed models following the nonparanormal approach.

With this purpose, we introduce the generalized linear mixed model and discuss the role of the random effect distribution.

Let $Y_i(t_j)$ be the longitudinal outcome of interest, measured in $j = 1,\dots,r_i$ occasions for the ith subject, $i = 1,\dots,n$, and $\mathbf{x}_i(t_j)$ the corresponding p-dimensional vector of explanatory variables. Moreover, let us indicate with $m_i(t_j)$ and $v_i(t_j)$ the expected value and the variance of $Y_i(t_j)$. For a given individual i, the response sequence is a r_i-vector $y_i(t_j) = [y_i(t_1),\dots,y_i(t_{r_i})]^\mathsf{T}$, and $Cov[y_i(t_j), y_i(t_k)] = v_i(t_{jk})$.

The generalized linear mixed models, see [1], are random effect models for responses with conditional distribution in the exponential family, [9]. In these models, the sources of unobserved individual-specific heterogeneity among individuals are represented by random variability in the regression coefficients. Models with a random intercept can be written as follows:

$$u(m_i(t_j)) = (\beta_0 + b_{i0}) + \beta_1 x_{i1}(t_j) + \ldots + \beta_p x_{ip}(t_j),$$

where $m_i(t_j) = E[Y_i(t_j)|b_{i0}, \mathbf{x}_i)]$ and $u(\cdot)$ is a given link function. When a set of random regression coefficients is used, we may write $\mathbf{b}_i \sim g(\mathbf{0}, \mathbf{D})$, where $g(\cdot)$ is a proper density function which can be parametric, and the covariance matrix $\mathbf{D}$ needs to be estimated.

It is known, see among others [5], [3], [15] and [18], that the fixed effect ML estimate may not be robust to misspecification of $g(\cdot)$, especially when the number of repeated measurements per individual is not high enough. The nonparametric maximum likelihood estimate, developed among others by [5], [10] and [11], is an appealing approach to achieve ML estimate consistency. Under this approach, assuming conditional independence of repeated measures corresponding to the same individual given the random effects, the longitudinal response distribution is written as follows:

$$\begin{aligned} f(y_i(t_j)) &= \int f(y_i(t_j|\mathbf{b}_i)g(\mathbf{b}_i)d\mathbf{b}_i = \int f(y_i(t_j|\mathbf{b}_i)dG(\mathbf{b}_i) \qquad (1) \\ &\approx \sum_{l=1}^{L} f(y_i(t_j|\mathbf{b}_l)\pi_l, \end{aligned}$$

and $g(\cdot)$ is approximated by a discrete distribution π_l on $L \leq n$ support points. Although this estimating method is theoretically strong and relatively simple to implement, it may be complicated by the high dimension of the random effects. Furthermore, some authors, see e.g. the discussion in [13] and [16] criticized this approach as unrealistic and have proposed a smooth version of the nonparametric mixing distribution, see [18], [19], [4] among others. In this perspective, the nonparanormal approach may be an alternative, since the assumption of continuous random effects still holds.

3 The Nonparanormal approach

Let us consider a multivariate random coefficient vector with dimension d for the ith subject, $\mathbf{b}_i = [b_{i1}, \dots, b_{id}]^\top$, and a transformed random variable $\mathbf{h}_i = h(\mathbf{b}_i) = [h(b_{i1}), \dots, h(b_{id})]^\top$ such that $h(\mathbf{b}_i)$ is multivariate Gaussian. This transformation leads to a nonparametric extension of the normal approach called *nonparanormal* distribution. This family of distributions requires the estimate of the univariate functions $h_{ik} = h(b_{ik})$, $k = 1, \dots, d$, and the covariance matrix $\mathbf{D}$.
The nonparanormal can be seen as an extension of copulas, [14], with fully nonparametric marginals; therefore, the estimation of univariate marginals h_{ik} may be done as follows:

$$h(b_{ik}) = \mu_k + \sigma_k \Phi^{-1}(G_k(b_{ik})), \tag{2}$$

where μ_k and σ_k are the kth component mean and standard deviation, $\Phi^{-1}(\cdot)$ is the inverse of the Gaussian distribution function and G_k is the distribution function of b_{ik}. Moreover, it is assumed that $E(h_{ik}) = E(b_{ik}) = \mu_k$ and $Var(h_{ik}) = Var(b_{ik}) = \sigma_k$. Once h_{ik} is estimated, we transform $\mathbf{b}_i = [b_{i1}, \dots, b_{id}]^\top$ to multivariate Gaussian random variable $h(\mathbf{b}_i) = [h(b_{i1}), \dots, h(b_{id})]^\top$ and apply methods for Gaussian graphical models to estimate the graph. It is worth noticing that in this case, the sparsity of the model is regulated through the precision matrix $\Omega = \mathbf{D}^{-1}$.

We say that $\mathbf{b}_i$ has a nonparanormal distribution, i.e. $\mathbf{b}_i \sim NPN(\mu, \mathbf{D}, \mathbf{h})$, when there exist functions $h(\cdot)$ such that $h(\mathbf{b}_i) \sim MVN(\mu, \mathbf{D})$. If $h_{ik} = h(b_{ik})$, $k = 1, \dots, d$, is differentiable, the joint density function of $\mathbf{b}_i$ is given by

$$g(\mathbf{b}_i|D) = \frac{1}{(2\pi)^{d/2}|\mathbf{D}^{1/2}|} \exp\left\{-\frac{1}{2}(h(\mathbf{b}_i) - \mu)^\top \mathbf{D}^{-1}(h(\mathbf{b}_i) - \mu)\right\} \prod_{k=1}^{d} |h^{'}(b_{ik})|, \tag{3}$$

where $\mu = [\mu_1, \dots, \mu_k]^\top$ and $|h^{'}(b_{ik})|$ is the jacobian of h_{ik}. It can be noticed that density in expression (3) is not identifiable. To make the family identifiable, it is required that h_{ik} preserves marginal means and variances. Hence, we fix $\mu_k = 0$ and $\sigma_k = \sigma_{0k}$, $k = 1, \dots, d$, and the sparsity of the model is identified by the estimation of the random effect covariances.

3.1 Estimation

For the estimation of h_{ik} and the precision matrix Ω, we follow the procedure described in [7], with the exception that we work with latent rather than observed variables. The procedure is similar to the one adopted by [6]. We assume that $\mathbf{b}_i$ is nonparanormal with marginals following a Dirichlet process, i.e. $G_k = DP(G_0, \alpha_k)$. Here, $G_0 \sim N(0,1)$ and α_k is a (component-specific) precision parameter, measuring the displacement of G_k from G_0. In summary, we have

$$Y_i(t_j) \mid \mathbf{b}_i \sim \mathtt{EF}(\eta_i(t_j))$$

$$\begin{aligned} \eta_i(t_j) &= \beta^\top \mathbf{x}_i(t_j) + \mathbf{b}_i^\top \mathbf{z}_i(t_j) \\ \mathbf{b}_i &\sim NPN(\mu, \Omega, \mathbf{h}) \\ G_k &\sim DP(G_0, \alpha_k) \quad , \end{aligned}$$

where EF stands for a distribution which belongs to the exponential family, and $\mathbf{z}_i(t_j)$ is a row vector of covariates associated to subject-specific effects. We can approximate the marginal distribution function for the k-th dimension by the following finite sum

$$G_k = \sum_{\ell=1}^{L} \pi_\ell^k \delta_{\theta_\ell^k}.$$

see for instance [6]. By developing an approach based on *stick breaking* processes, we assume that the locations θ_ℓ^k and the weights π_ℓ^k are distributed as follows:

$$\begin{aligned} \theta_\ell^k &\sim G_0 = N(0,1) \\ \pi_\ell^k &= \nu_\ell^k \prod_{h=1}^{\ell-1} (1-\nu_\ell^k) \\ \nu_\ell^k &\sim \mathtt{Beta}(1, \alpha_k) \quad . \end{aligned}$$

The longitudinal and random effect distributions are:

$$\begin{aligned} f(\mathbf{y}_i \mid \mathbf{b}_i) &= \prod_j f(y_i(t_j) \mid \mathbf{b}_i) \quad ; \\ g(\mathbf{b}_i \mid \nu) &\propto \mathtt{exp}\left(-\frac{1}{2}(\mathbf{h}_i - \mu)^\top \Omega (\mathbf{h}_i - \mu)\right) \prod_k |h'_{ik}| \quad . \end{aligned}$$

Within this modeling framework, $h(\cdot)$ can be written for the ℓ-th location and the k-th component as

$$h_{k\ell}(\theta_\ell^k) = \mu_k + \sigma_k \Phi^{-1}\left(G_k(\theta_\ell^k)\right) = \mu_k + \sigma_k \Phi^{-1}\left(\textstyle\sum_{h=1}^{\ell} \pi_h^k \delta_{\theta_h^k}\right).$$

The complete log-likelihood is then:

$$\log\left[\mathscr{L}_{\mathtt{c}}(\cdot)\right] = \sum_i \sum_\ell z_{i\ell} \left[\log f(\mathbf{y}_i \mid \theta_\ell) + \log g(\theta_\ell \mid \nu_\ell) + \sum_k \log p(\nu_\ell^k \mid \alpha_k)\right],$$

where $z_{i\ell} = 1$ if $\mathbf{b}_i$ comes from the ℓ-th component, and $z_{i\ell} = 0$ otherwise.
Parameter estimation is performed via an EM type algorithm.

References

1. Breslow, N.E., Clayton, D.G.: Approximate Inference in Generalized Linear Mixed Models. JASA **421**, 9-25 (1993)
2. Butler, S., Louis, T.: Random effects models with nonparametric priors. Statistics in Medicine **11**, 1981–2000 (1992)
3. Davies, R. B. : Mass point method for dealing with nuisance parameters in longitudinal studies. In Crouchley, R. (ed.): Longitudinal data analysis, Aldershot: Avebury (1987)
4. Ghidey, W., Lesaffre, E., Eilers, P.: Smooth random effects distribution in a linear mixed model. Biometrics **60**, 945–953 (2004)
5. Heckman, J.J., Singer, B. : A method for minimizing the impact of distributional assumptions in econometric models of duration. Econometrica. **52**, 271–320 (1984)
6. Heinzl, F., Tutz, G. : Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. Statistical Modelling, to appear. (2013)
7. Lafferty, J. , Liu, H. , Wasserman, L.: Sparse Nonparametric Graphical Models. Statistical Science. **27**, 519-537 (2012)
8. Laird, N. M.: Nonparametric maximum likelihood estimation of a mixing distribution. JASA **73**, 805–811 (1978)
9. Laird, N. M., Ware, J. H.: Random effects models for longitudinal data. Biometrics **38**, 963–974 (1982)
10. Lindsay, B.G. : The geometry of mixture likelihoods: a general theory. Annals of Statistics, **11**, 86-94 (1983)
11. Lindsay, B.G. : The geometry of mixture likelihoods, part II. Annals of Statistics: the exponential family. Annals of Statistics, **11**, 783-792. (1983)
12. Magder, L. S., Zeger, S. L.: A smooth nonparametric estimate of a mixing distribution using mixture of Gaussians. JASA **91**, 1141–1151 (1996)
13. McCulloch, C. E., Neuhaus, J. M. : Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. Statistical Science, **26**, 388-402. (2011)
14. Nelsen, R.: An Introduction to Copulas. Springer-Verlag, New York (1999)
15. Neuhaus, J.M., Hauck, W.W., Kalbfleish, J.D. : The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika. **79**, 755–762 (1992)
16. Neuhaus, J. M., McCulloch, C. E. : A Note on Type II Error Under Random Effects Misspecification in Generalized Linear Mixed Models. Biometrics, **67**, 654656. (2011)
17. Verbeke, G., Lesaffre, E.: The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. Computational Statistics and Data Analysis **23**, 541–556 (1997)
18. Verbeke, G., Lesaffre, E.: A linear mixed model with heterogeneity in the random effects population. JASA **91**, 217–221 (1996)
19. Zhang, D., Davidian, M.: Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics **57**, 795–802 (2001)