Modeling Through Latent Variables

Geert Verbeke^{1,2} and Geert Molenberghs^{2,1}

Interuniversity Institute for Biostatistics and statistical Bioinformatics

¹KU Leuven, Leuven, Belgium, B-3000

²Universiteit Hasselt, Hasselt, Belgium, B-3590

Annual Review of Statistics and Its
Application 2016. 3:1–18

This article's doi:
10.1146/((please add article doi))

Copyright © 2016 by Annual Reviews.
All rights reserved

Corresponding author:
Geert Verbeke, I-BioStat, Kapucijnenvoer
35, B-3000 Leuven, Belgium.
E-mail:geert.verbeke@kuleuven.be

Geert Molenberghs, I-BioStat,
Martelarenlaan 42, B-3500 Hasselt,
Belgium.
E-mail:geert.molenberghs@uhasselt.be

Keywords

bridge distribution, conditional model, gradient function, latent class model, marginal model, marginalized model, mixed model, mixing distribution, mixture model, nonparametric maximum likelihood

Abstract

In this review, a general overview is given of latent variable models. The general model is introduced and various inferential approaches are discussed. Afterwards several commonly applied special cases are presented, including mixture or latent class models, as well as mixed models. Many of these models are applied to a single dataset with simple structure, allowing for easy comparison of the results. This allows to discuss advantages and disadvantages of the various approaches, but also to illustrate several problems inherently linked to models incorporating latent structures. Finally, model extensions and applications are touched upon, and a number of issues often ignored when applying latent variable models highlighted.

Contents

1. INTRODUCTION	2
2. GENERAL MODEL FORMULATION	3
3. ILLUSTRATIVE CASE STUDY: LEUVEN DIABETES STUDY	5
3.1. Mixed Models	6
3.2. Latent Class Models	8
3.3. Non-parametric Maximum Likelihood (NPML)	8
4. APPLICATIONS, EXTENSIONS, AND ISSUES	9
4.1. Modeling Associations	9
4.2. Models with Latent Variables of Mixed Types	11
4.3. Parameter Interpretation	11
4.4. Marginalized Multilevel Models and Bridge Distributions	12
4.5. Identifiability	13
5. CONCLUSION	15

1. INTRODUCTION

Many models currently used in statistical practice assume the presence of latent unobserved variables which are used to account for overdispersion or to implicitly model association structures. For example, latent class models (McCutcheon 1987) assume the presence of a discrete latent variable representing the prior belief that the population studied consists of a finite number of sub-populations in which the outcome of interest follows a different statistical model. Mixed models (Verbeke and Molenberghs 2000, Molenberghs and Verbeke 2005), on the other hand, use continuous latent variables to implicitly model association in hierarchically structured data sets. While latent variable models are extremely flexible in the sense that they can handle complex data structures and provide straightforward extensions of classical models, the use of latent structures has important ramifications, which should be well understood by the user. First, because such models are formulated conditionally upon one or multiple latent variables, parameters in those models need to be interpreted conditionally on the latent structure, rather than marginally. For generalized linear or non-linear models, this implies that model comparisons should be at the marginal (population average) level and hence should not be based on a comparison of conditional-level parameters. Second, models with latent structures are not identifiable from the data alone but require non-verifiable assumptions. Third, and related to the previous issue, latent structures are, by definition, not observed, implying that assessing distributional assumptions about those structures is far from straightforward.

The aim of this paper is to show that many models which nowadays belong to the standard toolbox of every practicing statistician, and which have been developed independently in various subfields of statistical research, all fit within the general unified framework of latent variable models. In Section 2, the model is introduced in its most general form, and some special cases are presented, including latent class models and mixed models. In Section 3, the various models are illustrated in the analysis of a single data set. To focus ideas, and to simplify the comparison of the various model families, a very simple setting of a binomial outcome without any covariates is used. Section 4 presents a number of situations where latent variable models are applied, together with extensions of the basic models as

well as a discussion about the most important issues concerning parameter interpretation, identifiability and model assessment. In Section 4.1, it is shown that latent variables not only account for overdispersion but also implicitly model associations, making them particularly useful to jointly model multiple outcomes. Flexible modeling of overdispersion and associations may require inclusion of multiple (sets of) latent variables with different distributional assumptions (Section 4.2). In Section 4.3, a logistic model with normally distributed latent variables is used to explain why many standard latent variable models do not allow for marginal interpretations of the parameters. Some solutions proposed in the statistical literature are presented in Section 4.4. Section 4.5 illustrates that multiple latent variable models may imply identical marginal models, pointing towards issues for identifiability and model checking. Finally, some concluding remarks are given in Section 5.

2. GENERAL MODEL FORMULATION

Let $\mathbf{Y}_i$, $i = 1, \dots, N$ denote a sample of independent random vectors with components Y_{ij} , $j = 1, \dots, n_i$. While in many contexts all $\mathbf{Y}_i$ are of the same size, i.e., $n_i \equiv n$, this can be generalized, for example, to allow for unequal numbers of repeated measurements for different participants in longitudinal studies. We assume that $\mathbf{Y}_i$ (possibly appropriately transformed) satisfies

$$\mathbf{Y}_i | \mathbf{b}_i, \boldsymbol{\theta} \sim F_i(\mathbf{b}_i, \boldsymbol{\theta}), \quad (1)$$

i.e., $\mathbf{Y}_i$ follows a pre-specified distribution F_i , possibly depending on covariates (suppressed from notation), and parameterized through a vector $\boldsymbol{\theta}$ of unknown parameters, common to the models for all $\mathbf{Y}_i$, and a vector $\mathbf{b}_i$ that is $\mathbf{Y}_i$ -specific. Let the corresponding density be denoted by $f_i(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\theta})$. The distribution F_i can be any n_i -dimensional distribution. In practice, however, it is often assumed that the components Y_{ij} in $\mathbf{Y}_i$ are independent, conditional on $\mathbf{b}_i$. Under this ‘conditional independence assumption’ it suffices to specify the univariate distributions of all responses Y_{ij} , and the distribution function F_i in (1) becomes a product over the n_i independent elements in $\mathbf{Y}_i$.

Once the model in (1) has been specified, an obvious approach toward estimation is based on maximizing the likelihood $\prod_i f_i(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and all $\mathbf{b}_i$. However, Neyman and Scott (1948) showed that the so-obtained ML estimates may be inconsistent, i.e., may not converge to the true unknown value for some of the parameters, with increasing values for the sample size N . This is due to the fact that the number of parameters increases with N . This is a well-known result in the context of logistic regression for matched binary data, where, in the case of an increasing number of strata, the ML estimator $\hat{\psi}$ for the odds ratio converges to ψ^2 rather than to ψ . We refer to Breslow and Day (1989, Section 7.1) for an extensive discussion in this context. This shows that the parameters $\mathbf{b}_i$ should not be treated as fixed, unknown parameters, and that procedures are required to eliminate the $\mathbf{b}_i$ from the estimation process.

A first approach, also applied in the above discussed example of a logistic model for matched data, is conditional inference (see, e.g., Diggle *et al* 2002, Section 9.3). One then considers the parameters $\mathbf{b}_i$ as nuisance, and estimation of $\boldsymbol{\theta}$ is done by maximizing the likelihood of the data $\mathbf{y}_i$, conditional on sufficient statistics for the $\mathbf{b}_i$. The main advantage of conditional inference is that no additional assumptions are needed with respect to the parameters $\mathbf{b}_i$, which are then considered nuisance. A disadvantage is clearly that sufficient statistics for the $\mathbf{b}_i$ need to be found, and that the likelihood of the data conditional on

these statistics needs to be calculated and maximized. Also, all information on the $\mathbf{b}_i$ is lost, but information may also be lost on certain elements in $\boldsymbol{\theta}$, even though they may be of primary interest. An example can be found in Verbeke, Spiessens, and Lesaffre (2001) in the context of linear models for longitudinal data.

A second approach, allowing for inferences about the $\mathbf{b}_i$, assumes that the $\mathbf{b}_i$ were sampled independently from a general population of parameters. Let $Q(\mathbf{b}_i)$ denote distribution function of $\mathbf{b}_i$, often called the mixing distribution. Because the $\mathbf{b}_i$ are not observed they are traditionally called ‘latent’ random variables. Elimination of the $\mathbf{b}_i$ is then obtained from integrating them out, over their assumed distribution. More specifically, estimates and inferences for $\boldsymbol{\theta}$ follow from maximizing the log-likelihood function, based on the marginal density for $\mathbf{Y}_i$ given by

$$f_i(\mathbf{y}_i|\boldsymbol{\theta}, Q) = \int f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\theta}) dQ(\mathbf{b}_i), \quad (2)$$

where the integration is replaced by a sum in case of a discrete mixing distribution $Q(\mathbf{b}_i)$. Note that this approach can be interpreted as a flexible way of deriving multivariate marginal likelihoods, which makes it particularly appealing for modeling longitudinally measured outcomes Y_{ij} for a number of subjects. Indeed, even if two outcomes Y_{ij} and Y_{ik} are assumed independent, conditional on $\mathbf{b}_i$, marginally they are correlated because they share the same latent variable $\mathbf{b}_i$.

It follows from the classical maximum likelihood theory that, if Q is assumed to belong to some parametric family of distributions (i.e., a set of distributions indexed by a finite number of parameters), all parameter estimators are consistent and asymptotically normally distributed, from which inference immediately follows. If Q is assumed to be a discrete distribution, the implied model is called a ‘latent class model’ or ‘mixture model,’ see, e.g., McCutcheon (1987), McLachlan and Peel (2000), or Böhning (1999). Latent class models assume that the population studied consists of sub-populations characterised by a single common value for $\mathbf{b}_i$. If Q is assumed to be continuous, the implied model is often called a ‘mixed model,’ see, e.g., Verbeke and Molenberghs (2000) or Molenberghs and Verbeke (2005). In many cases, the $\mathbf{b}_i$ are assumed to be sampled from a (multivariate) normal distribution, but alternatives are possible, as will be illustrated in Section 4.2. It is also possible not to make any assumptions about the mixing distribution, and to estimate Q by the distribution which maximizes the likelihood over the class of all distributions, which is referred to as the nonparametric maximum likelihood estimator (NPMLE) of the mixing distribution. Kiefer and Wolfowitz (1956) have shown that the corresponding estimator $\hat{\boldsymbol{\theta}}$ is strongly consistent and also that $\hat{Q}$ converges with probability one to Q at every point of continuity of the latter. Later, Laird (1978) gave sufficient conditions such that the NPMLE would be a step function, and Lindsay (1983a–c) showed that, under general regularity conditions, the NPMLE exists, is discrete, and an upper bound for the number of points of support can be given. In some contexts, such as the Rasch model, conditions can be derived under which the nonparametric estimate for the population parameters coincide with the estimate obtained from a conditional approach (if estimable). More details on the relation between a NPML approach and the conditional inference approach can be found in Lindsay (1983b) and Lindsay, Clogg, and Grego (1991).

Figure 1

Leuven Diabetes Study. Histogram of the observed rates of patients on target, per primary care physician (PCP).

3. ILLUSTRATIVE CASE STUDY: LEUVEN DIABETES STUDY

To illustrate the various options available in a latent variable modelling approach, we analyse data from the Leuven Diabetes Study (Borgermans *et al* 2008). The purpose of the study was to assess clinician-level patterns of diabetes care in fee-for-service practices in which no previous systematic efforts to implement the Chronic Care Model had been attempted. All 379 primary care physicians (PCPs) that actively execute their profession within the Leuven (Belgium) area were invited to participate. The only inclusion criterion for the providers is the agreement to bring in all their known patients with type 2 diabetes mellitus. A total of 2218 patients treated by 106 PCP's were included, the number of patients per PCP varying between 1 and 65, with median value 16. One of the primary outcomes was glycosylated hemoglobin (HbA1c), a molecule in red blood cells that attaches to glucose. High values reflect a lot of glucose in the blood of the patient. In diabetes patients, HbA1c gives a good estimate of how well diabetes is being managed over the last 2 or 3 months. Non-diabetics have values between 4% and 6%, while a value above 7% is generally accepted as an indication that diabetes is poorly controlled, implying higher risk for long-term complications. We therefore study for each PCP the proportion of patients included in the study for which HbA1c is below 7%, i.e., the proportion of patients on target.

Let Y_i be the number of patients on target, treated by PCP i , $i = 1, \dots, 106$, and let n_i be the number of type 2 diabetes mellitus patients PCP i included in the study. Y_i then follows a Binomial $\text{Bin}(n_i, p_i)$. A histogram of the observed rates is shown in **Figure 1**. As a first model (Model 1) we consider a single Binomial model, $Y_i \sim \text{Bin}(n_i, p)$. The results are summarized in **Table 1**. A key assumption of the Binomial model is that it assumes a very strict relation between the mean and the variance, i.e.,

$$\left. \begin{aligned} \mu_i &= E(Y_i) = n_i p \\ \sigma_i^2 &= \text{Var}(Y_i) = n_i p(1 - p) \end{aligned} \right\} \implies \sigma_i^2 = \mu_i(n_i - \mu_i)/n_i. \quad (3)$$

As an informal check on the validity of the mean-variance link (3), the model was refitted

Table 1 Leuven Diabetes Study. Parameters (standard errors) and deviances for binomial and beta-binomial model. Standard errors have only been reported for parameters that take values on the entire real line.

Parameter	Model 1		Model 2		Model 3	
$p = \text{E}(p_i)$	0.673		0.673		0.666	
$\theta = \ln[p/(1 - p)]$	0.720	(0.045)	0.720	(0.060)	0.688	(0.062)
<u>Overdispersion:</u>						
ϕ			1.745			
ψ					0.035	
<u>Latent variable:</u>						
Beta shape α					18.489	
Beta shape β					9.289	
Deviance $-2\ell\ell$	511.5		511.5		489.4	

allowing for a scale factor ϕ , such that $\sigma_i^2 = \phi\mu_i(n_i - \mu_i)/n_i$ (Model 2), the results of which are also shown in **Table 1**. The Pearson χ^2 estimate of ϕ , defined by $\sum_i (y_i - \mu_i)^2 / [(N - 1)\sigma_i^2] = 1$, equals 1.745, indicating that there is 75% more variability in the data than what can be expected based on a single Binomial distribution.

One way to handle such ‘overdispersion’ is to allow for random success rates p_i . Indeed, let $E(p_i) = p$ and denote $\text{Var}(p_i)$ as $\psi p(1-p)$ for some ψ , it then immediately follows that $E(Y_i) = n_i p$ and

$$\text{Var}(Y_i) = n_i p(1-p)[1 + (n_i - 1)\psi].$$

Comparison to (3) immediately shows that overdispersion is obtained as soon as $\psi > 0$. Hence, latent variable models as introduced in Section 2 clearly account for overdispersion. Note also that no overdispersion is possible if $n_i = 1$, i.e., for Bernoulli data. In Sections 4.2 to 3.3, various approaches discussed in Section 2 will be applied.

3.1. Mixed Models

A very natural way to allow success rates p_i to be random is to assume them to be sampled from a Beta distribution with parameters α and β , i.e., the density of p_i equals

$$f(p_i|\alpha, \beta) = \frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(.,.)$ denotes the beta function. The mean and variance of p_i equal

$$E(p_i) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(p_i) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

from which it follows that the overdispersion parameter equals $\psi = (\alpha + \beta + 1)^{-1}$. The marginal density of Y_i is then given by

$$\begin{aligned} f_i(y_i|\alpha, \beta) &= \int \binom{n_i}{\mathbf{y}_i} p_i^{y_i} (1-p_i)^{n_i-y_i} f(p_i|\alpha, \beta) dp_i \\ &= \binom{n_i}{\mathbf{y}_i} \frac{B(y_i + \alpha, n_i - y_i + \beta)}{B(\alpha, \beta)}, \end{aligned} \quad (4)$$

Table 2 Leuven Diabetes Study. Parameters (standard errors) and deviances for various mixed models. Standard errors have only been reported for parameters that take values on the entire real line.

Parameter	Model 4		Model 5	
$p = \text{E}(p_i)$	0.666		0.665	
<u>Latent variable:</u>				
$\theta = \text{E}(b_i)$	0.715	(0.064)	0.441	(0.039)
$\sigma^2 = \text{Var}(b_i)$	0.168		0.062	
Deviance $-2\ell\ell$	489.7		489.5	

Parameter	Model 6		Model 7	
$p = \text{E}(p_i)$	0.665		0.667	
<u>Latent variable:</u>				
$\theta = \text{E}(b_i)$	0.723	(0.064)	0.445	(0.039)
$\sigma^2 = \text{Var}(b_i)$	0.166		0.060	
<u>Normal mixture:</u>				
Mean θ_1^*	0.523	(0.167)	0.323	(0.099)
Mean θ_2^*	1.371	(0.290)	0.824	(0.157)
Variance σ^{*2}	0.037		0.014	
Deviance $-2\ell\ell$	488.2		488.2	

the so-called beta-binomial model (Skellam 1948, Kleinman 1973). The results obtained from fitting this model to data from the Leuven Diabetes Study are reported in **Table 1** (Model 3). The overdispersion parameter was estimated as $\hat{\psi} = 0.035$, implying the ratio $1 + (n_i - 1)\psi$ of fitted variance over fitted mean to vary between 0, for a PCP with $n_i = 1$ patient, and 3.22, for a PCP with $n_i = 65$ patients in the study, with average value 1.694, close to the scale factor $\phi = 1.745$ obtained before.

An alternative, frequently used, class of mixed models assumes the success rates p_i to be normally distributed after an appropriate transformation, $b_i = \eta(p_i)$, selected such that the b_i take values on the entire real line. Frequently used link functions $\eta(\cdot)$ are the logit link $\eta(x) = \ln[x/(1 - x)]$ and the probit link $\eta(x) = \Phi^{-1}(x)$ in which $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Assuming the b_i to be normally $N(\theta, \sigma^2)$ distributed, the so-called ‘logistic-normal’ (Model 4) and ‘probit-normal’ (Model 5) models are obtained. The results found from fitting both models to the Leuven Diabetes Study are summarized in **Table 2**. In terms of deviance values, both models provide a very similar fit to the data. It has been shown (Agresti 1990) that the logit and the probit link functions behave very similarly, in the sense that for probabilities other than extreme ones (say, outside of the interval $[0.2; 0.8]$) logistic and probit regression models provide approximately the same parameter estimates, up to a scaling factor equal to $\pi/\sqrt{3}$.

Both models can be further extended by replacing the normality assumption for the latent variables b_i by a more flexible one. As an illustration, the logistic model (Model 6) and the probit model (Model 7) have been re-fitted, assuming the b_i to follow a mixture of two normals with common variance, i.e.,

$$b_i \sim \pi_1 N(\theta_1^*, \sigma^{*2}) + \pi_2 N(\theta_2^*, \sigma^{*2}),$$

with $\pi_1 + \pi_2 = 1$. The mean and variance of the latent variables then immediately follow

Table 3 Leuven Diabetes Study. Parameters and deviances for various latent class models.

Parameter	Model 8	Model 9	Model 10
$p = E(p_i)$	0.673	0.667	0.667
Latent variable:			
$\begin{pmatrix} p_1^* & \dots & p_g^* \\ \pi_1 & \dots & \pi_g \end{pmatrix}$	$\begin{pmatrix} 0.615 & 0.792 \\ 0.704 & 0.296 \end{pmatrix}$	$\begin{pmatrix} 0.606 & 0.756 & 0.869 \\ 0.629 & 0.329 & 0.042 \end{pmatrix}$	$\begin{pmatrix} 0.498 & 0.612 & 0.759 & 0.870 \\ 0.021 & 0.624 & 0.315 & 0.040 \end{pmatrix}$
Deviance $-2\ell\ell$	488.5	487.8	486.0

as

$$\begin{aligned} E(b_i) &= \pi_1 \theta_1^* + \pi_2 \theta_2^*, \\ \text{Var}(b_i) &= \sigma^{*2} + \pi_1 \theta_1^{*2} + \pi_2 \theta_2^{*2} - (\pi_1 \theta_1^* + \pi_2 \theta_2^*)^2. \end{aligned}$$

The results are also shown in **Table 2**. Admittedly, a formal comparison of the normal model with the normal mixture model cannot be based on the change in deviance due to the null-hypothesis being on the boundary of the parameter space, and parameters vanishing. Still, the very minor change in deviance, as well as very similar predictions for the average success rate p suggest that there is very little evidence in favor of the model extension.

Note that, due to the non-linear link functions, the marginal success rate p does not directly follow from the model and additional calculations,

$$p = E(p_i) = E[\eta^{-1}(\theta_i)] = \int \eta^{-1}(\theta_i) dQ(b_i),$$

are needed. The values reported in **Table 2** were obtained as sample averages $\sum_i \eta^{-1}(b_i)/n$ of $n = 10,000$ values b_i randomly generated from the fitted latent variable distributions (normal or normal mixture).

3.2. Latent Class Models

In latent class models, the success rates p_i are assumed to be sampled from a discrete distribution with g support points $p_1^*, \dots, p_g^*$ with corresponding probabilities $\pi_1, \dots, \pi_g$, $\sum_j \pi_j = 1$. The overall average success rate then equals $p = \sum_j \pi_j p_j^*$. As an illustration, we fit two latent class models with $g = 2$ (Model 8) and $g = 3$ (Model 9) support points. The results are summarized in **Table 3**. The fits are very similar to those obtained from the mixed models presented in Section 2.

3.3. Non-parametric Maximum Likelihood (NPML)

As explained in Section 2, one may prefer not to make any prior assumptions about the mixing distribution Q , and to estimate Q by the distribution that maximizes the likelihood over the class of all distributions. Under very general regularity conditions, the so-obtained nonparametric maximum likelihood estimator (NPMLE) of the mixing distribution will be discrete with a finite number of support points. For the Leuven Diabetes Study, the NPMLE is a 4-component model, the results for which are shown in **Table 3** (Model 10). As expected, the resulting deviance is the smallest amongst all models considered. Böhning

Figure 2

Leuven Diabetes Study. Gradient function corresponding to Model 10.

(1999) recommends to plot the gradient function

$$\Delta(\hat{Q}, p) = \frac{1}{N} \sum_{i=1}^N \frac{f_i(y_i|p)}{f_i(y_i|\hat{Q})}, \quad p \in [0; 1],$$

of the fitted mixing distribution $\hat{Q}$, to graphically check that the obtained solution is indeed NPML. $\hat{Q}$ is NPML if and only if its gradient function does not exceed the value 1. Moreover, the gradient will be equal to 1 in all support points of $\hat{Q}$. Finally, $\Delta(\hat{Q}, p) \equiv 1$ if and only if the NPMLE is not unique. The gradient function for Model 10 is shown in **Figure 2**, confirming that the mixing distribution obtained in Model 10 indeed is the unique nonparametric maximum likelihood estimate.

4. APPLICATIONS, EXTENSIONS, AND ISSUES

4.1. Modeling Associations

So far, focus has been on the use of latent variables to handle overdispersion, i.e., to model variability in a flexible way. In many contexts, latent variables are used to model associations as well. A typical example is the analysis of longitudinal data. Continuing with the general notation introduced in Section 2, let Y_{ij} be the j th measurement of some continuous outcome Y , taken at time-point t_{ij} , for subject i in a longitudinal study, $i = 1, \dots, N$, $j = 1, \dots, n_i$. Assuming that, apart from some residual variability, each subject can be well represented by a straight line, a potential model is

$$Y_{ij} = (\beta_1 + b_{i1}) + (\beta_2 + b_{i2})t_{ij} + \varepsilon_{ij}. \quad (5)$$

The latent vector $\mathbf{b}_i = (b_{i1}, b_{i2})'$ models how the evolution of subject i deviates from the average trend $\beta_1 + \beta_2 t$ in the population. Often, $\mathbf{b}_i$ is assumed to be $N(\mathbf{0}, D)$ distributed, for some covariance matrix D , and the residuals ε_{ij} are assumed all independent $N(0, \sigma^2)$.

Since the models of two measurements Y_{ij} and Y_{ik} from the same subject share the latent variables b_{i1} and b_{i2} , a marginal association

$$\text{Cov}(Y_{ij}, Y_{ik}) = d_{22}t_{ij}t_{ik} + d_{12}(t_{ij} + t_{ik}) + d_{11} + \sigma^2$$

is implied. Hence, the latent variables not only model the variability in the data, they also imply a specific association structure. Extensions of this so-called ‘linear mixed model’ (Verbeke and Molenberghs 2000) to non-Gaussian data can be obtained from latent variables to the linear predictor in a generalized linear models. The resulting model is called a ‘generalized linear mixed model’ (Breslow and Clayton 1993, Molenberghs and Verbeke 2005).

This principle is also used to model associations in other contexts. For example, Rizopoulos (2012) uses latent variables to implicitly model the association between longitudinally measured outcomes and some time-to-event outcome. In case of continuous longitudinal outcomes, a linear mixed model such as (5) is used, the latent variables b_{i1} and b_{i2} of which are used as covariates in the time-to-event model. Other examples include shared parameter models to jointly model longitudinal data and some dropout process (Molenberghs and Kenward 2007), random-effects models for multivariate longitudinal data (Fieuws and Verbeke 2006, Fieuws *et al* 2008, and Verbeke *et al* 2014), and random-effects models for joint modeling of outcomes of different types (Ivanova, Molenberghs, and Verbeke 2015).

All these models have in common that the latent variable structure implies a specific association model which may or may not be appropriate in a particular context. As an illustration, we consider an example from Verbeke *et al* (2014) on the joint modeling of two longitudinally measured continuous responses Y_{1i} and Y_{2i} , modeled using a linear mixed model similar to (5):

$$\begin{aligned} Y_{1ij} &= \beta_{11} + b_{i1} + \beta_{12}t_{ij} + \varepsilon_{1ij}, \\ Y_{2ij} &= \beta_{21} + \gamma b_{i1} + \beta_{22}t_{ij} + \varepsilon_{2ij}. \end{aligned} \quad (6)$$

The random intercept b_{i1} is shared by both longitudinal processes and implies an association structure between them. More specifically, the correlation between a measurement for Y_{1i} at time-point s and a measurement for Y_{2i} at time-point t is given by

$$\text{Corr}\{Y_{1i}(s), Y_{2i}(t)\} = \sqrt{\text{Corr}\{Y_{1i}(s), Y_{1i}(t)\}} \sqrt{\text{Corr}\{Y_{2i}(s), Y_{2i}(t)\}}, \quad \forall s \neq t, \quad (7)$$

implying that the correlation structures of the individual outcomes dictate the association between pairs of measurements from different outcomes. For example, the model would not allow that Y_{1i} and Y_{2i} to be independent if repeated measures of Y_{1i} and Y_{2i} are strongly correlated. Similar restrictions hold in more general models that involve shared random effect vectors beyond just a scalar random intercept b_{i1} . Such restrictions can be relaxed by allowing for different but correlated random effects for the various outcomes:

$$\begin{aligned} Y_{1ij} &= \beta_{11} + b_{i1} + \beta_{12}t_{ij} + \varepsilon_{1ij}, \\ Y_{2ij} &= \beta_{21} + b_{i2} + \beta_{22}t_{ij} + \varepsilon_{2ij}, \end{aligned} \quad (8)$$

in which b_{i1} and b_{i2} are now assumed to jointly follow a bivariate normal distribution. The correlation between a measurement for Y_{1i} at time-point s and a measurement for Y_{2i} at time-point t is now given by

$$\text{Corr}\{Y_{1i}(s), Y_{2i}(t)\} = \text{Corr}(b_{i1}, b_{i2}) \sqrt{\text{Corr}\{Y_{1i}(s), Y_{1i}(t)\}} \sqrt{\text{Corr}\{Y_{2i}(s), Y_{2i}(t)\}}, \quad \forall s \neq t,$$

showing that correlation between the outcome-specific latent variables generates a between-process association, but also that the restriction imposed by model (6) with shared latent variable is relaxed in the sense that the correlation structures of the individual outcomes no longer dictate the association between pairs of measurements from different outcomes. For example, the model would now perfectly allow Y_{1i} and Y_{2i} to be independent even if repeated measures of Y_{1i} and Y_{2i} would be strongly correlated. Finally, note that model (6) with shared latent variable can be obtained as a special case of (8) by specifying perfect correlation between b_{i1} and b_{i2} .

4.2. Models with Latent Variables of Mixed Types

In some cases multiple latent variables are needed to separately model overdispersion and association. For example, Molenberghs, Verbeke, and Demétrio (2007) analysed overdispersed longitudinal count data using a Poisson model with Gamma-distributed as well as normally distributed random effects. Using notation as in Section 4.1, they assumed that, conditionally on normally distributed latent variables $\mathbf{b}_i$ and Gamma(α, β), latent variables θ_{ij} , Y_{ij} followed a Poisson distribution with mean $\lambda_{ij} = \theta_{ij} \exp(x'_{ij}\beta + z'_{ij}\mathbf{b}_i)$, for two vectors x_{ij} and z_{ij} of known covariates. Special cases are obtained by omitting the $\mathbf{b}_i$ latent variables, the θ_{ij} latent variables, or both. The resulting models are the negative-binomial model, the poisson-normal mixed model, or the standard poisson regression model, respectively. As discussed in Section 4.1, the normal latent variables imply association amongst the repeated measurements from one subject. However, they also imply specific assumptions about the variance structure in the data. If any fine-tuning is needed, this is done using the latent variables θ_{ij} . The marginal density of $\mathbf{Y}_i$ can be derived analytically, and a easily implemented fitting algorithm was proposed (see Molenberghs, Verbeke, and Demétrio 2007). Later on, the same authors extended the approach to outcomes other than counts. We refer to Molenberghs *et al* (2010) for a presentation of the general family of generalized linear models for repeated measures with normal and conjugate random effects.

4.3. Parameter Interpretation

As already illustrated in Section 4.2 in the context of the Leuven Diabetes Study, parameters in latent variable models need to be interpreted conditional on the latent variables and, in general, do not have a marginal interpretation. For example, consider the logistic mixed model

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \quad \ln \left[\frac{\pi_{ij}}{1 - \pi_{ij}} \right] = (\beta_1 + b_i) + \beta_2 t_{ij}, \quad b_i \sim N(0, \sigma^2).$$

The conditional means $E(Y_{ij}|b_i)$, as functions of t_{ij} , are given by

$$E(Y_{ij}|b_i) = \frac{\exp[(\beta_1 + b_i) + \beta_2 t_{ij}]}{1 + \exp[(\beta_1 + b_i) + \beta_2 t_{ij}]} \quad (9)$$

The model assumes that the conditional means all satisfy a logistic model, with the same slope β_1 but with different intercepts $\beta_0 + b_i$ for all subjects. The marginal average evolution $E(Y_{ij})$ is obtained from averaging (9) over the latent variables b_i , i.e.,

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E \left[\frac{\exp(\beta_1 + b_i + \beta_2 t_{ij})}{1 + \exp(\beta_1 + b_i + \beta_2 t_{ij})} \right] \neq \frac{\exp(\beta_1 + \beta_2 t)}{1 + \exp(\beta_1 + \beta_2 t)}. \quad (10)$$

Figure 3

Graphical representation of a logistic model with latent intercepts, across a range of levels of the random intercept (black lines), together with the corresponding marginal curve (blue line).

A graphical representation of both (9) and (10) is given in **Figure 3**. Obviously, the marginal time trend is much less steep than each of the individual time trends, and the difference strongly depends on the variance of the latent variable: The larger the latent-variable variability, the more the slope of the average trend differs from the slopes of the trends conditional on any value for the latent variable b_i .

In general, this phenomenon holds for models where the conditional mean $E(\mathbf{Y}_i|\mathbf{b}_i)$ is a non-linear function of the latent variable $\mathbf{b}_i$. This is in strong contrast to linear models, where effects in the conditional mean model $E(\mathbf{Y}_i)$ keep the same interpretation in the marginal mean model $E(\mathbf{Y}_i|\mathbf{b}_i)$. For example, β_2 in (5) has a subject-specific interpretation as the slope of the median subject characterized by $\mathbf{b}_i = \mathbf{0}$, as well as a marginal interpretation as the slope of the average trend in the general population. For some models, e.g., the Poisson or Weibull model, where the conditional means $E(Y_{ij}|\mathbf{b}_i)$ are modeled as $\exp(x'_{ij}\beta + b_i)$ for scalar latent variables b_i , we have that the marginal means equal

$$E(Y_{ij}) = E[\exp(b_i)] \exp(x'_{ij}\beta),$$

showing that the interpretation of only the intercept is affected by the inclusion of the latent variables, while all other parameters in β have a conditional as well as marginal interpretation.

4.4. Marginalized Multilevel Models and Bridge Distributions

Several proposals have been made in the statistical literature to deal with the issues discussed in Section 4.3, i.e., to construct latent variable models that allow for a marginal interpretation of regression parameters β in a linear predictor $x'_{ij}\beta$.

Heagerty (1999) introduced ‘marginalized multilevel models’ in which latent variables are used to generate the distribution, including the association structure, for a vector $\mathbf{Y}_i$ of correlated outcomes Y_{ij} but with a conditional mean model that is sufficiently flexible for the implied marginal mean to be of a pre-specified desired form, allowing for a marginal interpretation of the parameters. More specifically, the marginal and conditional means are

modeled as

$$\eta[E(Y_{ij})] = x'_{ij}\beta, \quad (11)$$

$$\eta[E(Y_{ij}|\mathbf{b}_i)] = \Delta_{ij} + z'_{ij}\mathbf{b}_i, \quad (12)$$

for some prespecified link function $\eta(\cdot)$. The Δ_{ij} are functions connecting the marginal and conditional means (11) and (12), respectively, obtained from solving an integral equation

$$E(Y_{ij}) = \int \eta^{-1}(\Delta_{ij} + z'_{ij}\mathbf{b}_i)dQ(\mathbf{b}_i).$$

Wang and Louis (2003) specified a model through the constraint that the marginal mean and conditional mean are specified by the same link functions, with predictors that are the same up to a multiplicative factor ϕ and an offset k :

$$\eta[E(Y_{ij})] = k + \phi x'_{ij}\beta, \quad (13)$$

$$\eta[E(Y_{ij}|\mathbf{b}_i)] = x'_{ij}\beta + z'_{ij}\mathbf{b}_i. \quad (14)$$

Specification of (13) and (14) is similar to (11) and (12). The main difference with marginalized multilevel models is that now the distribution of $\mathbf{b}_i$ is not pre-specified but follows from the restrictions implied by the other model specifications, while in the approach of Heagerty (1999) the latent variable distribution is pre-specified from which the connector functions Δ_{ij} in (12) follow. Wang and Louis (2003) focused on the binary case, where $\eta(\cdot)$ is, for example, the logit, probit, or complementary log-log function. They restricted to models with random intercepts only. Extensions of the original model to settings where the mixing distribution depends on covariates or to settings with outcomes other than binary ones have been proposed by, e.g., Wang and Louis (2004) and Molenberghs *et al* (2013).

4.5. Identifiability

As explained in Section 2, latent variable models are defined through the conditional model $f_i(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)$ and the mixing distribution $Q(\mathbf{b}_i)$, and inference is based on the marginal model $f_i(\mathbf{y}_i|\boldsymbol{\theta}, Q)$. It should be emphasized that multiple combinations of conditional models and mixing distributions can lead to the same marginal model, implying identifiability issues.

As a simple example, consider the case where $\mathbf{Y}_i$ is bivariate, $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$, modeled as

$$\begin{aligned} Y_{i1} &= \beta_1 + b_i + \varepsilon_{i1}, \\ Y_{i2} &= \beta_2 + b_i + \varepsilon_{i2}, \end{aligned}$$

with the latent variable b_i normally $N(0, \sigma_b^2)$ distributed and with independent error terms both normally distributed with means zero and variances σ_1^2 and σ_2^2 , respectively. The implied marginal distribution for $\mathbf{Y}_i$ is bivariate normal with mean vector $(\beta_1, \beta_2)'$ and covariance matrix

$$V = \begin{pmatrix} \sigma_b^2 + \sigma_1^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_2^2 \end{pmatrix}. \quad (15)$$

As an alternative model, we consider

$$\begin{aligned} Y_{i1} &= \beta_1 + b_{i1} + \varepsilon_{i1}, \\ Y_{i2} &= \beta_2 + b_{i2} + \varepsilon_{i2}, \end{aligned}$$

in which the latent vector $\mathbf{b}_i = (b_{i1}, b_{i2})'$ follows a bivariate normal distribution with mean $(0, 0)'$, with variances σ_b^2 and $\psi\sigma_b^2$, respectively, and with covariance σ_b^2 . The implied marginal distribution for $\mathbf{Y}_i$ is then again bivariate normal with mean vector $(\beta_1, \beta_2)'$, but the covariance matrix equals

$$V = \begin{pmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 \\ \sigma_b^2 & \psi\sigma_b^2 + \sigma^2 \end{pmatrix}. \quad (16)$$

Obviously, the parametric models (15) and (16) are two different parameterizations with 3 parameters of the 2×2 covariance matrix V , and hence yield identical marginal fits to the data. This shows that different latent variable models can produce the same marginal model, implying that the mixing distribution $Q(\mathbf{b}_i)$ is only identified conditional on the specification of the conditional model $f_i(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)$.

This has been shown more generally by Molenberghs *et al* (2012), where a unified framework of enriched data was used to show that information required to identify models such as latent variables models is divided in information supplied by the data and information supplied externally, through assumptions and/or scientific knowledge. More specifically, they have shown that any model for enriched data can be replaced by an infinite family of models, all retaining the same fit to the observed data as achieved by the original model. In the latent variables context, this immediately can be seen as follows. Let $f(\mathbf{b}_i)$ be the density corresponding to the assumed mixing distribution $Q(\mathbf{b}_i)$. The marginal model then equals

$$\begin{aligned} f(\mathbf{y}_i|\boldsymbol{\theta}, Q) &= \int f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i \\ &= \int \frac{f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)f(\mathbf{b}_i)}{g(\mathbf{b}_i)}g(\mathbf{b}_i)d\mathbf{b}_i \\ &= \int g(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)g(\mathbf{b}_i)d\mathbf{b}_i = g(\mathbf{y}_i|\boldsymbol{\theta}, Q^*), \end{aligned}$$

with $g(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i) = f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)f(\mathbf{b}_i)/g(\mathbf{b}_i)$ and with $g(\mathbf{b}_i)$ the density corresponding to any other mixing distribution $Q^*(\mathbf{b}_i)$ with the same support as $Q(\mathbf{b}_i)$. It immediately follows that many latent variable models can be found with identical fit in terms of marginal likelihood, provided that one does not restrict to one single conditional model $f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)$ only. While this points to some identifiability problem, it is also computationally convenient for fitting latent variable models with non-standard mixing distributions, using standard software (Liu and Yu 2007). For example, Model 6 and Model 7 in Section 4.2 were fitted using standard software for models with Gaussian random effects, after replacing the original mixture density $f(\mathbf{b}_i)$ by the density $g(\mathbf{b}_i)$ of a standard normal density, and replacing the original binomial density $f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)$ by $g(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i) = f(\mathbf{y}_i|\boldsymbol{\theta}, \mathbf{b}_i)f(\mathbf{b}_i)/g(\mathbf{b}_i)$.

In summary, the mixing distribution $Q(\mathbf{b}_i)$ can only be identified if the distribution of the outcome $\mathbf{Y}_i$ conditional on $\mathbf{b}_i$ is assumed to be correctly specified. Under that assumption, Verbeke and Molenberghs (2013) developed a simple graphical technique to assess the fit of a selected mixing distribution, applicable in a wide class of latent variable models. As explained in Section 3.3, the gradient function can be used to graphically check whether a fitted mixing distribution $\hat{Q}$ is a non-parametric maximum likelihood estimate, i.e., if it maximizes the marginal log-likelihood over the class of all possible mixing distributions. The idea is to check whether the gradient function corresponding to the fitted model shows

a behavior close to what can be expected for the gradient function corresponding to the best fitting model, i.e., a gradient function not exceeding the value 1 and reaching 1 in all support points of the fitted mixing distribution. For example, in order to check whether a scalar latent variable b_i is normally distributed, it is sufficient to check whether the gradient function

$$\Delta(\hat{Q}, b, \hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \frac{f_i(y_i|\hat{\theta}, b)}{f_i(y_i|\hat{Q})},$$

with $\hat{Q}$ the fitted normal mixing distribution, equals the value 1 over the entire real line. Severe deviations from 1 would indicate that the model can be improved considerably by replacing the fitted mixing distribution $\hat{Q}$ by another one, implying non-normality of b_i . Furthermore, it can be shown that the gradient function does not need to be studied over the entire real line, but that attention can be restricted to the closed interval I which contains the unique modes $\hat{b}_i$ of all functions $f_i(y_i|\hat{\theta}, b_i)$, $i = 1, \dots, N$. Finally, since the gradient can be interpreted as a sample average, its asymptotic distribution is normal and pointwise confidence limits can easily be obtained from applying the central limit theorem with variance estimated by the sample variance of the likelihood ratio contributions $f_i(y_i|\hat{\theta}, b)/f_i(y_i|\hat{Q})$. Such limits are useful for assessing the importance of any deviations of the gradient from 1. The elegance of the gradient function to check the fit of a particular mixing distribution is that it does not require any computations additional to those needed to fit the latent variables model of interest, and the method is not restricted to any particular type of model. The disadvantage is that it provides an exploratory graphical check only. Efendi *et al* (2015) developed a formal test for the goodness-of-fit of a mixing distribution, based on the deviation from the gradient function from the value 1.

5. CONCLUSION

In this paper, a review of latent variable models has been given, and it has been shown that many frequently used statistical models can be viewed as special cases of this general unified class of models. These special cases include latent class models, as well as mixed models, but also the popular beta-binomial model for binomial counts. Many of these models have been developed independently in various subfields of statistical research, often leading to different terminologies used for the same model. For example, latent class models and finite mixture models both refer to models with discrete latent variables. Similarly, mixed models, random-effects models, and multi-level models all refer to models with continuous latent variables.

The common feature of all these models is that they model observed data conditional on one or more latent, unobserved, variables. Inference is based on the marginal distribution obtained from integrating out the latent variables over their assumed mixing distribution. Depending on the assumed conditional distribution and on the assumed mixing distribution a variety of models can be obtained. Therefore, latent class models are extremely flexible for handling complex data structures, including hierarchically structured data, correlated data, and overdispersed data. On the other hand, the fact that outcomes are modeled on structures which are not observed imply that caution is needed when interpreting inferences obtained from such models. Furthermore, since different latent variable models can lead to the same marginal model, and hence to the same fit to the observed data, the models are

not identified without additional constraints. Finally, the inclusion of unobserved variables in the model makes model assessment far from straightforward. We have discussed all those issues, presented potential solutions, and introduced various model extensions. Finally, the most commonly used models have been illustrated in the analysis of the Leuven Diabetes Study, in which 10 different latent variable models were used to estimate the success rate for a binomial outcome which cannot be modeled using a single Binomial distribution, due to overdispersion.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

LITERATURE CITED

- Agresti A. 1990. *Categorical data analysis*. John Wiley & Sons, Inc.
- Böhning D. 1999. *Computer-Assisted Analysis of Mixtures and Applications: meta-analysis, disease mapping and others*. Number 81 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.
- Borgermans L, Goderis G, Van Den Broeke C, Mathieu C, Aertgeerts B, Verbeke G, Carbonez A, Ivanova A, Grol R, Heyrman J. 2008. A cluster randomized trial to improve adherence to evidence-based guidelines on diabetes and reduce clinical inertia in primary care physicians in Belgium: Study protocol [NTR 1369]. *Implementation Science*, 3:42.
- Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.
- Breslow NE, Day NE. 1989. *Statistical methods in cancer research. Volume 1 : The analysis of case-control studies*. International Agency for Research on Cancer, Scientific Publications 32.
- Diggle PJ, Heagerty P, Liang KY, Zeger SL. 2002. *Analysis of longitudinal data*. Clarendon Press, Oxford.
- Fieuws S, Verbeke G. 2006. Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics*, 62:424–431.
- Fieuws S, Verbeke G, Maes B, Van Renterghem Y. 2008. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, 9:419–431.
- Heagerty PJ. 1999. Marginalized specified logistic-normal models for longitudinal binary data. *Biometrics*, 55:688–698.
- Ivanova A, Molenberghs G, Verbeke G. 2015. Mixed models approaches for joint modeling

- of different types of responses. *Journal of Biopharmaceutical Statistics*, 22:1–18 [Epub ahead of print].
- Kiefer J, Wolfowitz J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27:887–906.
- Kleinman J. 1973. Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, 68:46–54.
- Laird NM. 1978. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.
- Lindsay BG. 1983a. The geometry of mixture likelihoods, part I : A general theory. *The Annals of Statistics*, 11:86–94.
- Lindsay BG. 1983b. Efficiency of the conditional score in a mixture setting. *The Annals of Statistics*, 11:486–497.
- Lindsay BG. 1983c. The geometry of mixture likelihoods, part II : The exponential family. *The Annals of Statistics*, 11:783–792.
- Lindsay BG, Clogg CC, Grego J. 1991. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.
- Liu L, Yu Z. 2007. A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, 27:3105 – 3124.
- McCutcheon AL. 1987. *Latent Class Analysis*. Number 64 in Quantitative Applications in the Social Sciences Series. Thousand Oaks, California: Sage Publications.
- McLachlan GJ, Peel D. 2000. *Finite mixture models*. John Wiley & Sons, Inc.
- Molenberghs G, Kenward MG. 2007. *Missing Data in Clinical Studies*. John Wiley & Sons, New York.
- Molenberghs G, Kenward MG, Verbeke G, Iddi S, Efendi A. 2013. On the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models. *International Journal of Statistics and Probability*, 2:1–21.
- Molenberghs G, Njeru Njagi E, Kenward MG, Verbeke G. 2012. Enriched-data problems and essential non-identifiability. *International Journal of Statistics in Medical Research*, 1:16–44.
- Molenberghs G, Verbeke G. 2005. *Models for discrete longitudinal data*. Springer Series in Statistics. Springer, New-York.
- Molenberghs G, Verbeke G, Demétrio CGB. 2007. An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13:513–532.
- Molenberghs G, Verbeke G, Demétrio CGB, Viera A. 2010. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25:325–347.
- Neyman J, Scott EL. 1948. Consistent estimates based on partially consistent observations.

Econometrica, 16:1–32.

- Rizopoulos D. 2012. *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.
- Skellam JG. 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, 10:257–261.
- Verbeke G, Fieuws S, Molenberghs G, Davidian M. 2014. The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23:42–59.
- Verbeke G, Molenberghs G. 2000. *Linear mixed models for longitudinal data*. Springer Series in Statistics. Springer, New-York.
- Verbeke G, Molenberghs G. 2013. The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, 14:477–490.
- Verbeke G, Spiessens B, Lesaffre E. 2001. Conditional linear mixed models. *The American Statistician*, 55:25–34.
- Wang Z, Louis TA. 2003. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution. *Biometrika*, 90:765–775.
- Wang Z, Louis TA. 2004. Marginalizing binary mixed-effect models with covariate-dependent random effects and likelihood inference. *Biometrics*, 60:884–891.