Enriched-data problems and essential non-identifiability

Peer-reviewed author version

MOLENBERGHS, Geert; NJAGI, Edmund; Kenward, Michael G. & VERBEKE, Geert (2012) Enriched-data problems and essential non-identifiability. In: International Journal of Statistics in Medical Research, 1 (1), p. 16-44.

DOI: 10.6000/1929-6029.2012.01.01.02
Handle: http://hdl.handle.net/1942/14676

Enriched-data Problems and Essential Non-identifiability

Geert Molenberghs[1,2] **Edmund Njeru Njagi**[1]

Michael G. Kenward[3] **Geert Verbeke**[2,1]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[3] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E7HT, United Kingdom*

Abstract

There are two principal ways in which statistical models extend beyond the data available. First, the data may be coarsened, that is, what is actually observed is less detailed than what is planned, owing to, for example, attrition, censoring, grouping, or a combination of these. Second, the data may be augmented, that is, the observed data are hypothetically but conveniently supplemented with structures such as random effects, latent variables, latent classes, or component membership in mixture distributions. These two settings together will be referred to as *enriched data*. Reasons for modelling enriched data include the incorporation of substantive information, such as the need for predictions, advantages in interpretation, and mathematical and computational convenience. The fitting of models for enriched data combine evidence arising from empirical data with non-verifiable model components, *i.e.,* that are purely assumption driven. This has important implications for the interpretation of statistical analyses in such settings. While widely known, the exploration and discussion of these issues is somewhat scattered. The user should be fully aware of the potential dangers and pitfalls that follows from this. Therefore, we provide a unified framework for enriched data and show in general that to any given model an entire class of models can be assigned, with all of its members producing the same fit to the observed data but arbitrary regarding the unobservable parts of the enriched data. The implications of this are explored for several specific settings, namely that of latent classes, finite mixtures, factor analysis, random-effects models, and incomplete data. The results are applied to a range of relevant examples.

Some Keywords: Compound-symmetry; Empirical Bayes; Enriched data; Exponential random effects; Gamma random effects; Linear mixed model; Missing at random; Missing completely at random; Non-future dependence; Pattern-mixture model; Selection model; Shared-parameter model.

1 Introduction

It is common in statistics to use models that rely on assumptions that cannot be examined from the data under analysis. This is not a weakness, but an inevitable consequence of drawing statistical inferences in the settings in which such models are used. As a consequence, it is important that

the use of these models properly reflects the implied reliance on external information. A good example of the failure to appreciate the nature of such models is provided by the now well-known historical developments surrounding factor analysis in so-called general intelligence measurement[1]. Factor analysis dates back to Pearson and Spearman, though it is the latter that is credited with its introduction into psychology, a field in which it has held popularity for close to a century. Its arrival coincided with a time when psychologists were attempting to quantify 'mental worth' in a scientific sense. Motivated by the positive correlations exhibited by a set of mental tests, Spearman used the technique to develop the so-called "two-factor" theory. The theory implies that a set of mental tests represents an underlying general factor (g), in addition to each test's specific information. Spearman proceeded to accord g some real existence, terming it *general intelligence*. He further proceeded to identify g as an attribute, resident in the brain, which he called *general energy*. Some physical existence is also attributed to the test-specific information (s-factors): he identified them as specific engines in the brain, which are under the influence of the general energy. He also argued g as the theoretical basis of the IQ-testing, which was prevalent at the time: the IQ-test simply measures g, with each component test having a certain loading on g, and certain test-specific information, s. There was no corroborating structural neurological work to support this theory, however. The attribution of real existence to such mathematically constructed abstractions is an example of *reification.* A debate ensued, between two schools of thought: Spearman and Burt on one side and Thurstone on the other. We note that Burt, like Spearman, believed in the supremacy of g, though he also believed in the existence of *group factors*, subsidiary to g. Thurstone faulted Spearman's (and Burt's) method, and produced a solution which totally dispenses with g. The solution, which he called *simple structure*, is actually a rotation of Spearman's principal-components solution. The two solutions explain an identical amount of information, i.e., they fit the observed data equally well. Hence, they differ only in aspects of the model that cannot be verified from the data. As an anonymous referee has pointed out, the value of their respective solutions including their non-verifiable assumptions rests entirely on practical considerations. We show that this phenomenon is very common throughout statistical modeling, and extends across a range of common data-analytic settings, well beyond factor analysis, is the central theme of our paper.

At the time, rather than view this as an indication of the need to acknowledge sensitivity, and consequently refrain from reification, Thurstone proceeded to present his solution as a discovery of the correct explanation of the structure of the mind. The two schools of thought passionately

advocated the validity of their model as the proper representation of the mind. Gould[1] exposes a fundamental flaw which both parties failed to take notice of: that their respective solutions comprised of positioning of axes at locations which represented their *a priori* suppositions of the nature of the mind. Therefore, their respective models merely mirrored their prior belief.

In the following, we illustrate, through a range of settings, the common structure of problems like that of factor analysis above, and show how the practical implications of these rest on a division of information into that supplied by the data under analysis and that supplied externally. We distinguish two broad types of setting that fall under our general heading. The first can be termed *augmented* data, in the sense of supplementing the observed data with latent or unobserved quantities; examples include random-effects models, latent class and latent variable models, and finite-mixture models. The second, introduced by Heitjan[2,3] is a concept called *coarsening*, which refers broadly to situations where the observed data are coarser than the hypothetically conceived data structures, to which the models of interest apply. Examples include incomplete data and censored survival data. It is obvious that models for such augmented structures or coarsened data are identifiable only by virtue of making sometimes strong but always partially non-verifiable assumptions. Augmentation and coarsening taken together, and from now on termed *enriched data*, in line with Verbeke and Molenberghs[4], will be treated in a unified way, such that important, common features can be illuminated and scrutinized. There is a formal distinction between the two types. In the coarse-data setting, it is understood that a part of the data would ideally be observed but is not in practice (*e.g.*, actual survival time after censoring, outcomes after dropout, etc.). Augmented data refers rather to the addition of useful but artificial constructs to the data setting, such as random effects, latent classes, latent variables, factors, and mixture component membership. These can never be observed. Our focus will not be so much on the distinctions between coarse and augmented data on the one hand, nor on subtle distinctions within the coarsening and augmentation families on the other. Rather, we will review a selected range of each and bring out commonality.

Thus, in this paper, we focus on the general enriched-data case and establish that there will always be a part of the model that is totally unidentifiable from the observed data. This implies that the identification of such a part can come from assumptions only. This leads us to the main message of the paper. First, we set how models in enriched-data settings are identified by the triple: data, design-based assumptions (such as randomization), and further unverifiable assumptions. For this we focus on the model itself and its relationship to the data through likelihood. We are not concerned

with subsequent inferences; the same message holds whether we are being Bayesian or frequentist. In each setting considered we identify a part of the model for which, in the Bayesian case, the posterior depends only on the choice of prior (assuming appropriate independence relationships among components of the prior), and in the frequentist case that does not affect goodness-of-fit to the observed data. Second, while various forms of this are known in various sub-fields, to variable degrees, we emphasize the great similarity between these fields and settings; appropriate review of a number of selected areas is presented to facilitate study of the common features. We illustrate this by showing how non-identified parts can be replaced arbitrarily, without altering the fit to the observed data but with potentially non-trivial consequences for inferences and substantive conclusions. It should be clear that this can be dangerous and the user must carefully reflect on the arbitrary components. For example, they should be supported by substantive considerations or be made part of a sensitivity analysis. Therefore, acceptable goodness-of-fit to the observed data cannot be used as the sole justification for the analysis. In the absence of external corroborating knowledge or information, two alternative routes can be followed. First, it can be made clear that the conclusions drawn have meaning only under the external assumptions built into the analysis. For example, a researcher can choose to draw inferences given a set of scientifically plausible but otherwise non-verifiable causal relationships. It is then important not to divorce the data analysis from the assumptions made. Second, an appropriate sensitivity analysis can be conducted to augment the conclusions. By sensitivity analysis, we mean in this context, either a study of how unverifiable assumptions affect overall inferences, or an assessment of *traceability*[5,6], i.e., how unverifiable assumptions influence predictions for individual subjects. For example, analyses can be conducted under a number of alternative sets of hypothesized structures as well. This then allows the researcher to examine the sensitivity of the inferences concerning the scientific question to varying the underlying assumptions. See, for example, Part V of Molenberghs and Kenward[7].

The remainder of the paper is organized as follows. In the next section, we introduce seven illustrative examples. In Section 3, we introduce our general results concerning enriched data structures, in particular showing how components of the models can be chosen in an effectively infinite number of ways without affecting the fit to the observed data. In the subsequent sections, these general results are applied to five widely used settings, namely that of latent class models (Section 4), finite-mixture models (Section 5), factor analysis (Section 6), random effects models (Section 7), and incomplete data (Section 8), and practical implications are illustrated using the examples.

2 Motivating Data Sets

2.1 A Clinical Trial in Onychomycosis

The data introduced in this section were obtained from a randomized, double-blind, parallel group, multicentre study for the comparison of two oral treatments (in the sequel coded as A and B) for toenail dermatophyte onychomycosis (TDO)[8]. TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons[9]. Anti-fungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new such compounds, however, has reduced the treatment duration to 3 months. The aim of the present study was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A or with treatment B.

In total, 2×189 patients, distributed over 36 centres, were randomized. Subjects were followed during 12 weeks of treatment and followed further, up to a total of 48 weeks. Measurements were taken at baseline, every month during treatment, and every 12 weeks afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. We will restrict our analyses to only those patients for which the target nail was one of the two big toenails. This reduces our sample under consideration to 146 and 148 subjects, in group A and group B, respectively. One of the responses of interest was the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in mm. This outcome has been studied extensively in Verbeke and Molenberghs[10]. Figure 1 shows the observed profiles of 30 randomly selected subjects from treatment group A and treatment group B, respectively. In Table 1, the amount of missingness is brought to the forefront, by listing the number of repeated measures available per subject, for each of the two treatment arms separately. A linear mixed model will be considered, in which enrichment arises through the inclusion of random effects.

2.2 A Developmental Toxicity Study

This developmental toxicity study investigates the dose-response relationship in mice of the potentially hazardous chemical compound di(2-ethylhexyl)phthalate (DEHP), used in vacuum pumps[11] and as plasticizers for numerous plastic devices made of polyvinyl chloride. DEHP provides the finished plastic products with desirable flexibility and clarity[12]. It has been well documented that small

quantities of phthalic acid esters, of which DEHP is an instance, may leak out of polyvinyl chloride plastic containers in the presence of food, milk, blood, or various solvents. Due to their ubiquitous distribution and presence in human and animal tissues, considerable concern has developed as to the possible toxic effects of the phthalic acid esters. The developmental toxicity study, conducted in timed-pregnant mice during the period of major organogenesis and described by Tyl *et al*[13], has attracted much interest in the toxicity of DEHP. The doses selected for the study were 0, 0.025, 0.05, 0.1, and 0.15%, corresponding to a DEHP consumption of 0, 44, 91, 191, and 292 mg/kg/day, respectively. The dams were sacrificed, slightly prior to normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live fetuses were dissected from the uterus, anesthetized, and examined for external, visceral, and skeletal malformations, as well as for body weight. Our focus will be on the continuous weight outcome. Evidently, fetuses are clustered within mothers; hence the implied association needs to be accommodated in the analysis. When done through random effects, data enrichment arises. Summary data are presented in Table 2. Table 2 makes clear, when the number of viable fetuses (litter size) is compared to the number of implants, that there is a substantial amount of depletion and that it, not surprisingly, increases with dose.

2.3 The 2005 United States' National Youth Risk Behavior Survey data

This survey, conducted by the US Centers for Disease Control and Prevention, targets youths in grades 9–12, and the questions of interest are on various health-risk behaviors. These include alcohol and drug use, sexual behaviour, dietary habits, and physical activity. The Youth Risk Behavior Surveillance System aims at among others to monitor the trends of health-risk behaviour and to assess the impact of efforts to combat the same. We pay attention to 12 questions relating to smoking, alcohol consumption, consumption of other drugs, and sexual behaviour. Collins and Lanza[14] have previously extensively analyzed these variables, in the context of latent class models.

2.4 Accident Insurance Policies Data

Böhning[15] analyzes data on claims made by 9461 accident insurance policies issued by La Royale Belge Insurance Company. These data have been attributed to Thyrion[16]), and have also been used by Simar[17], as well as by Carlin and Louis[18]. The data are on the number of policies reporting a certain number of claims in a certain year. We use these data in the context of finite mixture models.

2.5 Data on Recurrent Asthma Attacks in Children

These data have also been used by Molenberghs *et al.*[19] and Duchateau and Janssen[20]. The setting is a prevention trial, where children, who are between 6 and 24 months, and who are at a high risk of developing asthma, are involved. They are randomized, before they experience Asthmatic attacks, to the study drug and placebo, and the attacks that occur are recorded. Since a patient will typically experience more than one event, there is clustering. Additionally, during the entire observation period, a patient will have different at risk times, separated by a period of attack or a period of no observation. We present part of the data in Table 4, in calendar-time format, where the time at risk is the time from the end of previous to the start of the next event. The end of each period will correspond to either an event or no event.

2.6 Time-to-insemination Data

These data are collected to assess factors associated with time to insemination in dairy heifer cows[21]. Dairy farmers aim for a calving interval between 12 and 13 months. The time from parturition to first insemination is a main factor determining this interval. Duchateau *et al.*[21] analyze data on the time-to-insemination for dairy cows, which were clustered within herds (farms). Some cows failed to get inseminated, and some were culled before insemination, thus there was censoring. We will focus on the covariate "parity," which is the number of times the cow has already calved, and which is dichotomized into "primiparous" and "multiparous" cows. Duchateau and Janssen[20] have also analyzed the data in terms of this covariate.

2.7 National Track Records for Women

Johnson and Wichern[22] present data on records for 7 women track events. For each of the seven events (100, 200, 400, 800, 1500, and 3000 metres, and the marathon). The record times are provided for $n = 54$ countries.

3 General Result About Counterparts in Enriched-data Structures

The result in this section is based upon Verbeke and Molenberghs[4]. Assume data $\boldsymbol{Z}_i$ for an independent unit $i = 1, \ldots, N$ are augmented with $\boldsymbol{c}_i$. The $\boldsymbol{c}_i$ can take any conventional enriched-data form. For example, the vector can refer to missing measurements, random effects, or perhaps a combination

of both. An example of a setting where the latter situation arises naturally is the shared-parameter framework, that will be considered in the next section.

Assume a joint model of the generic form $f(\boldsymbol{z}_i, \boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where covariates have been suppressed for notational simplicity. We assume the parameters to be disjoint, in the sense of Rubin[23], meaning that the parameter space of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ equals the set theoretic product of the individual parameter spaces. Consider the factorizations:

$$f(\boldsymbol{z}_i, \boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{z}_i|\boldsymbol{c}_i, \boldsymbol{\theta}) f(\boldsymbol{c}_i|\boldsymbol{\psi}), \quad (1)$$

$$= f(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) f(\boldsymbol{c}_i|\boldsymbol{z}_i, \boldsymbol{\theta}, \boldsymbol{\psi}). \quad (2)$$

Borrowing terminology from the hierarchical-models context, such as mixed models, which are given specific consideration in Section 7, every factor in both (1) and (2) can usefully be given a name. The left hand side is the *joint model*. Consider first the right hand sides. The first factor in (1) is the *hierarchical model* and the second one is the *prior density* for the enriched data. The first factor in (2) may be termed the *marginal model*, whereas the second one is the *posterior density* of the enriched data.

The above terminology makes clear the obvious link between (1)–(2) and the mixed-model setting. The link with incomplete data follows by setting $\boldsymbol{c}_i \equiv \boldsymbol{y}_i^m$ and $\boldsymbol{z}_i = (\boldsymbol{y}_i^o, \boldsymbol{r}_i)$.

These considerations immediately establish the following theorem.

Theorem 1 (A Family of Counterparts to a Given Model for Enriched Data.) *Let us assume that data $\boldsymbol{z}_i$ are enriched with $\boldsymbol{c}_i$. Then, any model (1) formulated for and fitted to such data, can be replaced by an infinite family of models, all retaining the fit to the observed data as achieved by the original model. This is done by preserving the marginal model $f(\boldsymbol{z}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$ and replacing the posterior density $f(\boldsymbol{c}_i|\boldsymbol{z}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$ by an arbitrary conditional density*

$$f(\boldsymbol{d}_i|\boldsymbol{z}_i, \boldsymbol{\gamma}). \quad (3)$$

Here, $\boldsymbol{d}_i$ rather than $\boldsymbol{c}_i$ is used to indicate that there need not be any connection between the original and substituted enriched data. Also, the new density (3) can be parameterized by a completely new parameter $\boldsymbol{\gamma}$.

While it might be argued that the conventional error term in an ordinary regression equation already is an instance of enrichment, we choose to view this as different from the theme of this paper. Broadly,

in a univariate regression context (encompassing linear regression, analysis of variance, regression based on generalized linear models, etc.) the response is split into signal and noise. While this surely depends on the posited model (e.g., a linear model with a certain mean function), it is verifiable from the data, using a realm of fit and diagnostic tools. Our situation of interest is different because of two aspects. First, in augmented-data settings, the noise is split into several sources of noise, a split which cannot be verified definitively from data. Second, in coarse-data settings, models describe the unobserved outcomes, given the observed ones, and predict the same; the models, by construction, are not verifiable from the data.

It may seem that the above derivations violate the so-called extended likelihood principle[24], which states that the extended likelihood $f(\boldsymbol{z}_i, \boldsymbol{c}_i)$ carries all information in the data about the unobservables. Of course, this is a very sensible principle to make inferences *given the posited model*. Our main point is not to take issue with the extended likelihood principle, but rather to demonstrate how models, coinciding in $f(\boldsymbol{z}_i)$ but differing in $f(\boldsymbol{c}_i|\boldsymbol{z}_i)$, are indistinguishable in terms of the data only. In contrast, the extended likelihood principle states that, *once a particular model has been chosen,* parametric inferences about the parameters governing the joint distribution, follow through the extended likelihood function.

4 Case I: Latent Classes and Latent Variables

Latent class (LC) models are widely used, especially in the social and behavioral sciences[25,26], where they are used to identify subgroups of individuals, based on phenomena defined in terms of categorical data. The observed variables are assumed to be a manifestation of some underlying categorical latent variable, the levels of which are believed to organize individuals into subgroups exhibiting distinct tendencies. In a latent variable (LV) model the unobservable is of a continuous nature. From a statistical perspective, use of the LC model may be viewed as a way of addressing heterogeneity among observations. A qualitative mixture distribution is assumed, and the observed, also called manifest or indicator, variables are assumed to be independent, conditional on the latent class. This is termed the local independence assumption.

When considered in a broad sense, LC and LV models exhibit connections with item-response theory models[27], shared-parameter models for incomplete data (Section 8), and factor analysis (Section 6). As in the previous cases, LC and LV models are based on unobservables. It is therefore impossible

to decide, in terms of the data alone, whether there are in fact such latent classes and, if we assume that there are, how many exist, and the number of categories in each. The 'identification' of the number of latent classes bears similarity to the identification of the number of components in mixture models (Section 5).

Suppose we observe response variables $Y_1, Y_2, \ldots, Y_T$, each with C_t categories, $t = 1, \ldots, T$, and assume a categorical latent variable, Z, with g levels. The basic latent class (LC) model takes the following form:

$$P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = P(Z = z) \prod_{t=1}^{T} P(Y_t = y_t | Z = z). \tag{4}$$

This is called the probabilistic representation of the model, the parameters of which are the conditional (item-response) probabilities: $P(Y_t = y_t | Z = z)$, and the latent class probabilities (prevalences): $P(Z = z)$. An equivalent representation of the basic LC model, called the log-linear representation, takes the form:

$$\log P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = \lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}. \tag{5}$$

The link between the conditional and log-linear model parameters is the following:

$$P(Y_t = y_t | Z = z) = \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{zy_t}^{ZY_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{zi}^{ZY_t}\right)}. \tag{6}$$

An iterative procedure, such as the Expectation-Maximization (EM) algorithm[28], is used for model estimation. The Akaike Information Criterion (AIC) and the likelihood ratio statistic (G^2) are typically used in evaluating the appropriate number of latent classes. The likelihood ratio statistic evaluates the proximity of the expected cell frequencies to the observed cell frequencies, whereas AIC adds penalty to this that depends on the number of parameters in the model. We now apply Theorem 1 to model (4), replacing the posterior distribution with two rather different choices: (a) the normal distribution and (b) a distribution corresponding to the posterior distribution of a model with $k \neq g$ latent classes.

We first set out the components in (1)–(2) for model (4). The joint model is simply the exponent of model (5), with expression

$$P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) = \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right). \tag{7}$$

From (6), and by the local independence assumption of the LC model, the hierarchical model is seen to take the form:

$$P(Y_1 = y_1, \ldots, Y_T = y_T | Z = z) = \prod_{t=1}^{T} P(Y_t = y_t | Z = z) = \prod_{t=1}^{T} \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{zy_t}^{ZY_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{zi}^{ZY_t}\right)} \tag{8}$$

and the prior distribution, the ratio of the joint to the hierarchical model, is:

$$P(Z = z) = \frac{\exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}{\prod_{t=1}^{T} \frac{\exp\left(\lambda_{y_t}^{Y_t} + \lambda_{zy_t}^{ZY_t}\right)}{\sum_{i=1}^{C_t} \exp\left(\lambda_i^{Y_t} + \lambda_{zi}^{ZY_t}\right)}}. \tag{9}$$

The marginal model is a weighted sum of the hierarchical probabilities:

$$P(Y_1 = y_1, \ldots, Y_T = y_T) = \sum_{z=1}^{g} P(Y_1 = y_1, \ldots, Y_T = y_T, Z = z) \tag{10}$$

and hence, the posterior distribution follows as:

$$P(Z = z | Y_1 = y_1, \ldots, Y_T = y_T) = \frac{\exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}{\sum_{z=1}^{g} \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}. \tag{11}$$

We now turn to each of the posteriors.

4.1 Normal Posterior

We retain the marginal model (10), but replace the sets of probabilities given in (11) with a unit-variance normal density and linear mean model:

$$f(h|\boldsymbol{Y}) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\left[h - (\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t)\right]^2}. \tag{12}$$

The new joint model follows as the product of (10) and (12), with now prior distribution

$$\frac{e^{\lambda}}{\sqrt{2\pi}} \sum_{y_1} \cdots \sum_{y_T} e^{\frac{1}{2}\left[h - (\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t)\right]^2} \sum_{z=1}^{g} \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right) \tag{13}$$

and hierarchical model

$$\frac{\frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\left[h - (\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t)\right]^2} \sum_{z=1}^{g} \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}{\frac{e^{\lambda}}{\sqrt{2\pi}} \sum_{y_1} \cdots \sum_{y_T} e^{\frac{1}{2}\left[h - (\alpha_0 + \sum_{t=1}^{T} \alpha_t Y_t)\right]^2} \sum_{z=1}^{g} \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^{T} \lambda_{y_t}^{Y_t} + \sum_{t=1}^{T} \lambda_{zy_t}^{ZY_t}\right)}. \tag{14}$$

We note here that (13) and (14) complete a new hierarchical specification, which gives the same marginal fit (10) achieved by the initial set (8) and (9). However, as will be clear from the data analysis below, there are consequences for ensuing inferences.

4.2 Distribution Corresponding to the Posterior of a Model With $k \neq g$ Latent Classes

We now couple (10) with

$$P(X = x|Y_1 = y_1, \ldots, Y_T = y_T) = \frac{\exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{x_t}^{X_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^k \exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{y_t}^{Y_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)}. \tag{15}$$

The prior distribution and the hierarchical model, respectively, can then be seen to take the following forms:

$$\begin{aligned} & \sum_{y_1} \cdots \sum_{y_T} \left[\sum_{z=1}^g \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^T \lambda_{y_t}^{Y_t} + \sum_{t=1}^T \lambda_{zy_t}^{ZY_t}\right)\right] \times \\ \times \quad & \frac{\exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{x_t}^{X_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^k \exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{y_t}^{Y_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)} \end{aligned} \tag{16}$$

and

$$\frac{f(y_1, \ldots, y_T)}{\sum_{y_1} \cdots \sum_{y_T} f(y_1, \ldots, y_T)}, \tag{17}$$

with

$$\begin{aligned} f(y_1, \ldots, y_T) \quad = \quad & \sum_{z=1}^g \exp\left(\lambda + \lambda_z^Z + \sum_{t=1}^T \lambda_{y_t}^{Y_t} + \sum_{t=1}^T \lambda_{zy_t}^{ZY_t}\right) \times \\ & \times \frac{\exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{x_t}^{X_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)}{\sum_{x=1}^k \exp\left(\beta + \beta_x^X + \sum_{t=1}^T \beta_{y_t}^{Y_t} + \sum_{t=1}^T \beta_{xy_t}^{XY_t}\right)}. \end{aligned}$$

Note that (16) and (17) complete yet another hierarchical specification, giving the same marginal fit (10).

4.3 Data Analysis

We now illustrate the above developments using the 2005 United States' National Youth Risk Behavior Survey data ($N = 13,840$), introduced in Section 2.3. We consider the 12 questions introduced earlier. Collins and Lanza[14] have previously extensively analyzed these variables, and chosen a 5-class LC model. We use the SAS procedure LCA, an add-on procedure in SAS, to re-analyze these data. In Table 5, we present the model's latent class prevalence and item-response probabilities.

These parameters are sufficient to calculate, by using Bayes' theorem, the posterior probability, where the latent classes and their corresponding probabilities act as (empirical) prior: the probability of belonging to a certain latent class, given a specific response pattern. Note that there is a total of

5×2^{12} posterior probabilities, where 5 represents the number of latent classes, and 2^{12} is the number of different response patterns possible. In practice, one needs to calculate only those that correspond to patterns actually occurring in the data. Classification of a respondent to any of the 5 latent classes, given his/her response pattern, is then based, for example, on the highest of the individual's set of 5 posterior probabilities: $P(Z = z|Y_1 = y_1, \ldots, Y_{12} = y_{12}), z = 1, \ldots, 5$. Alternatively, the set of posterior probabilities may be considered, which is especially instructive when a number of patterns differ only slightly in terms of posterior probability. As an example, consider the response pattern composed of a "Yes" response to all questions:

$$\begin{aligned} P(Z = z|Y_1 = \cdots = Y_{12} = \text{"Yes"}) &= \frac{P(Y_1 = \cdots = Y_{12} = \text{"Yes"}|Z = z)P(Z = z)}{P(Y_1 = \cdots = Y_{12} = \text{"Yes"})} \\ &= \frac{\prod_{t=1}^{12} P(Y_t = y_t|Z = z)P(Z = z)}{\sum_{z=1}^{5} P(Z = z)\left[\prod_{t=1}^{12} P(Y_t = y_t|Z = z)\right]}. \end{aligned}$$

Substituting the relevant parameters from Table 5, we obtain $(5.53E-18, 1.03E-08, 4.60E-06, 1.00, 7.33E-09)$ as the set of posterior probabilities, for latent classes 1–5 respectively. Clearly, classification of a respondent with such a response pattern would be to latent class 4, which, generally, has higher probabilities of a "Yes" response to the items than the other classes.

4.3.1 Normal Posterior

We now replace the posterior distribution with our first choice, the normal distribution. The parameters $\alpha_0, \ldots, \alpha_{12}$ play the role of sensitivity parameters; they can be freely specified, all without changing the marginal fit. Here, we set them to $\alpha_0 = \cdots = \alpha_{12} = 0.5$. Evidently, the concept of classifying a respondent to a particular latent class no longer exists. h in (9) is continuous, taking values on the whole real line, meaning that for any specific response pattern $(\boldsymbol{Y} = \boldsymbol{y})$, there exists an infinite collection of posterior densities. The prediction for the enrichment, $\widehat{h}$, is given as $E(h|\boldsymbol{Y}) = \alpha_0 + \sum_{t=1}^{12} \alpha_t y_t$. Letting y_t take the value 1 for a "Yes" response and 0 otherwise, we can calculate the prediction for h, for a specific response pattern. For example, for the response profile $(Y_1 = \cdots = Y_{12} = \text{"}Yes\text{"})$, mentioned earlier, $\widehat{h} = 0.5 + 12 \times 0.5 = 6.5$. The point we make is that having replaced the posterior distribution with our choice, we move to an entirely different setting, where, in contrast to the initial case where we could allocate to classes, we now work with a continuum. Once more, such manipulations are possible while the marginal fit remains unaffected.

4.3.2 Changing the Posterior With $k \neq 5$ Latent Classes

By choosing $k \neq 5$ in (9), we complete our choice. There is no information in the data about the β parameters; they can thus be specified freely. We only need to ensure that $\sum_x \beta_x^X = \sum_{y_1} \beta_{y_1}^{Y_1} = \cdots = \sum_{y_{12}} \beta_{y_{12}}^{Y_{12}} = \sum_{y_1} \beta_{y_1 x}^{Y_1 X} = \cdots = \sum_{y_{12}} \beta_{y_{12} x}^{Y_{12} X} = \sum_x \beta_{y_1 x}^{Y_1 X} = \cdots = \sum_x \beta_{y_{12} x}^{Y_{12} X} = 0$, so that the distribution is a genuine posterior from a latent class model. In so doing, we end up with completely different latent class allocations, though once more, the marginal fit remains the same.

5 Case II: Finite-mixture-model Component Membership

Finite mixture models[15] are often used to handle heterogeneity arising from the postulation of unknown sub-populations, which are treated as latent. We assume that the response variable X follows a finite mixture distribution, formalized as

$$f(x) = \sum_{j=1}^{g} \pi_j f_j(x), \tag{18}$$

π_j, $j = 1, \ldots, g$ being the mixing proportion, *i.e.*, the proportion of the j^{th} sub-population in the population, and $f_j(x)$, $j = 1, \ldots, g$ the component densities, characterized by the parameters $\lambda_1, \ldots, \lambda_j$, respectively. The π_j satisfy $0 < \pi_j \leq 1$ and $\sum_j \pi_j = 1$. Sub-population membership is considered a latent variable, Z, with a discrete distribution P with values λ_j and corresponding probabilities π_j, for $j = 1, \ldots, g$. Next, we specify all components in (1) and (2), then illustrate arbitrariness of the posterior distribution. The hierarchical model is

$$f(x|Z = z, z = 1, \ldots, g) = f(\lambda_z), \tag{19}$$

with $f(\lambda_z)$ denoting the density characterized by the parameter λ_z. For instance, for a finite mixture of Poisson distributions, $f(x|Z = z) = f(\lambda_z)$ would be

$$\text{Poi}(\lambda_z). \tag{20}$$

We let

$$P(Z = z) = \pi_z, \tag{21}$$

be the prior distribution. The marginal distribution is obtained by summing-out Z:

$$f(x) = \sum_{z=1}^{g} f(x|Z = z)\pi_z. \tag{22}$$

A finite mixture of Poisson distributions, for instance, yields

$$f(x) = \pi_1 \cdot \text{Poi}(\lambda_1) + \cdots + \pi_g \cdot \text{Poi}(\lambda_g). \tag{23}$$

The joint model takes the form

$$f(x, z) = f(x|Z = z)\pi_z. \tag{24}$$

We therefore have that the posterior distribution, the ratio of (24) to (22), takes the form

$$P(Z = \ell, \ell = 1, \ldots, g|x) = \frac{f(x|Z = l) \cdot \pi_\ell}{\sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z}. \tag{25}$$

This expression provides a channel through which data are a posteriori classified into the various sub-populations. A datum is classified into the sub-population for which $P(Z = l|x)$ is maximal. To illustrate sensitivity, we proceed as follows: retain the marginal model (22) but arbitrarily alter the posterior distribution (25). We replace the sets of probabilities in (25) by a continuous distribution,

$$f(g|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2}, \tag{26}$$

$\mu(x) = \gamma x$. We note here that the data contains no information about γ, and we will have the liberty to set it to some value. The new joint model follows as the product of (22) and (26):

$$f(g, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z. \tag{27}$$

For instance, for the Poisson mixture, we have

$$f(g, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} [\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j)]. \tag{28}$$

The prior distribution follows by integrating or summing over X, depending on whether it is continuous or discrete. For discrete X,

$$f(g) = \sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z. \tag{29}$$

For the Poisson mixture, where X is of course discrete, we have

$$f(g) = \sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \left[\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j) \right]. \tag{30}$$

The hierarchical distribution follows as the ratio of (27) to (29):

$$f(x|g) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z}{\sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \sum_{z=1}^{g} f(x|Z = z) \cdot \pi_z}. \tag{31}$$

For the Poisson mixture case,

$$f(x|g) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x))]^2} \left[\sum_{j=1}^{g} \pi_j \text{Poi}(\lambda_j)\right])}{\sum_x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[g-\mu(x)]^2} \left[\sum_{j=1}^{g} \pi_j \cdot \text{Poi}(\lambda_j)\right]}. \tag{32}$$

Thus, finally, (19) and (21) on the one hand, and (31) and (29) on the other, are two different hierarchical specifications, yielding the same marginal model, (22). Once more we have two models that are indistinguishable in terms of fit to the observed data, while the ensuing inferences are sensitive to the particular hierarchical formulation chosen. In the first formulation, it is possible to attribute *a posteriori* component membership to a given datum, through classification, based on (25). In the second formulation, however, this concept of classification disappears, because the posterior consists of a continuous density, naturally leading to prediction of the value of the (now continuous) latent variable, g, by noting that $E(g|x) = \gamma x$.

5.1 Data Analysis

The above developments are illustrated using the Accident Insurance Policies Data, introduced in Section 2.4. Böhning[15] employed the Non-parametric Maximum Likelihood Estimation method, as implemented in the package C.A.MAn (Computer-Assisted Analysis of Mixtures and Applications), to fit a finite mixture of Poisson distributions to these data, and reaches a three-component solution. We re-analyze these data and use the analysis to illustrate our result. The following model for X, the number of claims, is found:

$$f(x) = 0.4184\text{Poi}(0) + 0.5730\text{Poi}(0.3356) + 0.0087\text{Poi}(2.5454). \tag{33}$$

With this result, (25) can be used to allocate a specific datum, corresponding to x claims, to any of the mixture model components $z = 1, 2, 3$. For instance, for $x = 2$ counts, the set of probabilities $P(Z\ell|x)$, $\ell = 1, 2, 3$, is easily found to be $(0.0000, 0.9125, 0.0875)$. Based on the maximal posterior allocation criterion, such a datum would be allocated to component 2. On the other hand, for $x = 5$ counts, the set would be $(0.0000, 0.0234, 0.9766)$, in which case the datum would be allocated to the third component. Of course, this is a clear situation; in cases where there is not a clear winner among the three component, presenting all three would be more insightful. No need to add, though, that this does not remove the enrichment aspect of the problem.

We now move to our second hierarchical formulation, where we assume a normal posterior. Fix the parameter γ to 0.5. Given the continuous nature, the action parallel to the above mixture component

membership is to compute the predicted value for the latent variable, g. For $x = 2$ and $x = 5$ claims, respectively, we obtain 1 and 2.5, respectively.

In keeping with the theme of this paper, the choice between these very different routes is not possible in terms of the observed data. Rather, a researcher must carefully consider the substantive knowledge available, together with the scientific goal of the analysis.

6 Case III: Factor Analysis

In the introduction, we referred to reification as a typical consequence of naively using methods that combine data with external information, through unobservables. A very early context in this respect was the debate regarding general intelligence, based on different but, in terms of the data alone, indistinguishable forms of factor analysis. We now consider the factor-analytic case from a technical perspective.

We consider the following factor-analytic model:

$$Y_j - \mu_j = \sum_{m=1}^{k} \ell_{jm} F_m + \varepsilon_j, \tag{34}$$

$(j = 1, \ldots, p)$, where Y_j is a continuous response variable, with mean μ_j. The variable F_j is a latent continuous variable, called factor, and ε_j are errors. The coefficients ℓ_{jm} are called factor loadings. In matrix notation, the model is $\boldsymbol{Y} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}$. In line with convention, we make the following assumptions: $E(\boldsymbol{F}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\text{cov}(\boldsymbol{F}) = \boldsymbol{I}$ (the assumption of uncorrelated factors), and $\text{cov}(\boldsymbol{\varepsilon}, \boldsymbol{F}) = \boldsymbol{0}$. We make the distributional assumptions that $\boldsymbol{F}$ has a multivariate normal distributional with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{I}_k$, and that $\boldsymbol{\varepsilon}$ also has a multivariate normal distribution, with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Psi}$. The response vector for an individual is enriched with a vector of factors. We therefore first set out all the components in (1)–(2) for model (34). The prior distribution is:

$$\boldsymbol{F} \sim N(\boldsymbol{0}, \boldsymbol{I}_k). \tag{35}$$

The marginal distribution is readily shown to be:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}). \tag{36}$$

The joint distribution, when joint normality of $\boldsymbol{Y}$ and $\boldsymbol{F}$ is assumed, can also be represented as:

$(\boldsymbol{Y}', \boldsymbol{F}')'$ with mean vector $(\boldsymbol{\mu}', \boldsymbol{0}')'$ and covariance matrix

$$\begin{pmatrix} \boldsymbol{LL'} + \boldsymbol{\Psi} & \boldsymbol{L} \\ \boldsymbol{L'} & \boldsymbol{I}_k \end{pmatrix}.$$

Finally, by the conditional distribution property of subsets of multivariate normal distributions, the hierarchical and the posterior distributions, respectively, are

$$\boldsymbol{Y}|\boldsymbol{F} \sim N[\boldsymbol{\mu} + (\boldsymbol{LL'} + \boldsymbol{\Psi})\boldsymbol{I}_k(\boldsymbol{f} - \boldsymbol{\mu}_f), \boldsymbol{\Psi}], \tag{37}$$

$$\boldsymbol{F}|\boldsymbol{Y} \sim N[\boldsymbol{L}(\boldsymbol{LL'} + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{I}_k - \boldsymbol{L}(\boldsymbol{LL'} + \boldsymbol{\Psi})^{-1}\boldsymbol{L}]. \tag{38}$$

The mean of the posterior distribution provides the predictive distribution of the enrichment, given the data. This is ordinarily used in the estimation of factor scores, where the vector of factor scores for the i^{th} individual, $i = 1, \ldots, n$, is given by $\widehat{\boldsymbol{f}}_i = \widehat{\boldsymbol{L}}(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L'}} + \widehat{\boldsymbol{\Psi}})^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_y)$.

To illustrate arbitrariness of the posterior distribution, and the attendant consequences, we retain the marginal model (36) whilst replacing the posterior distribution (38). First note that $\boldsymbol{L}$, in the posterior is a $p \times k$ matrix. A particular way, therefore, to change the posterior distribution, is to change $\boldsymbol{L}$ to $\boldsymbol{L}_1$, with $\boldsymbol{L}_1$ being a $p \times k'$ matrix, $k \neq k'$. The new posterior, therefore, becomes

$$\boldsymbol{G}|\boldsymbol{Y} \sim N[\boldsymbol{L}_1(\boldsymbol{L}_1\boldsymbol{L}_1' + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \mathbf{I}_{k_1} - \boldsymbol{L}_1(\boldsymbol{L}_1\boldsymbol{L}_1' + \boldsymbol{\Psi})^{-1}\boldsymbol{L}_1]. \tag{39}$$

This corresponds to the posterior of a factor-analytic model with a different number of factors than the initial model (34). The new joint model follows as the product of (36) and (39):

$$f(\boldsymbol{Y}, \boldsymbol{G}) = 2\pi^{-(\frac{p+k_1}{2})}|\boldsymbol{\Sigma}_1|^{\frac{-1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}, \tag{40}$$

where $\boldsymbol{\Sigma}_1 = \text{cov}(\boldsymbol{Y})$, $\boldsymbol{\Sigma}_2 = \text{cov}(\boldsymbol{g}|\boldsymbol{Y})$, and $\boldsymbol{\mu}_{g|y} = E(\boldsymbol{g}|\boldsymbol{Y})$, components which are all described above. The new prior distribution follows as

$$\int_y 2\pi^{-(\frac{p+k_1}{2})}|\boldsymbol{\Sigma}_1|^{\frac{-1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}d\boldsymbol{y}. \tag{41}$$

The new hierarchical model is therefore

$$\frac{2\pi^{-\left(\frac{p+k_1}{2}\right])}|\boldsymbol{\Sigma}_1|^{\frac{-1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}}{\int_y 2\pi^{-\left(\frac{p+k_1}{2}\right])}|\boldsymbol{\Sigma}_1|^{\frac{-1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{-1}{2}}e^{\frac{-1}{2}\left[(\boldsymbol{y}-\boldsymbol{\mu}_y)'|\boldsymbol{\Sigma}_1|^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y)+(\boldsymbol{g}\boldsymbol{\mu}_{g|y})'|\boldsymbol{\Sigma}_2|^{-1}(\boldsymbol{g}-\boldsymbol{\mu}_{g|y})\right]}d\boldsymbol{y}.} \tag{42}$$

We note that (41) and (42) complete a new hierarchical formulation, which produces the same marginal model, hence with the same marginal fit, as that from the initial formulation composed of

(35) and (37). At the same time, however, important inferences ensuing from the two formulations are totally different. In particular, the estimation of factor scores, which uses the predictive distribution of the enrichment given the data, will be sensitive to the full model formulation. From the $(1 \times k)$ vectors of factor scores for the respondents, which would be the outcome of the initial formulation, we move to very different $(1 \times k')$ sets of factor scores, resulting from the new formulation. In addition, whereas ranking of individual respondents would be with respect to k components in the first formulation, it would be with respect to an arbitrary k', in the second. Vindication of any one formulation can only come through independent substantive information. Our illustration reiterates the fact that there is a completely unidentifiable part of the model. Therefore, we can view Gould's argument expressed through indeterminacy of the axis rotation, in the same way as the arbitrariness of the posterior in our enrichment terms. We emphasize that the enrichment view merely presents the well-known result about indeterminacy of factor rotation in a broader framework, underlining the commonality with other data-enrichment settings. Thus, also here, it follows that a good working knowledge of the difference between what can be learned from the data and what is identifiable through assumptions only, is a necessary part of the appropriate use of factor analysis.

6.1 Data Analysis

We analyze the track record data, described in Section 2.7, converting the record times into speed (in metres/second). We fit a 2-factor analysis model, using the maximum likelihood method, as implemented in SAS Version 9.2. In Table 6, the factor-loading pattern for the rotated solution is presented. Factor 1 loads rather highly on the distances from 800 metres to the marathon, while factor 2 loads highly on the distances from 100 metres to 400 metres. We may therefore deem factor 1 to represent the middle and long-distance events, with factor 2 representing the short-distance events. We now turn to the factor scores, and, indeed, consider ranking of the countries involved, based on the respective factors. The U.S.A., Germany, the Czech Republic, France, and Russia complete the list of the top 5 countries with respect to the short-distance factor, while Kenya, Ireland, China, North Korea, and Norway top the middle-and-long-distance factor. Note that, for each country, there is a 1×2 vector of factor scores. As described earlier, by arbitrarily replacing the number of columns in the matrix $\boldsymbol{L}$, in the posterior, from 2 to, say, 4, and, of course, leaving the marginal model unaltered, we would end up with, for each country, a 1×4 vector of factor scores. We also note that in that case, we would also have complete freedom to arbitrarily specify the

parameters in the now 7×4 matrix $\boldsymbol{L}_1$, in the new posterior, because the data carry no information about them. In view of these developments, the ranking that we initially conducted, based on the 2 factors, would completely change. Indeed, we would now be considering ranking based on completely different 1×4 vectors for the countries. Neither formulation is self-evidently appropriate and only independent substantive information can allow us to distinguish between them.

Thus, rather than extracting additional insight out of the data, our analysis shows that one has to be very aware of the arbitrary nature of at least some part of the conclusions.

7 Case IV: Random Effects Models

In Section 7.1, the linear mixed model will be considered for illustration. In Section A, the special but important case of clustered data will be considered, with constant mean within clusters and compound-symmetry variance-covariance structure.

7.1 Case IVA: The Standard Linear Mixed-effects Model

In line with Verbeke and Molenberghs[10], we consider the linear mixed-effects model, in all components featuring in (1)–(2), and then apply Theorem 1 to replace the posterior density of the random effects, ordinarily normal, by two versions of the exponential density.

7.1.1 Standard Formulation of the Linear Mixed Model

Using notation as in Section 8, the fully hierarchically specified linear mixed-effects model takes the form[10]:

$$\boldsymbol{Y}_i|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i), \tag{43}$$

$$\boldsymbol{b_i} \sim N(0, D), \tag{44}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, and X_i and Z_i are design matrices.

Based on (43) and (44), the marginal model and posterior distribution of the random effects can be derived:

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, V_i = Z_iDZ_i' + \Sigma_i), \tag{45}$$

$$\boldsymbol{b_i}|\boldsymbol{Y}_i \sim N[DZ_i'V_i^{-1}(\boldsymbol{Y}_i - X_i\boldsymbol{\beta}), (Z_i'\Sigma_i^{-1}Z_i + D^{-1})^{-1}]. \tag{46}$$

It is useful to present as well the empirical Bayes predictions[10,18]. For the random effects, these follow in a straightforward fashion as the mean of (46), i.e.,

$$\widehat{\boldsymbol{b_i}} = E(\boldsymbol{b_i}|\boldsymbol{Y}_i) = DZ_i'V_i^{-1}(\boldsymbol{Y}_i - X_i\boldsymbol{\beta}). \tag{47}$$

For the prediction of outcome $\boldsymbol{Y}_i$, the value in (47) is plugged into the mean of the hierarchical model (43):

$$\widehat{\boldsymbol{Y}_i} = (Z_iDZ_i') \cdot V_i^{-1}\boldsymbol{y}_i + (\Sigma_i) \cdot V_i^{-1}X_i\boldsymbol{\beta}, \tag{48}$$

the familiar "weighted average" of the observed outcomes $\boldsymbol{y}_i$ and the marginal mean $X_i\boldsymbol{\beta}$.

7.1.2 A First Normal-exponential Version of the Linear Mixed Model

To illustrate the arbitrariness of the posterior density, brought forward by Theorem 1 and in this case referring to the posterior density of the random effects, let us replace the normally distributed random effects by a vector of n_i independent gamma random effects, where each outcome component Y_{ij} is paired with a gamma random effect g_{ij}. The conventional density for a gamma variable ϕ is

$$f(\phi) = [\beta_*^{\alpha_*}\Gamma(\alpha_*)]^{-1}\,\phi^{\alpha_*-1}e^{-\phi/\beta_*}, \tag{49}$$

with $\alpha_*, \beta_* \geq 0$ parameters. For convenience, let us set $\alpha_* = 1$ and $\delta = 1/\beta_*$ in (49), producing

$$f(\phi) = \delta e^{-\phi\delta}, \tag{50}$$

which is the exponential density special case of the gamma family. Clearly, the mean of ϕ then is $E(\phi) = \delta^{-1}$. Note that the choice for an exponential distribution here is not aimed at proposing a viable model for data analysis. The choice is made to illustrate Theorem 1, in such a way that reasonably tractable closed-form solutions can be obtained, at the same time allowing for choice within the exponential framework. Indeed, the choice to be made next can be juxtaposed with the one of Section 7.1.3.

Our first choice is completed by choosing a conditional density of the form (50) for $\phi = g_{ij}$, with $\delta = \gamma_j y_{ij}$, where γ_j is an unspecified parameter. The marginal model (45) is retained and coupled with the posterior:

$$f(\boldsymbol{g}_i|\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \gamma_j y_{ij} e^{-g_{ij}\gamma_j y_{ij}}. \tag{51}$$

The joint density of $\boldsymbol{y}_i$ and $\boldsymbol{g}_i$ obviously follows as the product of the density corresponding to (45) and density (51), and hence, after some algebra, the hierarchical model and prior can be seen to take the forms:

$$f(\boldsymbol{g}_i) = \left(\prod_{j=1}^{n_i}\gamma_j\right) e^{\boldsymbol{\mu}_i'\boldsymbol{\theta}_i+\frac{1}{2}\boldsymbol{\theta}_i'V_i\boldsymbol{\theta}_i} M_{n_i}(\boldsymbol{\mu}_i+V_i\boldsymbol{\theta}_i,V_i), \tag{52}$$

$$f(\boldsymbol{y}_i|\boldsymbol{g}_i) = \frac{\left(\prod_{j=1}^{n_i} y_{ij}\right) e^{\boldsymbol{\theta}_i'(\boldsymbol{y}_i-\boldsymbol{\mu}_i)} e^{-\frac{1}{2}\left[(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)+\boldsymbol{\theta}_i'V_i\boldsymbol{\theta}_i\right]}}{(2\pi)^{n_i/2}|V_i|^{1/2}M_{n_i}(\boldsymbol{\mu}_i+V_i\boldsymbol{\theta}_i,V_i)}, \tag{53}$$

where $\boldsymbol{\mu}_i = X_i\boldsymbol{\beta}$, $\boldsymbol{\theta}_i$ has components $\theta_{ij} = -g_{ij}\gamma_j$, and $M_n(\boldsymbol{k},V) = E(Y_1\ldots Y_n;\boldsymbol{k},V)$, i.e., the sole nth order moment, relative to a normal distribution with mean $\boldsymbol{k}$ and variance V, where each component occurs exactly once. From Willink[29] it follows that a simple recursive relationship can be used, based on the concept of Hermite polynomials, to calculate such moments:

$$M_n(\boldsymbol{k},V) = k_nM_{n-1}(\boldsymbol{k},V) + \sum_{j=1}^{n-1} v_{jn}M_{1\ldots j-1,j+1\ldots n-1}(\boldsymbol{k},V),$$

where the last term is an $(n-2)$th order moment, with both the jth and nth components left out; k_j is the jth element of the vector $\boldsymbol{k}$ and v_{jn} is the (j,n)th entry of the matrix V.

The empirical Bayes predictions take the form:

$$\widehat{g_{ij}} = 1/(\gamma_j y_{ij}), \tag{54}$$

$$\widehat{\boldsymbol{y}_i} = \frac{\boldsymbol{P}_{n_i}(\boldsymbol{\mu}_i - V_i\boldsymbol{z}_i, V_i)}{M_{n_i}(\boldsymbol{\mu}_i - V_i\boldsymbol{z}_i, V_i)}, \tag{55}$$

where $\boldsymbol{P}_{n_i}(\boldsymbol{\mu}_i - V_i\boldsymbol{z}_i, V_i)$ is an n_i-dimensional vector with components defined by:

$$P_{nj}(\boldsymbol{k},V_i) = E(Y_1\ldots Y_{i,j-1}Y_{ij}^2Y_{i,j+1}\ldots Y_n;\boldsymbol{k},V). \tag{56}$$

Also here, the following recursive relationship is useful to calculate the components of (56)[29]:

$$\begin{aligned} P_{nj}(\boldsymbol{k},V) &= k_jM_n(\boldsymbol{k},V) + \sum_{k\neq j} v_{jk}E(Y_1\ldots Y_{i,j-1}Y_{ij}^2Y_{i,j+1}\ldots Y_{i,k-1}Y_{i,k+1}\ldots Y_n) \\ &\quad + v_{jj}E(Y_1\ldots Y_{i,j-1}Y_{i,j+1}\ldots Y_n). \end{aligned}$$

Finally, $\boldsymbol{z}_i$ is a vector with components $z_{ij} = 1/y_{ij}$.

There is an obvious consequence resulting from these developments regarding the meaning of model parameters. In specifying the original hierarchical model (43)–(44), the parameters $\boldsymbol{\beta}$, Σ_i, and D in

general, but D in particular, are part of a hierarchical specification. Since (45)–(46) taken together are equivalent to the original pair of equations, one might argue that the hierarchical interpretation still holds. The difference now is that all three sets of parameters occur in each of the two models, whereas in the original specification (43)–(44) there is a separation between $\boldsymbol{\beta}$ and Σ_i on the one hand and D on the other hand. However, it has been argued[10,30,31] that there is a fundamental difference in parameter interpretation, even to the point of bearing on the inferences made, when one solely considers the marginal model (45). This is clear when considering the model composed of (45) and, for example, either (51) or (57). Indeed, now all three parameters $\boldsymbol{\beta}$, Σ_i, and D feature in the marginal model only. The hierarchical parameters, γ_j in our particular instance, are completely separated from the marginal ones. This further implies that the so-called hierarchical parameter is estimable only because it also occurs in marginal model (45) for which, by definition, there is information in the data. Put differently, in the conventional hierarchical marginal model, all parameters are identifiable from marginal model (45), which is the only route by which the data convey information about these parameters. The model merely *appears* interpretable at a hierarchical, or enriched, level since (46) contains these, and only these parameters.

Note that the choice $\delta = \gamma_j y_{ij}$ is pragmatic, in the sense that δ should be non-negative. This is acceptable for a data set where the outcomes are sufficiently bounded away from zero, such as body length. However, it may be deemed less elegant, in which case it may make sense to square or exponentiate y_{ij}, motivating the following, alternative formulation.

If the Bayesian interpretation of the original model is maintained then $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$ is a conventional prior distribution, and arbitrariness pertains to the posterior distribution. While conventionally uncommon to specify the posterior first and then work back to the prior, it does help to illustrate the point that there is an observable and an unobservable part of the joint distribution. Also, it opens avenues for sensitivity analysis, as we will discuss further in Section 9.

7.1.3 A Second Normal-exponential Version of the Linear Mixed Model

We consider now an alternative choice for (50): $\delta = e^{\gamma_j y_{ij}}$. Straightforward algebra, thereby making use of the identity:

$$\prod_{j=1}^{n_i} e^{-q_{ij}e^{\gamma_j y_j}} = \sum_{m_1=0}^{\infty} \cdots \sum_{m_{n_i}=0}^{\infty} \frac{(-q_{i1})^{m_1} \cdots (-q_{in_i})^{m_{n_i}}}{m_1! \cdots m_{n_i}!} e^{m_1\gamma_1 y_{i1} + \cdots + m_{n_i}\gamma_{n_i} y_{in_i}},$$

leads to the following model equations, that are in the same order and with the same notation as in the first normal-exponential case:

$$f(\boldsymbol{q}_i|\boldsymbol{y}_i) = \prod_{j=1}^{n_i} e^{\gamma_j y_{ij}} e^{-q_{ij}e^{\gamma_j y_{ij}}}, \tag{57}$$

$$f(\boldsymbol{q}_i) = \sum_{\boldsymbol{m}} \left(\prod_{j=1}^{n_i} \frac{(-q_{ij})^{m_j}}{m_j!} \right) e^{\boldsymbol{\mu}_i'\boldsymbol{\lambda}_m + \frac{1}{2}\boldsymbol{\lambda}_m' V_i \boldsymbol{\lambda}_m}, \tag{58}$$

$$f(\boldsymbol{y}_i|\boldsymbol{q}_i) = \frac{\prod_{j=1}^{n_i} e^{\gamma_j y_{ij}} e^{-q_{ij}e^{\gamma_j y_{ij}}} e^{-\boldsymbol{\mu}_i'\boldsymbol{\lambda}_m - \frac{1}{2}\left[(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)+\boldsymbol{\lambda}_m' V_i \boldsymbol{\lambda}_m)\right]}}{(2\pi)^{n_i/2}|V_i|^{1/2}\sum_{\boldsymbol{m}} \left(\prod_{j=1}^{n_i} \frac{(-q_{ij})^{m_j}}{m_j!} \right)}, \tag{59}$$

$$\widehat{q_{ij}} = e^{-\gamma_j y_{ij}}, \tag{60}$$

$$\widehat{\boldsymbol{y}}_i = \frac{\sum_{\boldsymbol{m}} \left[\prod_{j=1}^{n_i} \frac{(-e^{-\gamma_j y_{ij}})^{m_j}}{m_j!} \right] e^{\boldsymbol{\mu}_i'\boldsymbol{\lambda}_m + \frac{1}{2}\boldsymbol{\lambda}_m' V_i \boldsymbol{\lambda}_m} (\boldsymbol{\mu}_i + V_i \boldsymbol{\lambda}_m)}{\sum_{\boldsymbol{m}} \left[\prod_{j=1}^{n_i} \frac{(-e^{-\gamma_j y_{ij}})^{m_j}}{m_j!} \right] e^{\boldsymbol{\mu}_i'\boldsymbol{\lambda}_m + \frac{1}{2}\boldsymbol{\lambda}_m' V_i \boldsymbol{\lambda}_m}}, \tag{61}$$

where $\boldsymbol{m}$ ranges over all non-negative integer vectors $\boldsymbol{m} = (m_1, \ldots, m_{n_i})$, and $\boldsymbol{\lambda}_m$ has components $\lambda_{mj} = (m_j + 1)\gamma_j$.

The specific but insightful case of exchangeable data with compound-symmetry covariance structure can be found in Appendix A.

7.1.4 Analysis of the Toenail Data

For the unaffected nail length, we specify a linear mixed-effects model (43)–(44):

$$Y_{ij}|(b_{i0}, b_{i1}) \sim N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})t_j + \beta_2 T_i + \beta_3 T_i t_j, \sigma^2), \tag{62}$$

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix} \right], \tag{63}$$

where $T_i = 0$ if patient i received standard treatment and 1 for experimental therapy ($i = 1, \ldots, 298$). Further, t_j is the time at which the jth measurement is taken ($j = 1, \ldots, 7$). Parameter estimates and standard errors, obtained through maximum likelihood[10], are presented in Table 7.

We are now able to partially replace the model specified by (62)–(63) with the exponential-defined models. We choose, for illustration, the second exponential model of Section 7.1.3. This implies that the marginal model resulting from (62)–(63) is retained:

$$\boldsymbol{Y}_i \sim N[X_i(\beta_0, \beta_1, \beta_2, \beta_3)', \sigma^2 I_{n_i} + Z_i' D Z_i], \tag{64}$$

and coupled with (57). Here, X_i and Z_i are the obvious $n_i \times 4$ and $n_i \times 2$ design matrices, respectively. Then, we can calculate empirical Bayes predictions under both the normal and the second exponential model. These produce two different subject-specific profiles, in addition to the observed-data and marginal mean profiles. Note that, for the posterior density (57), we have the freedom of specifying the parameters γ_j, because there is no information contained in the data. Indeed, they can be identified by additional assumptions only; they play the role of sensitivity parameters. We set them equal to $\gamma_j = 0.05$. Figure 2 presents these four profiles for four selected subjects, two from each treatment arm, respectively. This is a way to assess tracebility of the unverifiable assumptions. Other instances will be given using the other datasets as well. It is clear that the exponential choice produces predictions that lie much closer to the marginal mean profile and further away from the observed profile, than is the case with the normal random effects.

In theory, one could estimate the parameters γ_j, but the point here is that one can freely vary the parameters specific to the posterior distribution of the random effects, without affecting the marginal fit, i.e., without affecting what is verifiable directly from the data. One might argue that the gamma-based posterior is uncommon and few practitioners would consider it as their first option. This does not take away the risk of proceeding by selecting one particular, convenient model, that then happens to produce one particular description of the unobservables. The motivation for this is usually no other than that it is convenient to use, implemented in standard software, and therefore in use by a large research community. A better way forward is then to surround any given analysis by a sensitivity analysis. This point is taken up in the Concluding Remarks.

7.1.5 Analysis of the Developmental Toxicity Study

We consider the following hierarchically specified, exchangeable model for the DEHP data, introduced in Section 2.2:

$$Y_{ij}|b_i \sim N(\beta_0 + b_i + \beta_1 x_i, \sigma^2), \tag{65}$$

coupled with (105). Here x_i is rescaled dose, in the sense that the DEHP consumption doses of 0, 44, 91, 191, and 292 mg/kg/day are replaced by unit-interval standardized values 0.0000, 0.1507, 0.3116, 0.6541, and 1.0000, respectively. Parameter estimates and standard errors are presented in Table 8.

Following the developments in Section A, model (65), combined with (105), can be replaced by, for

example, the models with exponential posterior distributions, described in Sections 7.1.2 and 7.1.3, respectively. This implies that the marginal model is retained, with

$$Y_{ij} \sim N(\beta_0 + \beta_1 x_i, \sigma^2 + d), \tag{66}$$

but with alternative posterior distributions, and hence EB estimates for the random effects and predictions, as presented by (118) and (119), respectively. The results are graphically depicted in Figure 3. For 11 selected clusters, spread over the various dose groups, the figure shows (1): observed average weight per cluster (2): the estimated marginal mean as given by (66); (3), (4), and (5): predictions following the normal, first, and second exponential models, respectively. We observe that, in line with the analysis of the toenail data, the exponential predictions lie closer to the marginal averages than is the case with the normal model.

7.2 Case IVB: Frailty Models for Repeated Survival Outcomes

Whereas for linear models and, more broadly, for generalized linear models, hierarchies are often accommodated using normal random effects, repeated survival data are frequently modeled using so-called frailty models[20], which are random effects models with, typically, random effects drawn from distributions other than the normal. A common choice is the gamma, combined with a Weibull model for the outcomes.

While such models are now well established, there are non-trivial implications for their use. For example, Molenberghs and Verbeke[32] showed that the marginal distribution, generated from a Weibull-gamma frailty model, is of log-logistic type and only has a finite number of finite moments. There are examples where not even the second and first moments would be finite. However, this is an issue that takes us beyond the arbitrariness described above for the linear mixed model case, an analogy of which for the Weibull-gamma case will be described next.

The term "frailty", and its use in survival data, has its roots in gerontology. In the latter field, it is used to indicate the increased mortality and morbidity risks of the more frail patients; in line with natural history, it is expected to increase with age. In statistics, it is taken to be constant within a patient in general statistical modeling and rather describe heterogeneity between patients. The introduction of random effects in survival data modelling dates back to Beard[33], who, in modeling mortality, introduced the random effect in a univariate setting, and called it the "longevity factor." Vaupel, Manton, and Stallard[34] on the other hand, in attempting to allow individual differences in

mortality hazard rates, introduced the random effect and termed it "frailty." In illustrating our general result on the arbitrariness of the posterior in frailty models, we focus on the parametric proportional hazards Weibull-Gamma frailty model:

$$h_{ij}(t|u) = h_0(t)u_i \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta}), \tag{67}$$

where $h_{ij}(t|u)$ is the hazard of the j^{th} individual from the i^{th} cluster, $h_0(t) = \lambda\rho t^{\rho-1}, \lambda > 0, \rho > 0$, $\boldsymbol{X}_{ij}^t$ is the covariates' vector, $\boldsymbol{\beta}$ is the fixed effects vector, and u_i is the frailty for cluster i. The frailty distribution is gamma, which, in this context, is normally taken such that its mean equals one and hence the one-parameter gamma distribution is used:

$$f(u) = \frac{\alpha^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\alpha u}. \tag{68}$$

We now spell-out, for (67), all components in (1) and (2). The prior distribution, $f(u)$, is, of course, (68). For the hierarchical distribution, using the fact that $f_{ij}(t) = h_{ij}(t)S_{ij}(t)$, where

$$S_{ij}(t) = \exp\left(-\int_0^t \lambda\rho s^{\rho-1} u_i \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta}) ds\right),$$

it follows that the event times, given the frailty, are Weibull distributed with parameters $\lambda u_i \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta})$ and ρ. It follows that the hierarchical distribution is

$$f(t|u) = \lambda\rho t^{\rho-1} u \exp\left[\boldsymbol{X}_{ij}^t \boldsymbol{\beta}) \exp(-\lambda u t^\rho \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta})\right]. \tag{69}$$

The marginal distribution, $f(t)$, given as $\int_u f(t|u)f(u)du$, is easily shown to be

$$\frac{\lambda \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta})\rho t^{\rho-1}\alpha^{\alpha+1}}{[\alpha + \lambda \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta}) t^\rho]^{\alpha+1}}. \tag{70}$$

The posterior distribution, $f(u|t)$, follows as the product of (69) and (68), divided by (70). Evaluating this gives the posterior as

$$\text{Gamma}\left(\alpha + 1, \frac{1}{\zeta_\rho + \alpha}\right), \tag{71}$$

where $\zeta_\rho \equiv \lambda \exp(\boldsymbol{X}_{ij}^t \boldsymbol{\beta}) t^\rho$. This implies that

$$E(u|t) = \frac{\alpha + 1}{\zeta_\rho + \alpha} = \widehat{u}_i. \tag{72}$$

Prior to arbitrarily replacing (71), we derive some further quantities, which are of particular relevance in survival data settings. The population survival function, $S_f(t)$, corresponding to (67), is evaluated as $\int_0^\infty S_{ij}(t) f_u(u) du$, giving $[(1 + \zeta_\rho \alpha^{-1})^\alpha]^{-1}$. This implies that the population hazard function is

$$h_f(t) = \frac{\zeta_{\rho-1}\rho}{1 + \zeta_\rho \alpha^{-1}}. \tag{73}$$

We now return to our arbitrary replacement of the posterior. Specifically, we replace the Gamma posterior (71) with a normal posterior with mean μ and variance σ^2, for which we choose $\sigma^2 = 1$ and $\mu = \mu(t) = \varphi \boldsymbol{X}_{ij}^t t$, $\boldsymbol{X}_{ij}$ being as defined earlier. We note here that the normal posterior implies that

$$E(x|t) = \varphi \boldsymbol{X}_{ij}^t t = \widehat{x}_i. \tag{74}$$

The new joint model follows as:

$$\frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}. \tag{75}$$

The new prior distribution follows as

$$\int_0^\infty \frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}dt. \tag{76}$$

Hence, the new hierarchical model is

$$\frac{\frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}}{\int_0^\infty \frac{\zeta_{\rho-1}\rho\alpha^{\alpha+1}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{[\alpha+\zeta_\rho]^{\alpha+1}\sqrt{2\pi}}dt}. \tag{77}$$

We note that (77) and (76) complete a new formulation, which, as with the initial formulation consisting of (69) and (68), defines the same marginal distribution for the event times, given by (70). Thus our new, admittedly contrived, model is indistinguishable from the original model in terms of fit to the data. Like in the other enrichment settings, though, the prediction of h_{ij} (the conditional hazard), through the mean of the hierarchical model, is different from what it was before. This difference will have an impact on inferences, without ability for the data to testify whether this or the original formulation is better or worse.

7.3 Data Analysis

7.3.1 Data on Recurrent Asthma Attacks in Children

We now illustrate our results using the recurrent asthmatic attacks in children' data (Section 2.5). Though various time-representations exist to analyze data of this type, we hereby assume that interest is on the event rate in calendar time, leading to the model given below. Furthermore, for purposes of illustration, we restrict ourselves to risk times which culminate in an event (asthmatic attack), *i.e.*, we only consider that subset of data not consisting of censored observations (we only consider data points for which the corresponding "end-of-observation" period corresponds to an attack). Consider

the following model

$$h_{ij}(t|u) = \begin{cases} h_0(t)u\exp(X_i^t\beta) & \text{if } \mathsf{y}_{ij1} \leq \mathsf{t} \leq \mathsf{y}_{ij2}; \\ 0 & \text{otherwise,} \end{cases} \tag{78}$$

where $u \sim$ Gamma$(\alpha, 1/\alpha)$ and h_{ij} denotes the hazard for the i^{th} child, time j. We note that in the model specification above, the subscript j has been dropped in denoting the drug covariate, X, since a given child is under either study drug or placebo at all time points. Further, β is the parameter corresponding to the drug effect, (y_{ij1}, y_{ij2}), $j = 1, \ldots, n_i$, denotes the pairs corresponding to the beginning and end of each risk period for child j, and t is the time since entry into the trial. We optimize the marginal likelihood using the R 2.11.1 software. Following Duchateau and Janssen[20], we convert time from days to months, to avoid convergence issues arising when λ is too small. Parameter estimates obtained are $\widehat{\lambda}$=0.2306 (s.e. 0.0234), $\widehat{\rho}$=1.2576 (s.e. 0.0309), $\widehat{\beta}$=-0.0159 (s.e. 0.0749) and $\widehat{\theta}$= 0.1606 (s.e. 0.0290). We now partially replace the model defined above, by retaining its resultant marginal model, and coupling it with a normal posterior. We set $\varphi = 0.5$, which is required because the data contain no information about this parameter. We then consider predictions for the conditional hazard under the two model formulations. In Figure 4, we present, for the study drug and placebo, the population and conditional hazard functions, for the models composed of marginal model (70) with each of the two different posterior specifications. We note that the two formulations give totally different predictions for the conditional hazard. The Gamma choice produces a prediction which lies much closer to the population hazard than the normal choice, which, clearly, produces a prediction which is very different. These disparate inferences occur in disturbing conjunction with an unaltered marginal model, but is, in line with all other illustrations in this paper and the general result spelled out in Section 3.

7.3.2 Time-to-insemination Data

This data set was introduced in Section 2.6. For our purposes, we restrict attention to event times and only consider that subset of data not consisting of censored observations. For these data, we also consider (78), with now h_{ij} the hazard for the j^{th} cow in the i^{th} herd, X_{ij}^t the parity covariate, and β the corresponding parameter. We convert time to months. Optimization of the marginal likelihood is done in R 2.11.1 software. Due to computational challenges, we use, for our model fitting, at most 20 cows in a herd. Parameter estimates obtained are $\widehat{\lambda}$=0.0569 (s.e. 0.0035), $\widehat{\rho}$=2.538 (s.e. 0.082), $\widehat{\beta}$=-0.2210 (s.e. 0.0331) and $\widehat{\theta}$= 0.3248 (s.e. 0.0394). We now partially replace the model defined

by (78), by retaining its resultant marginal model, and coupling it with a normal density. Again, we set $\varphi = 0.5$ and consider predictions for the conditional hazard under the two model formulations. In Figure 5, we present, for each parity category, the population and conditional hazard functions, for the models composed of marginal model (70) with each of the two different posterior specifications. Also here, we note that the two formulations give totally different predictions for the conditional hazard. The prediction based on the gamma choice is closer to the population hazard than is the case for the normal choice; this is similar to what was observed for the linear mixed model (Section 7.1.4).

8 Case V: Incomplete Data

Let the random variable Y_{ij} denote the response of interest, for the ith study subject, designed to be measured at occasions t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Independence across subjects is assumed. The outcomes can conveniently be grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. In addition, define a vector of missingness indicators $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$ with $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise.

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured and grouped in a vector $\boldsymbol{x}_i$ but, throughout, are suppressed from notation.

We now sketch the modeling frameworks (Section 8), present the definition of MAR in each one of them (Section 8.1), and then establish that every MNAR model can be doubled up with a MAR counterpart that preserves the fit to the observed data (Section 8.2).

The full density function can be factored in different ways, each leading to a different framework, already briefly mentioned in the introduction.

The *selection model* (SeM) framework is based on the following factorization[23,35]:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}). \tag{79}$$

The first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. As an alternative, one can consider so-called *pattern-mixture models* (PMM)[36,37] using the reversed factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{\psi}). \tag{80}$$

The name is intimately linked to Heckman's[38] selection model, which has been popular in econometrics for a third of a century. As we will underscore in what follows, one has to be very careful with the non-verifiable assumptions made by such models. The *shared-parameter model*[39--44] assumes a vector of random effects $\boldsymbol{b}_i$, shared between both processes, conditional upon which the measurement and missingness processes are independent, and often taking the form of random effects with a specific parametric distribution. This *shared-parameter model* (SPM) is formulated by way of the following factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\psi}), \tag{81}$$

and hence

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\psi}) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}. \tag{82}$$

For our purposes, we will need a slightly more general SPM formulation, as presented by Creemers *et al*[45]. Indeed, while most formulations assume that a single, common set $\boldsymbol{b_i}$ drives the entire process, one can expand $\boldsymbol{b_i}$ to a set of latent structures.

We will now move beyond the above standard concepts by enlarging the family of shared-parameter models, then zoom in on missingness at random and study the impact of our general result for this particular case.

Definition 1 (A General Shared-parameter Model Family.) *A general shared-parameter model is defined as one of the form*

$$f(\boldsymbol{y}_i^o|\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{j}_i, \boldsymbol{\ell}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i, \boldsymbol{m}_i) f(r_i|\boldsymbol{g}_i, \boldsymbol{j}_i, \boldsymbol{k}_i\, \boldsymbol{n}_i), \tag{83}$$

where $\boldsymbol{g}_i$, $\boldsymbol{h}_i$, $\boldsymbol{j}_i$, $\boldsymbol{k}_i$, $\boldsymbol{\ell}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{n}_i$ are independent random-effects vectors, vectors of latent variables, etc.

Here, $\boldsymbol{y}_i^o$ ($\boldsymbol{y}_i^m$) refers to the observed (missing) components for subject i. While fixed effects are allowed to accompany each of the random effects, they are suppressed from notation.

Several remarks are in place. First, this is the most general random-effects model that can be considered in the sense that $\boldsymbol{g}_i$ is common to all three factors in (83), $\boldsymbol{h}_i$, $\boldsymbol{j}_i$, and $\boldsymbol{k}_i$ are shared between pairs of factors, and $\boldsymbol{\ell}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{n}_i$ are restricted to a single factor. Depending on the application, one may choose to either retain all random effects or to omit some. For example, $\boldsymbol{j}_i$ is present in the first factor but not in the second, with the reverse holding for $\boldsymbol{k}_i$. Retaining these is

useful when it is deemed plausible that, at the time of dropout, the process governing the outcome is sufficiently altered so as to modulate the effects of $\boldsymbol{g}_i$ and $\boldsymbol{h}_i$, which are common to both. Note also that $\boldsymbol{m}_i$ is never identifiable from data but is introduced as the basis for sensitivity analysis. It is important to understand the consequences of such simplifications, preferably also in terms of the missing-data mechanism operating. This is why it is useful to establish conditions under which MAR operates on the one hand, and missingness does not depend on future, unobserved measurements in a longitudinal context on the other hand[46]. Second, in full generality, model (83) may come across as somewhat contrived. The objective of formulating Definition 1 is not to postulate (83) as a model for use in every possible application of SPM, but rather as the most general SPM from which substantively appropriate models follow as sub-classes. Related to this, it may seem that (83) assumes two completely different distributions for the outcome vector, i.e., divorcing the observed from the missing components. This is not entirely the case because $\boldsymbol{g}_i$ and $\boldsymbol{h}_i$ still tie both components together. The impact of $\boldsymbol{j}_i$, $\boldsymbol{k}_i$, $\boldsymbol{\ell}_i$, and $\boldsymbol{m}_i$ is to modify an individual's latent process in terms of missingness. In other words, the most general model assumes that observed and missing components are governed in part by common processes and partly by separate processes. Third, in principle, we could expand (83) with the densities of the random effects. This is generally not necessary for our purposes, though. Fourth, the assumption of independent random-effects vectors is not restrictive, because association is captured through the sets common to at least two factors. Fifth, a conventional SPM formulation follows by removing all random effects but $\boldsymbol{g}_i$. For convenience, write

$$\boldsymbol{b_i} = (\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{j}_i, \boldsymbol{k}_i, \boldsymbol{\ell}_i, \boldsymbol{m}_i, \boldsymbol{n}_i). \tag{84}$$

8.1 Defining Missing at Random

The taxonomy of missing-data mechanisms, introduced by Rubin[23] and informally described in the introduction, is customarily formalized using the second factor on the right hand side of (79): A mechanism is MAR if

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi}), \tag{85}$$

i.e., the missing-data mechanism depends on the observed outcomes but, given these, not further on the unobserved ones. In the MNAR case, missingness depends on the unobserved outcomes $\boldsymbol{y}_i^m$, regardless of the observed outcomes and the covariates.

Molenberghs *et al*[47,48], among others, formulated MAR in the PMM sett ing:

Theorem 2 (Missingness at Random in the Pattern-mixture Framework.) *In the PMM framework, the missing-data mechanism is MAR if and only if*

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i, \boldsymbol{\theta}) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\theta}). \tag{86}$$

This means that the predictive distribution in every pattern is equal, and hence also equal to the one averaged over all patterns. By predictive distribution, we mean the conditional distribution of the unobserved components given the observed ones. Put differently, prediction of the unobserved outcomes can be done merely using the observed ones with no further information coming from the missing-data mechanism. Note that, owing to this result, MAR can be formulated in terms of R given Y, but also in terms of Y given R.

Creemers *et al*[45] characterized MAR in the SPM framework:

Theorem 3 (Characterization of MAR in the General Shared-parameter Family.) *A member of the general SPM family (83) is MAR if and only if*

$$\frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{j}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i) f(\boldsymbol{r}_i|\boldsymbol{g}_i, \boldsymbol{j}_i, \boldsymbol{k}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}}{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i, \boldsymbol{j}_i) f(\boldsymbol{r}_i|\boldsymbol{g}_i, \boldsymbol{j}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}}$$

$$= \frac{\int f(\boldsymbol{y}_i^o|\boldsymbol{g}_i, \boldsymbol{h}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i) f(\boldsymbol{b_i})\, d\boldsymbol{b_i}}{f(\boldsymbol{y}_i^o)}. \tag{87}$$

Note that the random effects $\boldsymbol{\ell}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{n}_i$, pertaining to a single factor only, are suppressed from notation but are allowed to be present. Clearly, this result is not as intuitive as the SeM and PMM versions and, as such, the above result has little immediate data-analytic value. Therefore, fortunately, Creemers *et al.*[46] also showed that the following family satisfies the MAR property:

Definition 2 (A Sub-class of SPM Models.) *Define a sub-class of shared-parameter model (83):*

$$f(\boldsymbol{y}_i^o|\boldsymbol{j}_i, \boldsymbol{\ell}_i) f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{m}_i) f(r_i|\boldsymbol{j}_i, \boldsymbol{n}_i), \tag{88}$$

where $\boldsymbol{j}_i$, $\boldsymbol{\ell}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{n}_i$ are independent random-effects vectors.

At the same time, they established that there are members of the SPM family satisfying Theorem 3 but that are not of the (88) type. It is thus a proper sub-set. From their example of a model satisfying Theorem 3 but not belonging to the sub-class, one can infer that these typically would be rather contrived. Definition (88) has the advantage of having a clear, intuitive interpretation.

8.2 Every MNAR Model Has an MAR Counterpart

In this section, based on the argument of Molenberghs *et al*[48], we restate that for every MNAR model fitted to a set of data, there is a unique MAR counterpart providing exactly the same fit to the data. Whereas Molenberghs *et al*[48] confined attention to the missing-data setting, in the next section we will provide a much more general result, pertaining to all data-enriched structures.

The concept of model fit should be understood as being measured using such conventional methods as deviance measures, as applied to the observed data. The following steps are involved: (1) fitting an MNAR model to the data; (2) reformulating the fitted model in PMM form; (3) replacing the density or distribution of the unobserved measurements given the observed ones and given a particular response pattern by its MAR counterpart; (4) establishing that such an MAR counterpart uniquely exists.

In the first step, we fit an MNAR model to the observed set of data. The observed data likelihood equals

$$L = \prod_i \int f({\boldsymbol{y}_i}^o, {\boldsymbol{y}_i}^m, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d{\boldsymbol{y}_i}^m. \tag{89}$$

Upon denoting the obtained parameter estimates by $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$ respectively, the fit to the hypothetical full data is

$$f({\boldsymbol{y}_i}^o, {\boldsymbol{y}_i}^m, \boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f({\boldsymbol{y}_i}^o, {\boldsymbol{y}_i}^m | \widehat{\boldsymbol{\theta}}) f(\boldsymbol{r}_i | {\boldsymbol{y}_i}^o, {\boldsymbol{y}_i}^m, \widehat{\boldsymbol{\psi}}). \tag{90}$$

To undertake the second step, full density (90) can be re-expressed in PMM form as:

$$f({\boldsymbol{y}_i}^o, {\boldsymbol{y}_i}^m | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f({\boldsymbol{y}_i}^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}). \tag{91}$$

Note that the final term on the right hand side of (91), $f({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$, is not identified from the observed data. In this case, it is determined solely from modeling assumptions, the latter of which may or may not be inspired by substantive knowledge. Within the PMM framework, identifying restrictions have to be considered[37,47,49].

The third step requires replacing this factor by the appropriate MAR counterpart. Now, using Theorem 2, it is clear that $f({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$ needs to be replaced with

$$f^*({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o, \boldsymbol{r}_i) = f^*({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o) = f({\boldsymbol{y}_i}^m | {\boldsymbol{y}_i}^o, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}), \tag{92}$$

where the $f^*(\cdot)$ notation is used for shorthand purposes. Note that the density in (92) follows from the SeM-type marginal density of the complete data vector. Sometimes, therefore, it may be more

convenient to replace the notation ${\boldsymbol{y}_i}^o$ and ${\boldsymbol{y}_i}^m$ by one that explicitly indicates which components are observed and missing in pattern $\boldsymbol{r}_i$ under consideration:

$$f^*({\boldsymbol{y}_i}^m|{\boldsymbol{y}_i}^o, \boldsymbol{r}_i) = f^*({\boldsymbol{y}_i}^m|{\boldsymbol{y}_i}^o) = f[(y_{ij})_{r_j=0}|(y_{ij})_{r_j=1}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}]. \tag{93}$$

Thus, (93) provides a unique way of extending the model fit to the observed data, within the MAR family. As stated before, the above construction does not lead to a member of a conventional parametric family. While this obviously implies limitations on its use, its use is similar to the construction of some semi- and non-parametric estimators. Also, it helps to understand that an overall, definitive conclusion about the nature of the missing-data mechanism, solely based on the observed outcomes, is not possible, even though one can make progress if attention is confined to a given parametric family, in which one puts sufficiently strong prior belief[50]. Molenberghs *et al*[48] showed formally that the fit remains the same, leading to:

Theorem 4 (MAR Counterpart to MNAR Models.) *Every fit to the observed data, obtained from fitting an MNAR model to a set of incomplete data, is exactly reproducible from an MAR decomposition.*

The characterization of Theorem 3 allows us to construct an MAR counterpart to an arbitrary SPM of the form (83). It is necessary to (a) retain the fit of the model to the observed data, while (b) ensuring that (87) holds. This is easily done by *a-posteriori integrating* over the shared random effects in the densities describing the unobserved measurements, given the observed ones. Practically, integration takes place over the densities of $\boldsymbol{g}_i$, $\boldsymbol{h}_i$, and $\boldsymbol{k}_i$, where fitted parameters are plugged into the densities.

Theorem 5 (An MAR Counterpart to a General SPM.) *The MAR counterpart, to an arbitrary general SPM of the type (83) is found by replacing* $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i, \boldsymbol{m}_i)$ *with*

$$f^*(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{m}_i) = \int_{\boldsymbol{g}_i}\int_{\boldsymbol{h}_i}\int_{\boldsymbol{k}_i} f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i, \boldsymbol{m}_i) f(\boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i) d\boldsymbol{g}_i d\boldsymbol{h}_i d\boldsymbol{k}_i. \tag{94}$$

First, it is clear that this marginalization describes only the model-based prediction of the unobserved outcomes, given the observed ones. Hence, the choice for $f^*(\cdot)$ does not alter the fit. Second, observe that using $f^*(\cdot)$ in (87), instead of $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i, \boldsymbol{h}_i, \boldsymbol{k}_i, \boldsymbol{m}_i)$, of Theorem 3, reduces the equation to an identity, and hence the MAR condition is also satisfied. The importance of this result is that (94)

provides an MAR scenario for the missing-data mechanism, consistent with the previously achieved model fit.

Some comments are in place. Note that our general result follows quite easily in this case by observing that any missing-data model can be recast as a full PMM, as in (91). This framework readily allows the construction of an MAR substitute (92), which renders Theorem 4 almost trivial. Indeed, a PMM factors the joint distribution of observed measurements, missing measurements, and missing-data indicators such that the predictive distribution of what is unobserved, given what is observed, is an explicit factor in the model.

Note that the same feature is employed in Theorem 5, relative to the SPM. Indeed, also here the distribution of what is unobserved, given what is observed, is used. Two remarks are worth making.

First, the right hand side of (94) does not condition on $\boldsymbol{r}_i$, in spite of it being observed. Now, this absence is a key characteristic of SPM and therefore entirely logical.

Second, uniqueness results in the missing-data case come from the requirement that the counterpart is of MAR type. This can be relaxed by observing that, in (91), the factor $f({\boldsymbol{y}_i}^m|{\boldsymbol{y}_i}^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$ may be replaced by *any* valid density. This well-known result is: (1) placed in the broader context of enriched data; (2) also phrased in a shared-parameter context; (3) is illustrated in an insightful way.

8.3 Analysis of the Toenail Data

Consider a general model of the form (83), with random effects confined to $\boldsymbol{g}_i$, i.e., common to all three components. For the measurement model, assume a linear mixed model[10], with general form:

$$\boldsymbol{Y}_i|\boldsymbol{g}_i \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{g}_i, \Sigma_i), \tag{95}$$

$$\boldsymbol{g}_i \sim N(0, D). \tag{96}$$

Based on (95) and (96), the so-called marginal model can be derived

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{97}$$

Note that Section 7 focuses on random effects as such, whereas here the random effects play the role of shared parameters in the generalized shared-parameter model. To compute the model's prediction for the unobserved data, given the observed measurements, the corresponding density needs to be

derived. To this end, construct the conditional density, with obvious notation:

$$\begin{aligned} \boldsymbol{Y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{g}_i &\sim N\Big[(X_i^m - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1} X_i^o)\boldsymbol{\beta} + \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1}\boldsymbol{y}_i^o + (Z_i^m - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1} Z_i^o)\boldsymbol{g}_i, \\ &\qquad \Sigma_i^{mm} - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1}\Sigma_i^{om}\Big]. \end{aligned} \tag{98}$$

Now, (98) corresponds to the model as formulated, and will typically be of the MNAR type. To derive the MAR counterpart, we need to integrate over the random effect. With similar logic that leads to (97), now applied to (98), we obtain:

$$\begin{aligned} \boldsymbol{Y}_i^m|\boldsymbol{y}_i^o &\sim N\Big[(X_i^m - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1} X_i^o)\boldsymbol{\beta} + \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1}\boldsymbol{y}_i^o, \\ &\qquad (Z_i^m - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1} Z_i^o)D(Z_i^m - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1} Z_i^o)' \\ &\qquad +\Sigma_i^{mm} - \Sigma_i^{mo}\{\Sigma_i^{oo}\}^{-1}\Sigma_i^{om}\Big]. \end{aligned} \tag{99}$$

Hence, (99) is the MAR counterpart to (98). For the unaffected nail length, we choose for (95)–(96):

$$E(Y_{ij}|g_i, T_i, t_j, \boldsymbol{\beta}) = \beta_0 + g_i + \beta_1 T_i + \beta_2 t_j + \beta_3 T_i t_j, \tag{100}$$

$g_i \sim N(0, d)$, and $\Sigma_i = \sigma^2 I_7$, where I_7 is a 7×7 identity matrix. Further, $T_i = 0$ if patient i received standard treatment and 1 for experimental therapy ($i = 1, \ldots, 298$). Finally, t_j is the time at which the jth measurement is taken ($j = 1, \ldots, 7$).

Given these choices, (98) and (99) simplify to

$$\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o, g_i \sim N(X_i\boldsymbol{\beta} + Z_i^m g_i, \sigma^2 I_i), \tag{101}$$

$$\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o \sim N(X_i\boldsymbol{\beta}, dJ_i + \sigma^2 I_i), \tag{102}$$

with I_i an identity matrix and J_i a matrix of ones, with dimensions equal to the number of missing measurements for subject i. As a result of the conditional independence assumption, the simplification is dramatic.

Next, we formulate a model for the missingness mechanism in (83). The sequence $\boldsymbol{r}_i$ can take one of two forms in our case. Either, it is a length-7 vector of ones, for a completely observed subject, or it is a sequence of k ones followed by a sole zero $1 \leq k \leq 6$, for someone dropping out. Note that k is 1 at least, since for everyone the initial measurement has been observed. It is convenient to assume a logistic regression of the form:

$$\text{logit}\,[P(R_{ij} = 1|R_{i,j-1} = 0, g_i, T_i, t_j, \boldsymbol{\gamma})] = \gamma_0 + \gamma_{01} g_i + \gamma_1 T_i + \gamma_2 t_j + \gamma_3 T_i t_j, \tag{103}$$

($j > 1$), where γ_{01} is a scale factor for the shared random effect in the missingness model; forcing the variance in the measurement and dropout indicator sequences to be equal would make no sense. As a result, $\gamma_{01} g_i \sim N(0, \gamma_{01}^2 d)$.

Parameter estimates and standard errors are displayed in Table 9. Note that the scale factor γ_{01} has a negative estimate, even though it is not significant. While we should not overly stress its importance, there is some indication that a higher subject-specific profile of unaffected nail length corresponds to a lower dropout probability, which is not surprising. The magnitude of the scale factor allows us to 'translate' the subject-specific effect from the continuous outcome scale, expressed in mm, to the unit-less logit scale on which the probability of missingness is described. The random-intercept variance is highly significant among unaffected nail length outcomes; the same is not true for the dropout model, with $p = 0.2487$, using a $50 : 50$ mixture of a χ_0^2 and χ_1^2 distribution[10].

Figure 6 displays the incomplete profiles, extended beyond the time of dropout, using prediction based on: (1) the original model (dashed lines); (2) the MAR counterpart (solid lines). Within each of the treatment arms, three profiles are highlighted. The MAR counterpart reduces all predictions to the same profile, whereas the MNAR model predicts different evolutions for different subjects, implied by the presence of the random effect. The simple MAR-based prediction structure follows directly from the conditional independence assumption, present in (101). When deemed less plausible, the fully general structure (98) can be implemented. But the most important realization is that no distinction between both whatsoever is possible, based on the data.

9 Concluding Remarks

In this paper, we have used the unified framework of enriched data, encompassing coarse and augmented data, to bring out the common feature of unobservables, shared by all. The information required to identify such models is divided in that supplied by the data and that supplied externally, through assumptions and/or scientific knowledge. This implies that entire classes of models exist, coinciding in their description of the observed data, but different in their representation of the unobservables given the observed data.

For the data analyst, this means that every model in an enriched-data setting can be factored into a product of two components: the first one, termed the marginal model, fully identifiable from the observed data; the second one, the predictive distribution, entirely arbitrary. Failure to appreciate

this could result in grave errors. As a consequence, the conventional modeling route, consisting of formulating a model and judging its quality based on goodness-of-fit alone, is inadequate. This is because the inferences drawn and the assumptions made about the unobservables cannot be divorced. As we see it, there are therefore two alternative modes of analysis, that nevertheless fully exploit the information contained in the empirical data. In the first one, non-verifiable assumptions are based on substantive knowledge and/or statistical design. In the second one, sensitivity of the inferences to the non-verifiable assumptions are assessed formally.

As we have illustrated, this common issue is pertinent in a range of seemingly disparate settings, namely, incomplete data, random-effects and frailty models, latent classes, latent variables, factor analysis, and mixture models. Of course, various of these are interconnected and can be placed under the general umbrella of structural equations modeling[51]. Bringing it out has some value, we believe, as there are instances in the literature where it is missed[52].

We have not been exhaustive in our coverage of enrichment. Other areas include censored survival data, which is very similar to the incomplete-data case, grouped data, and situations where more than one type of enrichment occur simultaneously, such as, for example, incomplete data in random-effects models. Another major omission is that of methodology for causal inference. The issues raised here have been widely discussed in the appropriate literature. For example, Pearl[53] states: "Alternative causal models usually exist that make contradictory claims and, yet, possess identical statistical implications. Statistical test (sic) can be used for rejecting certain kernels, in the rare cases that such cases have testable implications, but the lion's share of supporting causal claims falls on the shoulders of untested causal assumptions." By kernel, Pearl refers broadly to a minimal set of assumptions required to identify the underlying causal model. Within the causal framework, we can include inferences drawn from randomized clinical trials[55]. In the incomplete-data setting sensitivity analysis is particularly well developed, for both the parametric and non-parametric settings[7,10,55,56,57]. For example, Creemers *et al*[45] showed how the generalized shared-parameter model for incomplete data can be used, not only to demonstrate that one cannot choose based on the data between MAR and MNAR, but also how it can be used as a vehicle for sensitivity analysis. Much work on sensitivity analysis can also be found in the causal literature[56,47]

Data sets and programs are available from the authors and through the journal's web pages.

Acknowledgment

The authors gratefully acknowledge support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy). We are grateful to Professor G. Opsomer for kindly making available the time-to-insemination data.

References

[1] Gould JS. The Mismeasure of Man. New York: WW Norton and Company; 1981.

[2] Heitjan DF Rubin DB. Ignorability and coarse data. Ann Stat 1991; 19:2244–53.

[3] Zhang J, Heitjan DF Impact of nonignorable coarsening on Bayesian inference. Biostatistics 2007; 8: 722–43.

[4] Verbeke G, Molenberghs G Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. Stat Mod 2010: 10: 391–419.

[5] Molenaar PCM A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. Measurement 2004: 2; 201–18.

[6] Molenaar PCM. On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. Dev Psychobiol 2008; 50: 60–9.

[7] Molenberghs G, Kenward MG Missing Data in Clinical Studies. Chichester: Wiley; 2007.

[8] De Backer M, De Keyser P, De Vroey C, Lesaffre E. A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day–a double-blind comparative trial. Br J Dermatol 1996; 124: 16–7.

[9] Roberts DT. Prevalence of dermatophyte onychomycosis in the United Kingdom: Results of an omnibus survey. Br J Dermatol 1992; 126 (Suppl 39), 23–7.

[10] Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York: Springer; 2000.

[11] Windholz M The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals. 10th ed. Rahway, NJ: Merck and Co; 1983.

[12] Shiota K, Chou, MJ, Nismimura, H. Embryotoxic effects of di-2-ethylhexyl phthalate (DEHP) and di-n-butyl phthalate (DBP) in mice. Environm Res 1980; 22: 245–53.

[13] Tyl RW, Price CJ, Marr MC, Kimmel CA. Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. Fund Appl Toxicol 1988; 10: 395–412.

[14] Collins LM, Lanza ST Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. New Jersey: Wiley; 2009.

[15] Böhning D. Computer-Assisted Analysis of Mixtures and Applications: meta-analysis, disease mapping, and others. Boca Raton: Chapman & Hall/CRC; 2000.

[16] Thyrion P Contribution à létude du bonus pour non sinistre en assurence automobile. ASTIN Bull 1960; 1: 142–62.

[17] Simar L. Maximum likelihood estimation of a compound Poisson process. Ann Stat 1976; 4: 1200–9.

[18] Carlin BP, Louis TA Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman & Hall; 1996.

[19] Molenberghs G, Verbeke G, DemétrioCGB, Vieira A. A family of generalized linear models for repeated measures with normal and conjugate random effects. Stat Sci 2010; 25: 325–347.

[20] Duchateau L, Janssen P *The Frailty Model*. New York: Springer; 2008.

[21] Duchateau L, Opsomer G, Dewulf J, Janssen P. The non-linear effect (determined by the penalised partial-likelihood approach) of milk-protein concentration on time to first insemination in Belgian dairy cows. Prevent Vet Med 2005; 68: 81–90.

[22] Johnson RA, Wichern DW Applied Multivariate Statistical Analysis. New Jersey: Pearson; 2007.

[23] Rubin DB Inference and missing data. *Biometrika* 1976: 63, 581–92.

[24] Bjørnstad JF. On the generalization of the likelihood function and likelihood principle. J Am Stat Assoc 1996; 91: 791–806.

[25] Goodman LA. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A Modified latent structure approach. Am J Sociol 1974; 79: 1179–259.

[26] Xu H, Craig BA A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. Biometrics 2009; 65, 1145–55.

[27] Tatsuoka KK. Toward an integration of item-response theory and cognitive error diagnosis. In: Frederiksen N, Glazer R, Lesgold A, Shafto MG, editors. Diagnostic monitoring of skill and knowledge acquisition Hillsdale, NJ: Lawrence Erlbaum Associates; 1990: p. 453–88.

[28] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 1977; 39: 1–38.

[29] Willink R. Normal moments and Hermite polynomials. Stat Prob Let 2005; 73, 271–75.

[30] Verbeke G, Molenberghs G. The use of score tests for inference on variance components. Biometrics 2003, 59: 254–62.

[31] Molenberghs G, Verbeke G. Likelihood ratio, score, and Wald tests in a constrained parameter space. Am Stat 2007; 61: 1–6.

[32] Molenberghs G, Verbeke G. On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. J Stat Planning Inference 2011; 141: 861–8.

[33] Beard RE. Note on some mathematical mortality models. In: Wolstenholme GEW, O'Connor M, editors. The lifespan of animals. Ciba colloquium on Aging. Brown: Boston,; 1959; p. 302–11.

[34] Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454.

[35] Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: Wiley; 2002.

[36] Little RJA. Pattern-mixture models for multivariate incomplete data. J Am Stat Assoc 1993; 88: 125–34.

[37] Little RJA. A class of pattern-mixture models for normal incomplete data. Biometrika 1994; 81: 471–83.

[38] Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Measurement 1976; 5: 475–92.

[39] Wu MC, Bailey KR. Analysing changes in the presence of informative right censoring caused by death and withdrawal. Stat Med 1988; 7: 337–46.

[40] Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. Biometrics 1989; 45: 939–955.

[41] Wu MC, Carroll RJ Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. Biometrics 1988; 44: 175–188.

[42] TenHave TR, Kunselman AR, Pulkstenis EP, Landis JR. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. Biometrics 1988; 54: 367–83.

[43] Follmann D, Wu MC. An approximate generalized linear model with random effects for informative missing data. Biometrics 1995; 51: 151–68.

[44] Little RJA. Modelling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995; 90: 1112–21.

[45] Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward MG. Generalized shared-parameter models and missingness at random. Stat Mod 2011; 11: 279–311.

[46] Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward MG. A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. Biometrical J 2011; 52: 111–25.

[47] Molenberghs G, Michiels B, Kenward MG, Diggle PJ. Monotone missing data and pattern-mixture models. Stat Neerl 1998; 52: 153–61.

[48] Molenberghs G, Beunckens C, Sotto C, and Kenward MG. Every missing not at random model has got a missing at random counterpart with equal fit. J Roy Stat Soc B 2008; 70: 371–88.

[49] Kenward MG, Molenberghs G, Thijs H. Pattern-mixture models with proper time dependence. Biometrika 2003; 90: 53–71.

[50] Jansen I, Hens N, Molenberghs G, Aerts M, Verbeke G, Kenward MG. The nature of sensitivity in missing not at random models. Comput Stat Data Analysis 2006; 50: 830–858.

[51] Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling. London: Chapman & Hall/CRC; 2004.

[52] Bollen K. *Structural Equations with Latent Variables*. New York: Wiley; 1989.

[53] Pearl J. An introduction to causal inference. Int J Biostat 2010; 6: 1–58.

[54] Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. J Educat Psychol 1974; 66: 688–701.

[55] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York: Springer; 2005.

[56] Beunckens C, Sotto C, Molenberghs G, Verbeke G. A multifaceted sensitivity analysis of the Slovenian Public Opinion Survey data. Appl Stat 2009; 58: 171–196.

[57] Chickering D, Pearl J. A clinician's tool for analyzing non-compliance. Compu Sci Stat 1997; 29: 424–431.

[58] Daniels MJ, Hogan JW. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Boca Raton: Chapman Hall/CRC; 2008.

[59] Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). J Roy Stat Soc A 2005; 168: 267–306.

Table 1: *Toenail Data. Number of available repeated measurements per subject, for each treatment arm separately.*

	Group A		Group B	
# Obs.	N	%	N	%
1	4	2.74%	1	0.68%
2	2	1.37%	1	0.68%
3	4	2.74%	3	2.03%
4	2	1.37%	4	2.70%
5	2	1.37%	8	5.41%
6	25	17.12%	14	9.46%
7	107	73.29%	117	79.05%
Total:	146	100%	148	100%

Table 2: *Developmental Toxicity Study (DEHP). Summary data by dose group.*

	# dams with		# live	average	
dose	implants	viable implants	fetuses	litter size	weight
0 mg/kg/day	30	30	330	13.2	0.9483
44 mg/kg/day	26	26	288	11.1	0.9592
91 mg/kg/day	26	26	277	10.7	0.8977
191 mg/kg/day	24	17	137	8.1	0.8509
292 mg/kg/day	25	9	50	5.6	0.6906

Table 3: *Accident insurance policies data of Thyroin (1960).*

Count (No. of claims)	0	1	2	3	4	5	6	7
Frequency (No. of policies)	7840	1317	239	42	14	4	4	1

Table 4: *Asthma data: The first four data points for the first two patients.*

Patient ID	Drug	Begin	End	Status
1	0	0	15	1
1	0	22	90	1
1	0	96	325	1
1	0	329	332	1
2	1	0	180	1
2	1	189	267	1
2	1	273	581	1
2	1	582	600	0

Table 5: *National Youth Risk Behavior Survey Data. Latent class model parameters.*

	Latent Class				
	1	2	3	4	5
Latent class prevalence	0.6741	0.1383	0.0910	0.0546	0.0420
	Probability "Yes"				
Driving after taking alc.	0.0058	0.4208	0.1488	0.4537	0.1098
Smoked before age 13	0.0422	0.1083	0.7584	0.6387	0.1738
Smoked daily for 30 days	0.0202	0.2670	0.3144	0.6588	0.1247
First alc. drink before age 13	0.1433	0.2075	0.7875	0.6790	0.3928
$\geq$5 alc. drinks/day, in past 30 days	0.0805	0.7421	0.4789	0.7875	0.1621
Took marijuana first before age 13	0.0074	0.0286	0.4596	0.5530	0.2173
Have ever used cocaine	0.0040	0.1919	0.0716	0.8800	0.0255
Tried glue sniffing, etc to get high	0.0550	0.1886	0.2153	0.5778	0.0420
Used methamphetamines	0.0035	0.0997	0.0245	0.7271	0.0102
Used ecstasy	0.0035	0.1093	0.0630	0.6429	0.0556
$<$13 years at first sexual intercourse	0.0138	0.0015	0.1753	0.2957	0.8073
Have had sex with at least 4 people	0.0639	0.2859	0.2409	0.5641	0.8348

Table 6: *National Track Records for Women: Factor analysis.*

Distance	Factor 1	Factor 2
100 metres	0.4406	0.8376
200 metres	0.4352	0.8908
400 metres	0.4116	0.8164
800 metres	0.7266	0.5673
1500 metres	0.8592	0.4822
3000 metres	0.9138	0.3859
Marathon	0.7654	0.3888

Table 7: *Toenail Data. (Unaffected nail length outcome). Parameter estimates (standard errors) for the model specified by (62) and (63).*

Effect	Parameter	Estimate	(Standard error)
Fixed effects:			
Intercept	β_0	2.46	(0.24)
Time effect	β_1	0.59	(0.05)
Dose effect	β_2	0.28	(0.34)
Dose by time interaction	β_3	0.04	(0.06)
Variance components:			
Random intercept variance	d_{00}	7.32	(0.70)
Random slope variance	d_{11}	0.22	(0.02)
Random effects covariance	d_{01}	-0.50	(0.10)
Residual variance	σ^2	3.15	(0.13)

Table 8: *Developmental Toxicity Study (DEHP). Parameter estimates (standard errors) for the model specified by (65) and (105).*

Effect	Parameter	Estimate	(Standard error)
Fixed effects:			
Intercept	β_0	0.9733	(0.0138)
Dose effect	β_1	-0.2563	(0.0327)
Variance components:			
Random intercept variance	d	0.0086	(0.0015)
Residual variance	σ^2	0.0195	(0.0009)

Table 9: *Toenail Data. Continuous, longitudinal unaffected-nail-length outcome. Parameter estimates (standard errors) for the model specified by (100) and (103).*

	Unaffected nail length		Dropout	
Effect	Parameter	Estimate (s.e.)	Parameter	Estimate (s.e.)
Mean structure parameters				
Intercept	β_0	2.510 (0.247)	γ_0	-3.127 (0.282)
Treatment	β_1	0.255 (0.347)	γ_1	-0.538 (0.436)
Time	β_2	0.558 (0.023)	γ_2	0.035 (0.041)
Treatment-by-time	β_3	0.048 (0.031)	γ_3	0.040 (0.061)
Variance-covariance structure parameters				
Residual variance	σ^2	6.937(0.248)		
Scale factor			γ_{01}	-0.076 (0.057)
Rand. int. variance	τ^2	6.507 (0.630)	$\gamma_{01}^2\tau^2$	0.038 (0.056)

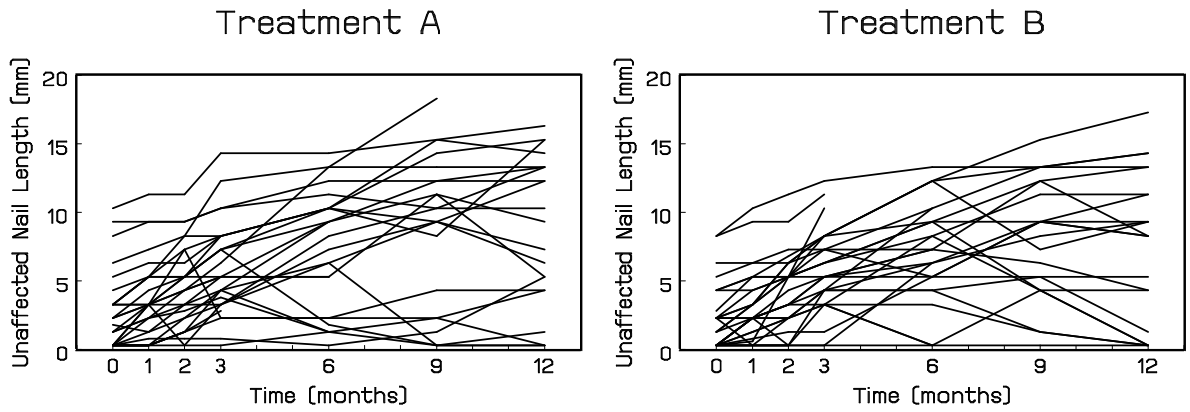

Figure 1: *Toenail Data. Individual profiles of 30 randomly selected subjects in each of the treatment groups in the toenail experiment.*

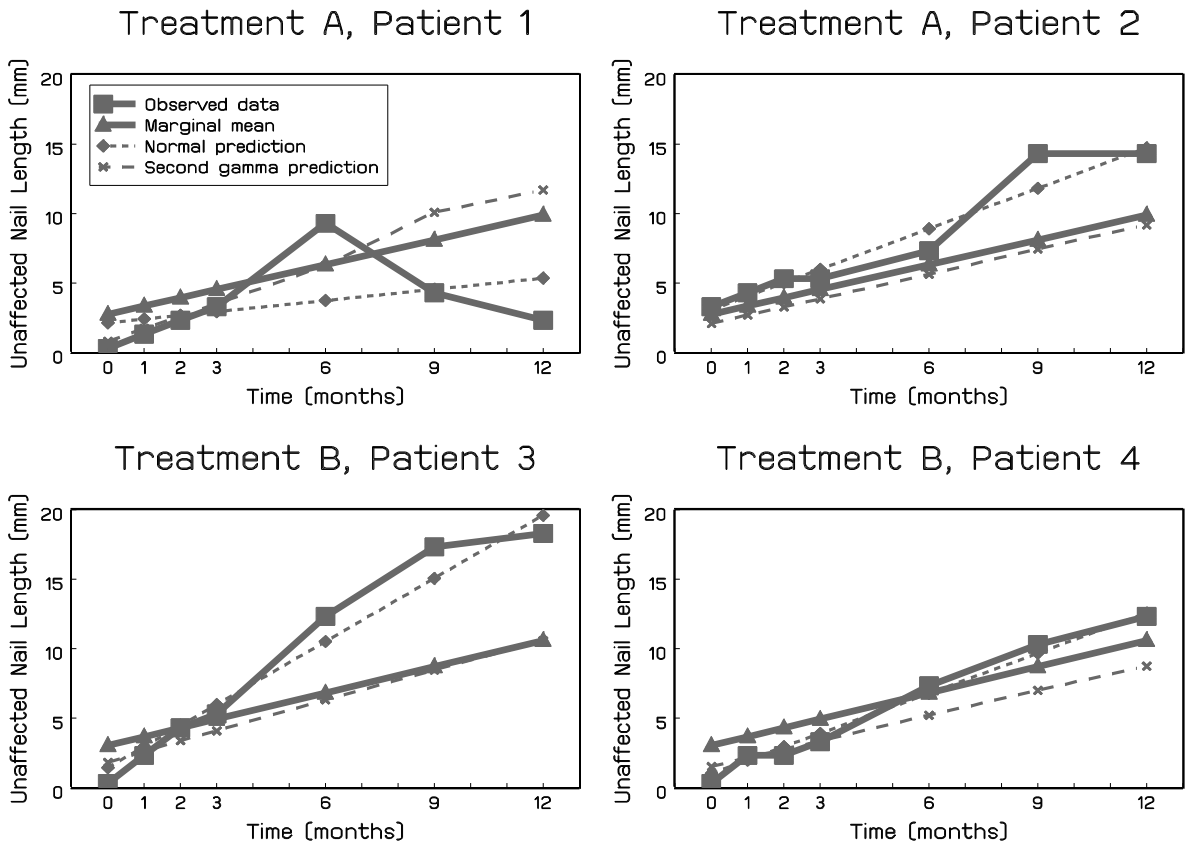

Figure 2: *Toenail data. For 4 selected subjects, two per treatment arm: (1): observed profile; (2) marginal mean profile (which solely depends on treatment); (3) prediction from the normal model (48); (4) prediction from the second exponential model (61).*

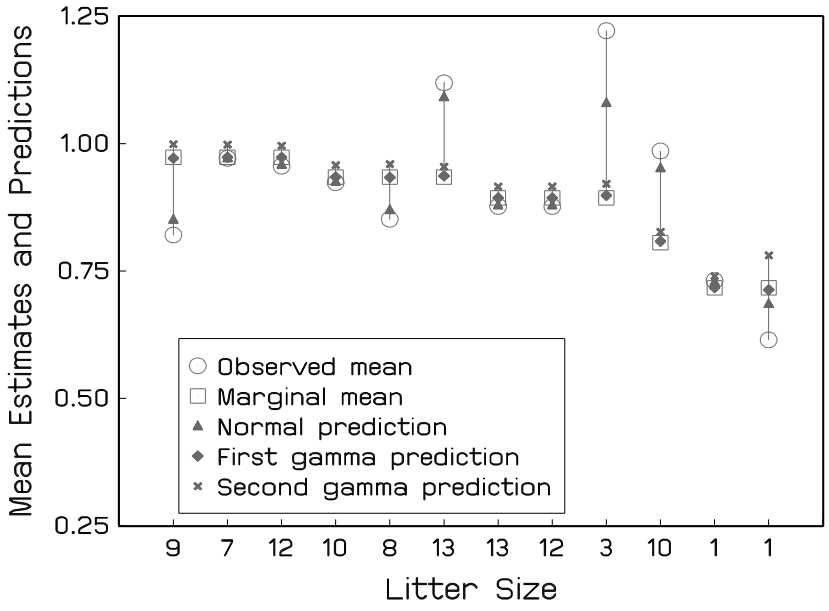

Figure 3: *Developmental Toxicity Study (DEHP). For 12 selected clusters from the control group (for which the size is shown in the x-axis): (1): observed average weight per cluster (2): the estimated marginal mean as given by (66); (3) prediction from the normal model (109); (4) prediction from the first exponential model (114); and (5): prediction from the second exponential model (119).*

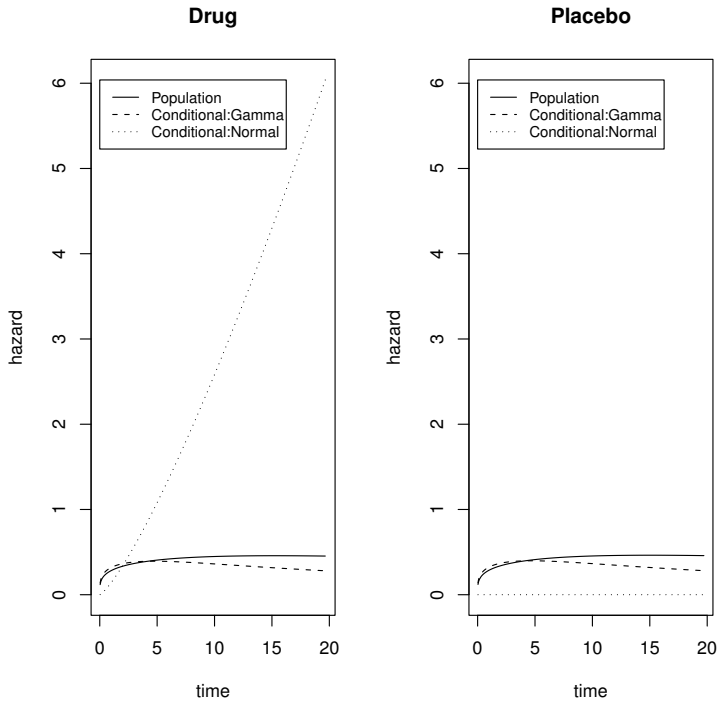

Figure 4: *Population and Conditional Hazard Functions.*

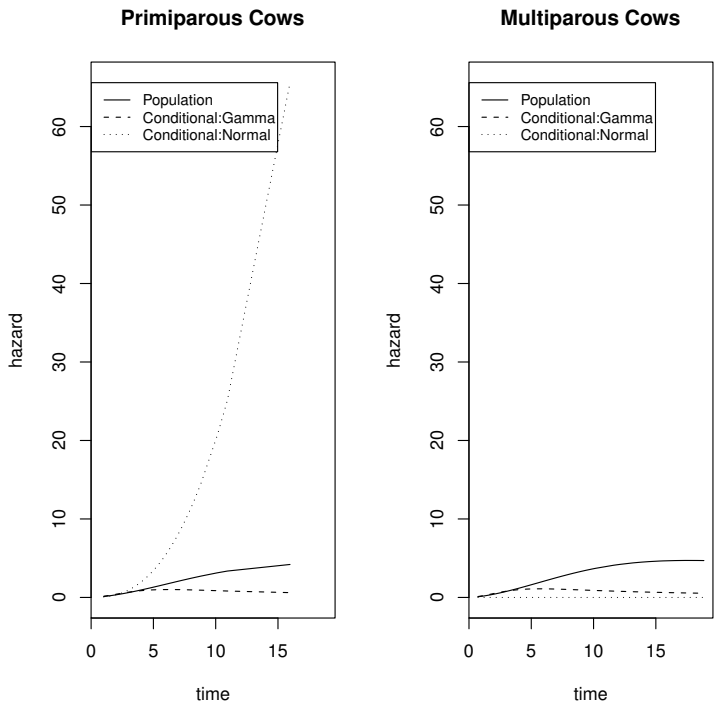

Figure 5: *Population and Conditional Hazard Functions.*

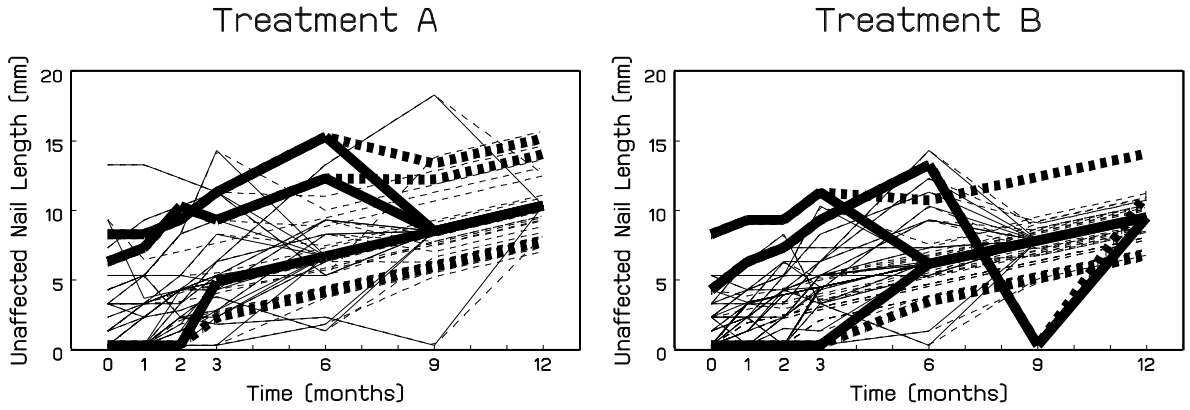

Figure 6: *Toenail Data. Individual profiles of subjects with incomplete data, for each treatment arm, extended using MNAR Model (100) (dashed line) and using the model's MAR counterpart (solid line). In each group, three subjects are highlighted.*

Enriched-data Problems and Essential Non-identifiability

Geert Molenberghs[1,2] **Edmund Njeru Njagi**[1]

Michael G. Kenward[3] **Geert Verbeke**[2,1]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[3] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E7HT, United Kingdom*

Supplementary Materials

A Exchangeable Data With Compound-symmetry Covariance

We now consider the special but enlightening case of exchangeable, compound-symmetry data, in the sense that all members of a cluster have the same mean μ_i and the variance-covariance matrix is of a compound symmetry structure $V_i = \sigma^2 I_{n_i} + dJ_{n_i}$, where I_{n_i} is an n_i-dimensional identity matrix and J_{n_i} is an $n_i \times n_i$ matrix consisting of ones. We will simply refer to this setting as the "exchangeable" one.

For each of the three model formulations in Section 7, we present the six model equations considered there, for the special case of interest here.

A.1 The Standard Linear Mixed-effects Model

Let $\mathbf{1}_{n_i}$ be a length n_i vector of ones and denote by $\overline{y_i}$ the average of the components of the outcome vector $\boldsymbol{y}_i$. Further, the following expressions are useful:

$$V_i^{-1} = \frac{1}{\sigma^2}\left(I_{n_i} - \frac{d}{dn_i + \sigma^2} J_{n_i}\right), \qquad |V_i| = \sigma^{2n_i} + n_i \sigma^{2(n_i-1)} d.$$

The exchangeable versions of (44)–(48) are:

$$\boldsymbol{Y}_i|b_i \sim N(\mathbf{1}_{n_i}\mu_i + \mathbf{1}_{n_i}b_i, \sigma^2 I_{n_i}), \tag{104}$$

$$b_i \sim N(0, d), \tag{105}$$

$$\boldsymbol{Y}_i \sim N(\mathbf{1}_{n_i}\mu_i, V_i = \sigma^2 I_{n_i} + dJ_{n_i}), \tag{106}$$

$$b_i|\boldsymbol{Y}_i \sim N\left[\frac{n_i d}{\sigma^2 + n_i d}(\overline{y_i} - \mu_i), \frac{\sigma^2}{\sigma^2 + n_i d}d\right], \tag{107}$$

$$\widehat{b_i} = \frac{n_i d}{\sigma^2 + n_i d}(\overline{y_i} - \mu_i), \tag{108}$$

$$\widehat{\boldsymbol{Y}_i} = \frac{n_i d\overline{y_i} + \sigma^2\mu_i}{\sigma^2 + n_i d} \cdot \mathbf{1}_{n_i}. \tag{109}$$

A.2 A First Normal-exponential Version of the Linear Mixed Model

It now makes sense to assume, like in Section 7.1.2, that there is a single, exponentially distributed, random effect. This alters the model from Section A.1 a bit, in addition to obvious simplification. This means that (106) will be coupled with

$$f(g_i|\boldsymbol{y}_i) = \gamma\overline{y_i}e^{-g_i\gamma\overline{y_i}}. \tag{110}$$

We obtain the following sequence of model equations:

$$f(g_i) = \gamma e^{-g_i\mu_i\gamma + \frac{1}{2}\frac{g_i^2\gamma^2}{n_i}(\sigma^2 + n_i d)}\left[\frac{n_i\mu_i - g_i\gamma(\sigma^2 + n_i d)}{n_i}\right], \tag{111}$$

$$f(\boldsymbol{y}_i|g_i) = \frac{n_i\overline{y_i}e^{-\frac{1}{2}\left[\frac{1}{\sigma^2}(\boldsymbol{y}_i - \mathbf{1}_{n_i}\overline{y_i})'(\boldsymbol{y}_i - \mathbf{1}_{n_i}\overline{y_i}) + \frac{n_i}{\sigma^2 + n_i d}(\overline{y_i} - \mu_i)^2\right] - g_i\gamma(\overline{y_i} - \mu_i)}}{(2\pi)^{n_i/2}|V_i|^{1/2}e^{\frac{1}{2}\frac{g_i^2\gamma^2}{n_i}(\sigma^2 + n_i d)}[n_i\mu_i - g_i\gamma(\sigma^2 + n_i d)]} \tag{112}$$

$$\widehat{g_i} = 1/(\gamma\overline{y_i}), \tag{113}$$

$$\widehat{\boldsymbol{y}_i} = \frac{\left\{\left[n_i\mu_i - \frac{1}{\overline{y_i}}(\sigma^2 + n_i d)\right]^2 + n_i(\sigma^2 + n_i d)\right\}\mathbf{1}_{n_i}}{n_i\left[n_i\mu_i - \frac{1}{\overline{y_i}}(\sigma^2 + n_i d)\right]}. \tag{114}$$

A.3 A Second Normal-exponential Version of the Linear Mixed Model

Now, (105) will be coupled with

$$f(q_i|\boldsymbol{y}_i) = e^{\gamma\overline{y_i}}e^{-q_i e^{\gamma\overline{y_i}}}. \tag{115}$$

This then produces the following sequence of model equations:

$$f(q_i) = \sum_{m=0}^{\infty} \frac{(-q_i)^m}{m!} e^{\mu_i \gamma (m+1) + \frac{1}{2}\frac{\gamma^2 (m+1)^2}{n_i}(\sigma^2 + n_i d)}, \tag{116}$$

$$f(\boldsymbol{y}_i|q_i) = \frac{e^{-\frac{1}{2}\left[\frac{1}{\sigma^2}(\boldsymbol{y}_i - \mathbf{1}_{n_i}\overline{y_i})'(\boldsymbol{y}_i - \mathbf{1}_{n_i}\overline{y_i}) + \frac{n_i}{\sigma^2 + n_i d}(\overline{y_i} - \mu_i)^2\right] + \gamma \overline{y_i} - q_i e^{\gamma \overline{y_i}}}}{(2\pi)^{n_i/2}|V_i|^{1/2} \sum_{m=0}^{\infty} \frac{(-q_i)^m}{m!} e^{\mu_i \gamma (m+1) + \frac{1}{2}\frac{\gamma^2 (m+1)^2}{n_i}(\sigma^2 + n_i d)}}, \tag{117}$$

$$\widehat{q_i} = e^{-\gamma \overline{y_i}}, \tag{118}$$

$$\widehat{\boldsymbol{y}_i} = \frac{\sum_{m=0}^{\infty} \frac{(e^{-\gamma \overline{y_i}})^m}{m!} e^{\mu_i \gamma (m+1) + \frac{1}{2}\frac{\gamma^2 (m+1)^2}{n_i}(\sigma^2 + n_i d)} \left[\mu_i + \frac{\gamma (m+1)}{n_i}(\sigma^2 + n_i d)\right] \mathbf{1}_{n_i}}{\sum_{m=0}^{\infty} \frac{(e^{-\gamma \overline{y_i}})^m}{m!} e^{\mu_i \gamma (m+1) + \frac{1}{2}\frac{\gamma^2 (m+1)^2}{n_i}(\sigma^2 + n_i d)}}. \tag{119}$$