Response Transformations for Random Effect and Variance Component Models

Amani Almohaimeed [1,2], **Jochen Einbeck** [2,3]

[1] Department of Mathematics, College of Science and Arts, Qassim University, Oyoon Aljawa, Saudi Arabia

[2] Department of Mathematical Sciences, Durham University, UK

[3] Durham Research Methods Centre, Durham University, UK

Address for correspondence: Amani Almohaimeed, Department of Mathematics, College of Science and Arts, Qassim University, Oyoon Aljawa, Qassim, Saudi Arabia.

E-mail: ama.almohaimeed@qu.edu.sa.

Phone: (+966) 16 3016953.

Fax: (+966) 16 3803070.

Abstract: Random effect models have been popularly used as a mainstream statistical technique over several decades; and the same can be said for response transformation models such as the Box-Cox transformation. The latter aims at ensuring that the assumptions of normality and of homoscedasticity of the response distribution are fulfilled, which are essential conditions for inference based on a linear model or a linear mixed model. However, methodology for response transformation and *simultaneous* inclusion of random effects has been developed and implemented only scarcely, and is so far restricted to Gaussian random effects. We develop such methodology, thereby

not requiring parametric assumptions on the distribution of the random effects. This is achieved by extending the "Nonparametric Maximum Likelihood" towards a "Nonparametric Profile Maximum Likelihood" (NPPML) technique, allowing to deal with overdispersion as well as two-level data scenarios.

Key words: Box-Cox transformation; Random effects model; Variance component model; Nonparametric maximum likelihood; EM algorithm

1 Introduction

In regression analysis, meeting the assumptions of normality and homoscedasticity of the response distribution and linearity of the model often requires transforming the response variable. The power transformation that was proposed by Box and Cox (1964) allows the response variable to achieve at least approximately a normal distribution, implicitly making the variance more nearly constant across data points around the regression line. Osborne (2010) suggested that normalizing data via the Box–Cox transformation to be a stage in data cleaning routines.

The Box-Cox transformation has been widely used in applied data analysis. The objective of the transformation is to select an appropriate parameter λ which is then used to transform data so that they follow a normal distribution more closely than the untransformed data. The transformation of the responses y_i, $i = 1, \ldots, n$, takes the form

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & (\lambda \neq 0), \\ \log y_i & (\lambda = 0), \end{cases} \tag{1.1}$$

where the restriction $y_i > 0$ applies. This family of transformations includes many traditional transformations to meet the needs of the data (Osborne , 2010), in particular $\lambda = 1$ means that no transformation is needed and hence produces results identical to the original data, $\lambda = 1/2$ is the square root transformation, $\lambda = 1/3$ corresponds to the cube root transformation, $\lambda = 0$ is the natural log transformation, $\lambda = -1/2$ yields the reciprocal square root transformation, and $\lambda = -1$ is the inverse transformation.

There is some close connection between such transformations and variance-stabilizing transformations. Sakia (1992) pointed out that the variance of a Box-Cox transformed variable can be approximated by

$$\text{Var}\left(y_i^{(\lambda)}\right) = \text{Var}(y_i)E(y_i)^{2\lambda-2}. \tag{1.2}$$

That is, the transformation $\lambda = 1/2$ is variance-stabilizing if the variance is proportional to the mean, such as for Poisson models, whereas $\lambda = 0$ has a variance-stabilizing effect if the variance is a quadratic function of the mean.

Box and Cox (1964) introduced their transformation originally for the linear model, where it is assumed that a set of explanatory variables x_i, $i = 1, \ldots, n$, and a response variable y_i are linearly related such that $y_i = x_i^T\beta + \varepsilon_i$, with independent errors ε_i which are usually taken to be Gaussian and homoscedastic. The transformation $y_i^{(\lambda)}$ is designed to mitigate violations of the latter two properties. However, not all types of violations can be mitigated through this route. It is often the case that the population from which the data are sampled consists of heterogeneous subpopulations. If these subpopulations are known, then they can simply be accounted for through an additional covariate in the model. However, frequently the subpopulations are *latent*, *i.e.* it is not possible to identify to which subpopulations the observations

of a sample belong (Wang , 2004). Under the resulting unobserved heteorogeneity, the errors cease to be independent, and their distribution tends to be multimodal. Fortunately, there is a well–known solution to this problem: The contribution by the latent subpopulation is captured by a random effect, conditional on which the errors restore their independence.

In this work, we intend to connect and combine both approaches, *i.e.* we assume that there is a value of λ so that the transformed responses are independently and normally distributed with mean function $E(y_i^{(\lambda)}|z_i) = x_i^T\beta + z_i$, conditionally on the random effect z_i. In explicit notation, one has then

$$y_i^{(\lambda)}|z_i \sim N(x_i^T\beta + z_i, \sigma^2), \tag{1.3}$$

where z_i is a random effect term with some density $g(\cdot)$. Under the presence of a random effect, the intercept term can be omitted from $x_i^T\beta$, so that, in what follows, $\beta \in \mathbb{R}^p$ denotes the vector of regression parameters excluding the intercept. For the distribution of $g(\cdot)$, several choices are possible, among them the normal distribution, as in the classical literature on linear mixed models. The extension of the transformation under this scenario was proposed by Gurka et al. (2006), and extended to the longitudinal data setting by Maruo et al. (2017) whose main interests were in robust estimation of fixed (treatment) effects.

However, a normal distribution is by definition unimodal, and hence may fail to capture the full heterogeneity of the latent subpopulations. An obvious concern is whether there are any harmful effects of this potential misspecification. Agresti et al. (2004) showed that a misspecification of the random effects distribution may affect the prediction accuracy of the random effects as well as the fixed effects, and suggest

that the "safest approach might seem to be always to use a nonparametric rather than a parametric approach for the random effects distribution." In consideration of the random effects misspecification, Wang et al. (2012) argued that even when the estimation of the fixed effect is robust, the estimation of the random effects could be invalid.

Accordingly, we follow in this work the concepts laid out by Aitkin (1996), which allows leaving the density $g(\cdot)$ unspecified. For estimation purposes, $g(\cdot)$ is then approximated by a finite discrete mixture with masses π_k at mass points z_k, $k = 1, \ldots, K$. These mixture parameters can be estimated alongside the other regression parameters in a usual EM algorithm. While it could, superficially, be argued that a 'discrete random effect' constitutes an even stronger limitation than a normal random effect, there is solid evidence that this is not the case. Methodologically, what is being approximated is the marginal likelihood,

$$L = \prod_{i=1}^{n} \int f(y_i|z_i)g(z_i)dz_i \approx \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i|z_k) \tag{1.4}$$

(where in our context $f(y_i|z_i)$ is the conditional density of the raw — not the transformed — data, which can be obtained from (1.3) using the transformation formula for probability density functions). It is known from early work by Laird (1978), Bock and Aitkin (1981) and Lindsay (1983), that this integral can be approximated with very high accuracy, and that the NPML estimate of the mixing distribution involves a finite number K of mass-points and corresponding masses. In practical applications, this integer K is typically very small, with values between $K = 2$ and 10.

In the context of model (1.3), the parameter λ needs to be estimated on top of the regression and mixture parameters, which leads us to an approach which one can consider as a 'nonparametric profile maximum likelihood' (NPPML) technique, in a

direct extension of the profile maximum log–likelihood estimation technique discussed by Box and Cox (1964). Box and Cox also suggested a way of producing confidence intervals for λ based on the χ^2-distribution.

Within the context of NPML estimation, Böhning et al. (2006) remarked that "profile likelihood ratios will not have standard χ^2-distributions", therefore, they suggested using model selection criteria for determining the number of components. Piepho and McCulloch (2004) considered the model selection in mixed models with transformations as "a difficult problem". Gurka (2004) suggested the use of likelihood-based measures such as Akaike's information criterion (AIC; Akaike, 1998) and the Bayesian information criterion (BIC; Schwarz , 1978) in the context of transformation models. Furthermore, graphical measures can be used for exploring normality such as control charts, probability plots, or histograms of residuals. Piepho and McCulloch (2004) suggested to fit a number of models and compare their fits by plotting the residual on the transformed and untransformed scale.

The rest of the paper is organized as follows. In Section 2 we introduce the NPPML technique by combining the Box-Cox transformation and the NPML estimation technique. Specifically, Subsection 2.1 lays out the maximum likelihood problem and derives explicit equations for the required EM algorithm. Subsection 2.2 is dedicated to model selection (for K), and the remaining subsections summarize further relevant technicalities. We extend the proposed technique to the two-level variance component model in Section 3. Simulated and real data applications are used to verify the proposed approach in Sections 4 and 5, respectively. Finally, we summarize our findings in Section 6. An implementation of the methodology is available in R package **boxcoxmix** (Almohaimeed and Einbeck , 2020).

2 Box-Cox transformation in random effect models

2.1 NPPML estimation

In this section, we consider estimation of the parameters in model (1.3). While the main goal is to estimate λ and β under the presence of the random effect, the discrete mixture approximation to the random effect distribution will be implicitly estimated, and may also be of interest in its own right in special applications.

Assuming positive responses y_i, and taking account of the Jacobian of the transformation from y_i to $y_i^{(\lambda)}$, the conditional probability density function of y_i given z_i is

$$f(y_i|z_i) = \frac{y_i^{\lambda-1}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i^{(\lambda)} - x_i^T\beta - z_i)^2\right]. \tag{2.1}$$

Under the nonparametric maximum likelihood estimation approach, the distribution of the random effect will be approximated by a discrete distribution at mass points $z_1, \ldots, z_K$, with masses $\pi_1, \ldots, \pi_k$ (Aitkin et al. , 2009), under the obvious constraints $\pi_k \geq 0$, $k = 1, \ldots, K$, and $\sum_{k=1}^K \pi_k = 1$. Along the lines of (1.4), the likelihood in relation to the original observations can be approximated as

$$L(\lambda, \beta, \sigma^2, z_1, ..., z_k, \pi_1, ..., \pi_k) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_{ik} \tag{2.2}$$

where $f_{ik} = f(y_i|z_k)$. Defining indicators $G_{ik} = 1$ if case i stems from cluster k and 0 otherwise (which constitute the 'missing information' for EM purposes), the complete log-likelihood takes the shape

$$\ell^* = \log L^* = \sum_{i=1}^n \sum_{k=1}^K \left[G_{ik} \log \pi_k + G_{ik} \log f_{ik}\right], \tag{2.3}$$

where $L^* = \prod_{i=1}^n \prod_{k=1}^K (\pi_k f_{ik})^{G_{ik}}$. If $K = 1$, the log-likelihood would be the usual log-likelihood of the Box-Cox model without random effects. Of course, ℓ^* depends

on λ. For fixed λ, one proceeds via a standard EM algorithm, where in the E-step expectations of G_{ik} are obtained via

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}, \tag{2.4}$$

and in the M-step the expected complete likelihood is maximized, yielding

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^T \right)^{-1} \sum_{i=1}^{n} x_i \left(y_i^{(\lambda)} - \sum_{k=1}^{K} w_{ik} \hat{z}_k \right), \tag{2.5}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{w_{ik} (y_i^{(\lambda)} - x^T \hat{\beta} - \hat{z}_k)^2}{n}, \tag{2.6}$$

$$\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik} (y_i^{(\lambda)} - x_i^T \hat{\beta})}{\sum_{i=1}^{n} w_{ik}}, \tag{2.7}$$

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} w_{ik}}{n}. \tag{2.8}$$

The estimates $\hat{z}_k$, $k = 1, \ldots, K$, and $\hat{\beta}$ are obtained by iterating between Equations (2.7) and (2.5) a small number of times within each M-step, where the resulting estimates of the previous EM iteration are used as starting values in this inner loop. This inner loop does not detrimentally impact the convergence behavior of the overall EM algorithm; note that the EM algorithm will still converge even if the M-step fails to identify the maximum of the complete likelihood, as long as an improvement of the latter is made (Dempster , 1997).

With view to the presence of $\sum_{i=1}^{n} w_{ik}$ in the denominator of (2.7), it is noted that this sum approaching the value 0 would correspond to $\hat{\pi}_k \longrightarrow 0$ and so ℓ^* approaching $-\infty$, that is, clearly not maximizing the (expected) complete log-likelihood. Hence, if K is larger than necessary, it appears more attractive for the EM algorithm to produce many identical $\hat{z}_k$'s with split probability mass, rather than allocating individual components the probability 0, even though the latter behavior has also been reported (Lukočienė , 2010). As long as K does not exceed the NPML solution (Böhning ,

2000), components with probability 0 (and denominators equal to 0 in (2.7)) are not to be expected, and indeed we did not observe such issues with the above algorithm.

So far λ has been fixed, and of course it is possible to stop here, in which case one has now completed the maximum likelihood estimation of mixture parameters under a given transformation parameter. However, the more interesting case is that λ needs to be estimated. In this case one repeats the procedure above over a grid of λ values, each time plugging the estimates (2.5) to (2.8) obtained for a given fixed λ into f_{ik} and then into the right-hand term of equation (2.2). This produces the profile-likelihood function $L_P(\lambda)$, or its logarithmic version $\ell_P(\lambda) = \log(L_P(\lambda))$. The non-parametric profile maximum likelihood (NPPML) estimator is therefore given by

$$\hat{\lambda} = arg \max_{\lambda} \ell_P(\lambda), \tag{2.9}$$

which can be found through a grid search over λ.

2.2 Model selection

In the original sense of NPML estimation, the value K is estimated by maximizing the likelihood successively for $K = 1, 2, 3 \ldots$ until there is no further improvement of the maximized likelihood (Laird , 1978; Aitkin et al. , 2009). Leroux and Puterman (1992) indicated that the NPML estimate may require an unnecessarily high number of components to maximize the likelihood whereas well-fitting models with a small number of components are usually preferred. Hence, it has become common to base the selection of K on a model selection criterion rather than the likelihood itself. In this work we follow Lukočienė (2010), who suggested an approach in which the number of components is increased until no further improvement is possible for the criterion used for model selection.

Commonly used criteria are Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Adapted to NPPML estimation, these criteria take the shape

$$\text{AIC} = -2\ell_P(\hat{\lambda}) + 2 \times (p + 2K - 1 + c) \tag{2.10}$$

$$\text{BIC} = -2\ell_P(\hat{\lambda}) + \log(n) \times (p + 2K - 1 + c), \tag{2.11}$$

i.e. the disparity, $-2\ell_P(\hat{\lambda})$, penalized by a quantity involving the total number of parameters estimated in the model. The constant c takes the value 1 if the transformation parameter is estimated and zero otherwise (in which case $\hat{\lambda} \equiv \lambda$, the given fixed value of the transformation parameter). The model with the selected number of classes is the one with the minimum AIC or BIC value.

Note that, even though $\hat{\sigma}$ depends on z_k and λ, the parameter σ is of no relevance for the problem of model selection, therefore, it is not included in the degrees of freedom (df) of the model. See Table 1 for an overview over all model parameters and associated df. As such, given a set of models, the best model in terms of relative quality will be the one with minimum AIC or BIC value.

While AIC is sometimes used in the literature to infer the number of components, it is known for its tendency to overestimate this number (McLachlan and Peel , 2004). The theoretical assumptions underlying both AIC and BIC break down for mixture models, but there is still some evidence supporting the use of the BIC, including a consistency result when considering the mixture as a density estimator (Celeux et al. , 2018). In direct comparison, the BIC has meanwhile established itself as the preferred choice (Steele and Raftery , 2010). The BIC still tends to overestimate the number of components especially if the model assumption for the component densities is invalid (McLachlan and Peel , 2004). Complications with BIC can also arise in variance

component models (Lukočienė and Vermunt , 2009).

Possible strategies for tailoring model selection criteria to the mixture problem are outlined in Celeux et al. (2018) or Naik et al. (2007). The latter paper, which introduces the 'mixture regression criterion' as a variant of the AIC, also indicates that the gain in accuracy of estimating the number of components as compared to BIC is not very large. For our purposes, the analysis which follows will consider AIC and BIC in order to select K, eventually confirming the established preference hierarchy of these two criteria in the context of NPPML estimation.

Parameters	df
$\hat{z}_1, \ldots, \hat{z}_K$	K
$\hat{\pi}_1, \ldots, \hat{\pi}_{K-1}$	$K-1$
$\hat{\beta}_1, \ldots, \hat{\beta}_p$	p
$\hat{\lambda}$	1

Table 1: Model parameters and degrees and freedom for use in information criteria

2.3 Starting point selection and the first cycle

In the first cycle of the EM algorithm, the model is fitted initially as

$$y_i^{(\lambda)} = \beta_0 + x_i^T \beta + \varepsilon_i. \tag{2.12}$$

The usual least squares solution of (2.12) delivers starting values $\beta^0 \in \mathbb{R}^p$ for β and σ^0 for σ. We set the initial estimates π_k^0 as equal probabilities $1/K$, so that it remains to choose the starting mass points z_k^0. There are several ways in which this can be done. Firstly, one can make use of Gauss-Hermite quadrature points (Einbeck and

Hinde , 2006). Under this approach, one sets

$$z_k^0 = \beta^0 + \texttt{tol} \times \sigma^0 \times g_k, \tag{2.13}$$

where `tol` is a scaling parameter which is typically restricted to the choice $0 \leq$ `tol` ≤ 2, and g_k are Gauss-Hermite quadrature points.

Alternatively, one could also make use of a quantile-based version

$$z_k^0 = \bar{y}^{(\lambda)} + \texttt{tol} \times q_k^{(\lambda)} \tag{2.14}$$

where $\bar{y}^{(\lambda)}$ is the mean of the transformed responses $y_i^{(\lambda)}$ and $q_k^{(\lambda)}$ are $(\frac{k}{K} - \frac{1}{2K})$-quantiles of the empirical distribution of $y_i^{(\lambda)} - \bar{y}^{(\lambda)}$. All application studies in this paper make use of the first of these two methods. The R package **boxcoxmix** (Almohaimeed and Einbeck , 2020) implements both approaches.

In either case, following the definition of the z_k^0 one obtains the extended linear predictor for the k-th component $E(y_i^{(\lambda)}|z_k^0) = x_i^T\beta^0 + z_k^0$ and associated densities $f(y_i|z_k^0)$ according to (2.1). Together with π_k^0, one is now able to compute initial weights $w_{ik}^0 = \pi_k^0 f(y_i|z_k^0)/\sum_\ell f(y_i|z_\ell^0)$, completing the initial E-step.

The subsequent M-step finds the parameter estimates by computing Equations (2.5) to (2.8), using $w_{ik} = w_{ik}^0$. From the resulting estimates of this cycle, one gets an updated value of the weights, and so on.

A comment is needed on the selection of the tuning parameter `tol`. For each fixed λ, one could in principle run the procedure described in Section 2.1 for a grid of `tol` values, and then choose the value of `tol` which returns the minimal disparity. However, this procedure is computationally expensive. Hence, we suggest a simpler approach where, for the fixed setting $\lambda = 1$, one finds the value of `tol` which minimizes

the disparity. This value of `tol` is then used across all considered λ values.

2.4 Summary of NPPML estimation procedure

For a fixed value of K, the previously described elements can be summarized as follows.

1. Decide on a range over which the optimization of λ will occur. We suggest the range from $\lambda_{min} = -3$ to $\lambda_{max} = 3$, with a grid of size $10 \times (\lambda_{max} - \lambda_{min})$. (Extreme transformations beyond this range are usually considered ineffective due to poor restoration of normality and other problems, see Osborne (2013) for related discussion.)

2. For the fixed setting $\lambda = 1$, find the value of `tol` which minimizes the disparity. Use this value of `tol` then across the whole grid of λ values.

3. For each fixed value of λ in the grid

   (a) Carry out the procedure described in Section 2.3 to identify suitable starting points.

   (b) Run the EM algorithm described by the E-step (2.4), and the M-step given by (2.5)–(2.8), noting the additional iteration between (2.7) and (2.5) required within each M-step.

   (c) Stop the algorithm if the difference of disparities $(-2\ell_P(\hat{\lambda}))$ between two subsequent iterations falls below a small threshold, such as 0.0001.

   (d) The resulting ML estimates $\hat{z}_k, \hat{\sigma}^2, \hat{\beta}$, $\hat{\pi}_k$ are used to produce $\ell_P(\lambda)$.

4. The optimal choice for λ is the one that maximizes $\ell_P(\lambda)$.

It is noted that this procedure is not equivalent to fitting a set of transformation models over a grid of λ values and then carrying out a simple grid search over λ to identify the 'best' model in some sense; such a procedure would be incorrect since the individual likelihoods would be based on the respective transformed data, not the raw data as in our approach, and hence not be comparable.

Concerning the selection of K, our practical advice would be to initially produce the values of the model selection criterion, such as BIC, until $K = 3$, and then increase K further if it has not yet stopped decreasing.

3 Box-Cox transformation in variance component models

For data with a two-level structure, such as longitudinal data, correlation of responses within upper–level units can be induced by adding a random effect z_i to the linear predictor $x_{ij}^T\beta$, with the upper-level indexed by $i = 1, \ldots, r$, and the lower-level indexed by $j = 1, \ldots, n_i$, $\sum n_i = n$. We assume that there is a value of λ for which

$$y_{ij}^{(\lambda)}|z_i \sim N(x_{ij}^T\beta + z_i, \sigma^2) \tag{3.1}$$

where z_i is a random effect with an unspecified mixing distribution $g(z_i)$. Under this model, which is also known as a variance component model, the responses $y_{ij}^{(\lambda)}$ are assumed to be conditionally independent given the random effect, with mean function

$$E(y_{ij}^{(\lambda)}|z_i) = x_{ij}^T\beta + z_i. \tag{3.2}$$

The marginal likelihood can again be approximated using NPML estimation (Aitkin et al. , 2009),

$$L(\lambda, \beta, \sigma^2, g) = \prod_{i=1}^{r} \int \left[\prod_{j=1}^{n_i} f(y_{ij}|z_i) \right] g(z_i) dz_i \approx \prod_{i=1}^{r} \sum_{k=1}^{K} \pi_k m_{ik}, \tag{3.3}$$

where $m_{ik} = \prod_{j=1}^{n_i} f(y_{ij}|z_k)$. The complete log-likelihood is thus

$$\ell^* = \log L^* = \sum_{i=1}^{r} \sum_{k=1}^{K} [G_{ik} \log \pi_k + G_{ik} \log m_{ik}] \tag{3.4}$$

where $L^* = \prod_{i=1}^{r} \prod_{k=1}^{K} (\pi_k m_{ik})^{G_{ik}}$. We apply the expectation-maximization (EM) approach similar as before, with the following adjustments:

1. In the E-step, the weights w_{ik} replace f_{ik} by m_{ik}.

2. In the M-step, the four estimators are now:

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} x_{ij}^T \right)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} \left(y_{ij}^{(\lambda)} - \sum_{k=1}^{K} w_{ik} \hat{z}_k \right), \\
\hat{\sigma}^2 &= \frac{\sum_{i=1}^{r} \sum_{k=1}^{K} w_{ik} \left[\sum_{j=1}^{n_i} (y_{ij}^{(\lambda)} - x_{ij}^T \hat{\beta} - \hat{z}_k)^2 \right]}{\sum_{i=1}^{r} n_i}, \\
\hat{z}_k &= \frac{\sum_{i=1}^{r} w_{ik} \left[\sum_{j=1}^{n_i} (y_{ij}^{(\lambda)} - x_{ij}^T \hat{\beta}) \right]}{\sum_{i=1}^{r} n_i w_{ik}}, \\
\hat{\pi}_k &= \frac{\sum_{i=1}^{r} w_{ik}}{r}.
\end{aligned}$$

As with the random effect models, we iterate between $\hat{z}_k$ and $\hat{\beta}$ a small number of times in each M-step to obtain their values. Substituting the results into Equation (3.3) we get the non-parametric profile log-likelihood function

$$\ell_P(\lambda) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \hat{\pi}_k \hat{m}_{ik} \right). \tag{3.5}$$

The NPPML estimator is therefore given by

$$\hat{\lambda} = arg \max_{\lambda} \ell_P(\lambda), \tag{3.6}$$

which can be found through a grid search over λ as in Equation (2.9). The procedures described in Subsections 2.3 and 2.4 extend accordingly to this scenario.

Concerning the choice of K, the comments relating to the relative merits of AIC and BIC apply similarly to variance component models (Lukočienė and Vermunt , 2009); however, as mentioned earlier there is a difficulty in the use of BIC which needs addressing. Note from (2.11) that the penalty term contains the quantity n. By strict use of BIC this would be the total sample size, that is the sum $n = \sum_{i=1}^{r} n_i$. However, noting that the mixture components operate on the upper level, Lukočienė and Vermunt (2009) made the case for employing instead the number of upper-level-units, r, for use in BIC, and endorsed this argument by a simulation study which shows that BIC may underfit the number of components otherwise. The possibility for BIC to underfit was also mentioned by McLachlan and Peel (2004) (albeit not in the context of variance component models), who stated that such behavior is possible if the samples sizes are 'not very large' and the component densities are 'valid'. However, we did not observe underfitting of the BIC criterion in real data sets, and this even though the samples sizes considered were rather small. Hence, we decided to use, and report, the BIC criterion for variance component models only according to its original definition, with $n = \sum_{i=1}^{r} n_i$ in (2.11).

4 Simulation studies

We are interested in examining the method's ability to estimate the true parameter values. Therefore, we first simulate data by applying the Box-Cox transformation 'backwards', through a transformation $\tilde{y}(\cdot)$ as defined below, from a dataset that fol-

lows a normal distribution. Specifically, we consider the following simulation designs:

Random effect model. We consider sample sizes $n = 100$ and 200, and $K = 1, 2, 4$ and 8. For each combination of n and K, and each of four given values λ_ℓ, $\ell = 1, 2, 3, 4$, we generate 1000 datasets. In each dataset, the i-th observation, $i = 1, \ldots, n$, for each $\ell = 1, \ldots, 4$, is generated as

$$\tilde{y}(\eta_i, \lambda_\ell) = \begin{cases} \left(1 + \lambda_\ell \eta_i\right)^{1/\lambda_\ell} & \text{for } \lambda_\ell \neq 0, \\ e^{\eta_i} & \text{for } \lambda_\ell = 0 \end{cases} \tag{4.1}$$

$$\eta_i = 3\, x_{1,i} + 0.5\, x_{2,i} + z_i + \varepsilon_i, \;\; \varepsilon \sim N(0, 0.5^2)$$

$$X_1 \sim U(-1, 1), \; X_2 \sim U(-3, 3)$$

$$\lambda_1 = 0, \;\; \lambda_2 = 0.5, \;\; \lambda_3 = 1, \;\; \lambda_4 = 2$$

$$z_i \sim \text{Multinomial}\{1, (z_1, \ldots, z_K) | \pi_1, \ldots, \pi_K\}$$

$$\pi_k = 1/K, \; k = 1, ..., K,$$

$$(z_k)_{k=1..K} = \begin{cases} (20) & \text{for } K = 1 \\ (20, 35) & \text{for } K = 2 \\ (15, 20, 30, 35) & \text{for } K = 4 \\ (20, 30, 35, 40, 50, 55, 60, 70) & \text{for } K = 8 \end{cases}$$

Variance component model. For the variance compopnent model, we consider $K = 2$ and $K = 4$. The i-th replicate, $i = 1, \ldots, n_j$, in the j-th upper-level unit,

$j = 1, \ldots, J$, for each $\ell = 1, \ldots, 4$, is generated as

$$\tilde{y}(\eta_{ij}, \lambda_\ell) = \begin{cases} \left(1 + \lambda_\ell \eta_{ij}\right)^{1/\lambda_\ell} & \text{for } \lambda_\ell \neq 0, \\ e^{\eta_{ij}} & \text{for } \lambda_\ell = 0 \end{cases} \tag{4.2}$$

$$\eta_{ij} = 3\, x_{ij} + z_i + \varepsilon_{ij}$$

$$x_{ij} \sim U(-4, 4), \;\; \varepsilon_{ij} \sim N(0, 0.5)$$

$$\lambda_1 = 0, \;\; \lambda_2 = 0.5, \;\; \lambda_3 = 1, \;\; \lambda_4 = 2$$

$$z_i \sim \text{Multinomial}\{1, (z_1, \ldots, z_4) | \pi_1, \ldots, \pi_4\}$$

$$(z_k)_{k=1..K} = \begin{cases} (35, 50) & \text{for } K = 2 \\ (15, 20, 30, 35) & \text{for } K = 4 \end{cases}$$

Clearly, the generated data possess random effects and variance component structures, respectively, due to the random effect terms z_i.

In the estimation step, we estimate λ and β (using a grid for λ as described in Section 2.4), yielding for each (true) value of λ a total of 1000 estimates of $\hat{\lambda}$ and $\hat{\beta}$ for each model. For the random effect model with $K = 4$, Figure 1 shows the boxplots for the regression and transformation parameter estimates, for samples sizes $n = 100$ and $n = 200$. The reference lines in the figures indicate the actual values of the parameters. The means and medians of the estimated λ and β parameters are also provided in Table 2. We find that the median of the estimated λ and β is approximately equal to the true value in each case, with the estimates being closer to the true values for $n = 200$. There are some outliers in each of the plots; in fact the outliers in the transformation estimates can cause even larger outliers in the regression estimates. This is, for instance, visible in the biased mean values of $\hat{\beta}_1$ in Table 2. It is clear, once that the estimate of λ is biased, then the estimate of β_1 *has* to be biased as the biased transformation shifts the scale of the linear predictor. Mitigatingly, for such

cases, one should say that the individual estimates of the regression parameters may still be useful *given* the respective estimated values of λ; they are just very poor in relation to the true parameters for the *true* values of λ. The corresponding tables for $K = 1$, $K = 2$ and $K = 8$ are provided in Appendix A. We see that the problem of outlying β estimates is not restricted to random effect models, and also occurs for $K = 1$. The results are generally best for $K = 2$. For $K = 8$, estimates get poor except if the true λ is equal to 0.

We also investigate the standard errors of the regression parameter estimates. An empirical but robust measure of spread of the estimated β can be obtained by computing the IQR of (the non-logarithmic version of) each of the four columns in Figure 1. Via normal reference, the IQR can be mapped back to the scale of the standard deviations by division through 1.349. We call the resulting estimate $\mathrm{RESD}(\hat{\beta})$, reading Robust Estimate of Standard Deviation. Tables 2 displays $\mathrm{RESD}(\hat{\beta})$ values along with means and medians of EM-based standard errors, $\mathtt{SE}(\hat{\beta})$, which were obtained by extraction from the model fitted in the last M-step. It is conceptually clear that such EM-based standard errors cannot be 'correct' as they ignore the variation caused by the EM algorithm itself *and* the variation caused by the estimation of λ, but we see from Table 2 that at least for $\lambda \leq 0.5$, they are still satisfyingly close to their empirical counterparts, with the approximation getting closer (and the standard errors generally getting smaller) for larger n. A look at the boxplots in Figure 1 shows that the variance of the estimates of λ increases as the value λ gets larger. This in turn causes the increased variability of the parameter estimates, yielding biases of their estimated standard errors for larger values of λ. If λ is assumed fixed and known, we would get $\mathrm{Mean}(\hat{\mathtt{SE}}(\hat{\beta}))$ values which are nearly equal to $\mathrm{RESD}(\hat{\beta})$ (not shown); that is, it is not the presence of λ by itself which causes the increased variance of $\hat{\beta}$, but

the need to estimate it.

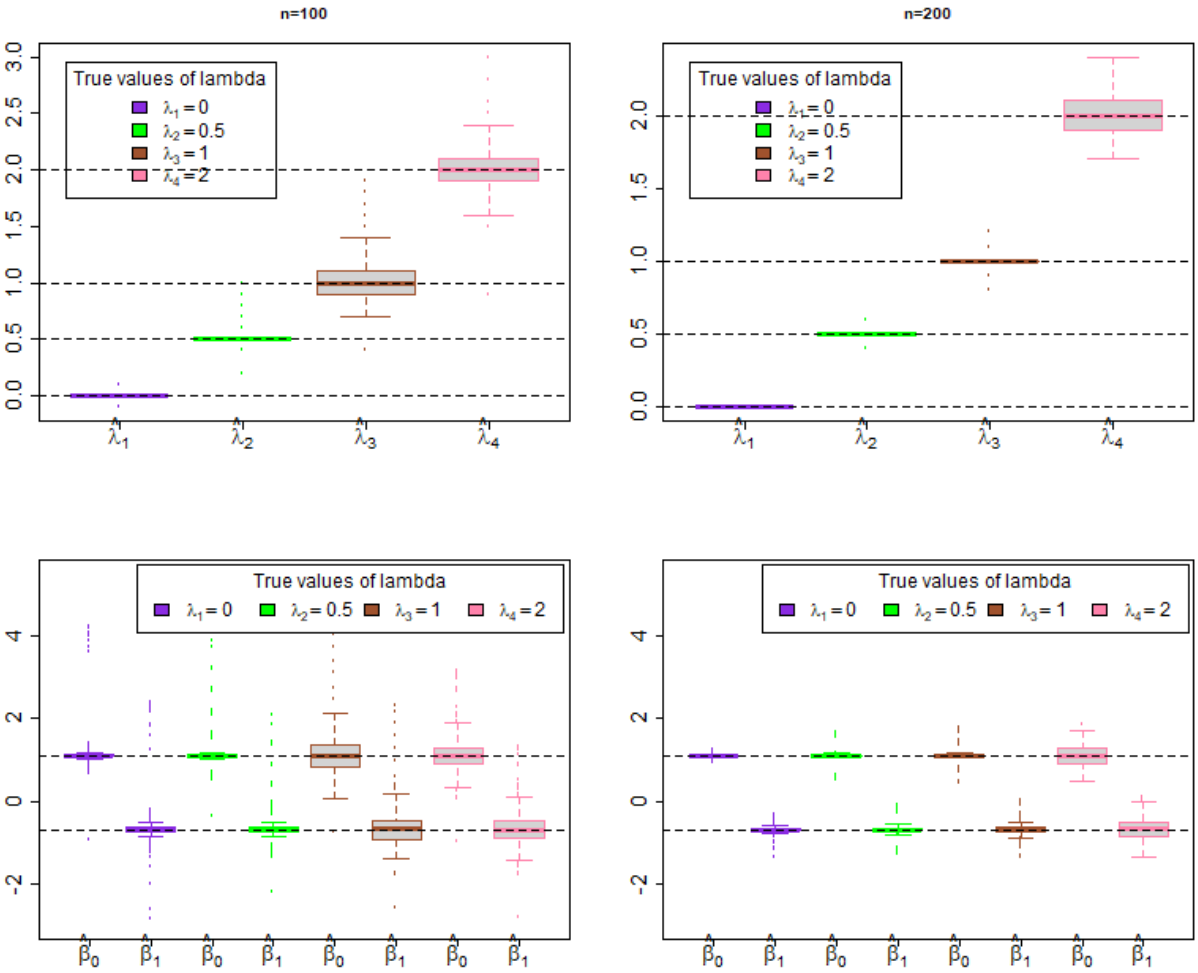

Figure 1: Simulation results for random effects model with $K = 4$, for $n = 100$ (left) and $n = 200$ (right): Estimates $\hat{\lambda}$ (top) and $\hat{\beta}$ (bottom; logarithmic scale), in each panel for true $\lambda_\ell = 0, 0.5, 1, 2$ (from left to right). Horizontal lines indicate the true values.

For the variance component model, we set initially $J = 20$ and $n_j = 5, j = 1, \ldots, J$, and investigate the cases $K = 2$ and $K = 4$. Results are provided in Figure 2 and Table 3. We see again an almost perfect match of the median estimates of transformation and regression parameters to their true values. It is also again visible that,

$K = 4$	$n = 100$				$n = 200$			
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
Mean($\hat{\lambda}$)	0.0006	0.5022	1.0046	2.0061	0	0.4992	1.0004	2.0007
Median($\hat{\lambda}$)	0	0.5	1	2	0	0.5	1	2
β_1	3	3	3	3	3	3	3	3
Mean($\hat{\beta}_1$)	3.3958	3.2882	3.3449	3.2626	2.9987	3.0240	3.0614	3.0718
Median($\hat{\beta}_1$)	3.0047	2.9990	3.0075	3.0073	2.9973	2.9975	2.9981	2.9983
RESD($\hat{\beta}_1$)	0.0983	0.1268	1.2527	0.8864	0.0611	0.0676	0.1135	0.8245
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.1387	0.0967	0.0988	0.0985	0.0660	0.0614	0.0621	0.0623
Median($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.0854	0.0843	0.0841	0.0842	0.0608	0.0607	0.0607	0.0607
β_2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mean($\hat{\beta}_2$)	0.5566	0.5472	0.5599	0.5432	0.5000	0.5049	0.5112	0.5129
Median($\hat{\beta}_2$)	0.5011	0.5011	0.5024	0.4998	0.5004	0.5003	0.5006	0.5019
RESD($\hat{\beta}_2$)	0.0339	0.0456	0.1702	0.1476	0.0206	0.0231	0.0371	0.1288
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0461	0.0322	0.0330	0.0328	0.0221	0.0205	0.0207	0.0208
Median($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0285	0.0282	0.0281	0.0282	0.0204	0.0203	0.0203	0.0203

Table 2: Summary of simulation results for random effects model with $K = 4$.

for $K = 2$, the estimates are more precise than for $K = 4$ (of course, assuming that the true K is used for estimation). We also see that the empirical and approximate standard errors are more similar as compared to the random effect model, with a very close correspondence if the median values of the latter are considered. In Appendix A, we also study the effect of a larger sample size in two ways, with firstly considering the double number of lower-level units, and then the double number of upper-level units (in each case totalling to $n = 200$). From Figure 6 we see that, qualitatively, not much seems to have changed; however from detailed analysis in Table 11 we find that standard errors of regression parameter estimates have become smaller, and

the correspondence of approximated to empirical standard errors has become better, especially in the second case where $J = 40$.

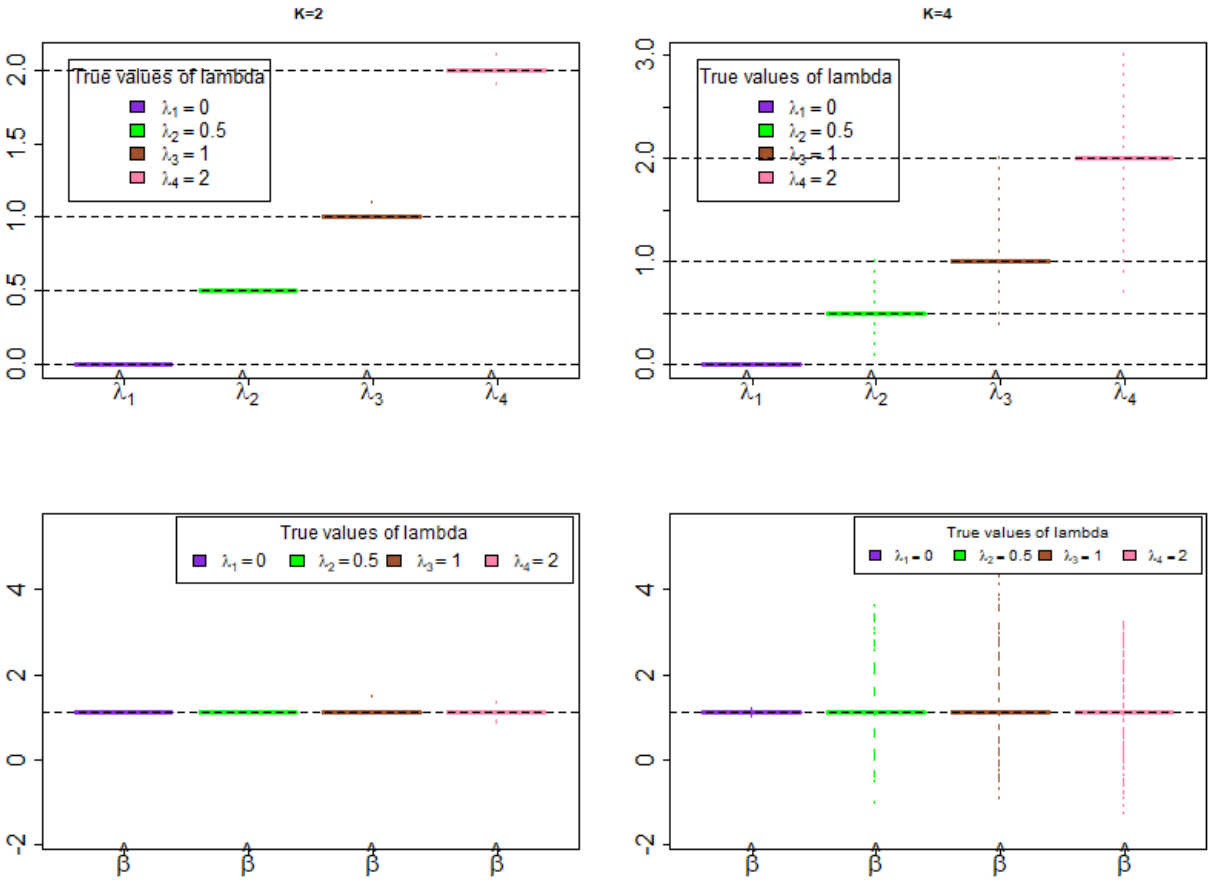

Figure 2: Simulation results for the variance component model with $K = 2$ (left) and $K = 4$ (right): Estimates $\hat{\lambda}$ (top) and $\hat{\beta}$ (bottom; logarithmic scale), in each plot for true $\lambda_\ell = 0, 0.5, 1, 2$ (from left to right). Horizontal lines indicate the true values.

5 Applications to real data

In this section, we illustrate the application of the proposed approaches using real data examples.

	$K=2$				$K=4$			
True values	$\lambda=0$	$\lambda=0.5$	$\lambda=1$	$\lambda=2$	$\lambda=0$	$\lambda=0.5$	$\lambda=1$	$\lambda=2$
β	3	3	3	3	3	3	3	3
Mean($\hat{\lambda}$)	0	0.5	1.0002	1.9994	0.0000	0.5091	1.0220	2.0279
Median($\hat{\lambda}$)	0	0.5	1	2	0	0.5	1	2
Mean($\hat{\beta}$)	3.0001	3.0001	3.0029	2.9984	2.9984	3.5760	3.9958	3.8369
Median($\hat{\beta}$)	3.0001	3.0001	3.0001	3	3.0001	3.0002	3.0003	3.0002
RESD($\hat{\beta}$)	0.0060	0.0060	0.0060	0.0061	0.0069	0.0072	0.0072	0.0072
Mean($\widehat{\mathtt{SE}}(\hat{\beta})$)	0.0059	0.0059	0.0059	0.0059	0.0191	0.03160	0.0434	0.0395
Median($\widehat{\mathtt{SE}}(\hat{\beta})$)	0.0057	0.0057	0.0057	0.0057	0.0061	0.00613	0.0061	0.0061

Table 3: Summary of simulation results for variance component model, for $K = 2$ (left) and $K = 4$ (right).

5.1 Internet Usage data

We firstly consider the `WWWusage` data from the **R** library **datasets** (R Core Team , 2016) which is a time series which records, over 100 minutes, how many users an internet server had every minute. The graphical representations by Qarmalah et al. (2018) indicated that the data follows a mixture of either three or four normal distributions. However, it is a relevant question, already alluded to in McLachlan and Peel (2004), whether heterogeneity can be reduced by considering an adequate transformation such as a log-normal model. In the context of our work, this corresponds to the problem of finding the best transformation parameter in order to fit a Gaussian mixture model (without any predictors) to the transformed data. In the notation of (1.3), this situation is described as $y_i^{(\lambda)}|z_i \sim N(z_i, \sigma^2)$. Note that in this scenario the NPPML estimation is slightly simplified since (2.6) and (2.7) are replaced by

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(y_i^{(\lambda)} - z_k)^2$ and $\hat{z}_k = \frac{\sum_{i=1}^{n} w_{ik} y_i^{(\lambda)}}{\sum_{i=1}^{n} w_{ik}}$ respectively, and hence no additional iteration within the M-step is required.

We investigate this problem by applying the Box-Cox transformation. Following discussion in Section 2.2, K is informed by considering the AIC and BIC from fitting mixture models for different numbers of classes. The model is initially fitted with $\lambda = 1$ and this value of `tol` is henceforth used for all considered λ's. The results are illustrated in Table 4. For both raw and transformed data, AIC achieves its minimum at $K = 8$, and BIC at $K = 4$. It is emphasized that the values of the criteria for the transformed and untransformed models are indeed comparable, since the likelihoods are always with reference to the original data, and hence operate on the same scale. This means, for this data set, the untransformed BIC solution at $K = 4$ (995.43) is preferable to the transformed BIC solution (999.97).

The trade-off between λ and K, as illustrated in Figure 3, is interesting: For $K = 1$ or 2, a log-transformation (or similar) appears suitable, while for $K \geq 3$ the estimates of λ settle quite robustly at around $\lambda \approx 1$, clearly indicating that there is no need for transformation once that heterogeneity is accounted for through an increased number of components. That supports the suggestion by Qarmalah et al. (2018) that the `WWWusage` data follows a normal distribution subject to heterogeneity.

5.2 Fabric data

In this example, we consider a data set available as part of the **R** package **npmlreg** (Einbeck et al. , 2007), which consists of 32 observations concerning faults in rolls of fabric. We are interested in the effect of the number of faults y on the log of the

		$\lambda = 1$			$\lambda = \hat{\lambda}$			
K	`tol`	$-2\ell_P(\lambda)$	AIC	BIC	$\hat{\lambda}$	$-2\ell_P(\lambda)$	AIC	BIC
1	–	1021.56	1023.56	1026.17	0.14	1015.76	1019.76	1024.97
2	1.1	1016.71	1022.71	1030.53	0.14	1014.75	1022.75	1033.18
3	0.6	992.32	1002.32	1015.35	1.02	992.57	1004.57	1020.20
4	0.2	963.19	977.19	**995.43**	0.9	963.13	979.13	**999.97**
5	0.1	963.19	981.19	1004.64	0.9	963.13	983.13	1009.18
6	0.1	958.00	980.00	1008.66	0.61	957.73	981.73	1012.99
7	0.2	955.68	981.68	1015.55	1.37	953.94	981.94	1018.41
8	0.1	**938.81**	**968.81**	1007.89	0.725	936.75	**968.75**	1010.43
9	0.1	955.68	989.68	1033.97	0.78	**936.18**	972.18	1019.07

Table 4: Comparison of results from the untransformed and transformed `WWWusage` data using K from 1 to 9. Minimal values for each column given in bold face.

length of the roll given by the variable x. McLachlan and Peel (2004) and Aitkin et al. (2009) observed overdispersion of the simple Poisson regression model, and used NPML with two and three mass-points to produce Poisson mixture regression models. Aitkin (1996) and Hinde and Demetrio (1998) fitted several further related models to these data. We approach this modelling problem through transformation models. In order to account for the overdispersion, a random effect z_i with an unspecified mixing distribution $g(z)$ is added to the linear predictors,

$$y_i^{(\lambda)} = \beta_1 x_i + z_i + \epsilon_i. \tag{5.1}$$

The optimal `tol` values, the disparities, AIC and BIC values for each K are given in Table 5. The first observation to make is that the values of the selection criteria for the transformed model are throughout well below their untransformed counterparts, giving clear evidence that a transformation is beneficial.

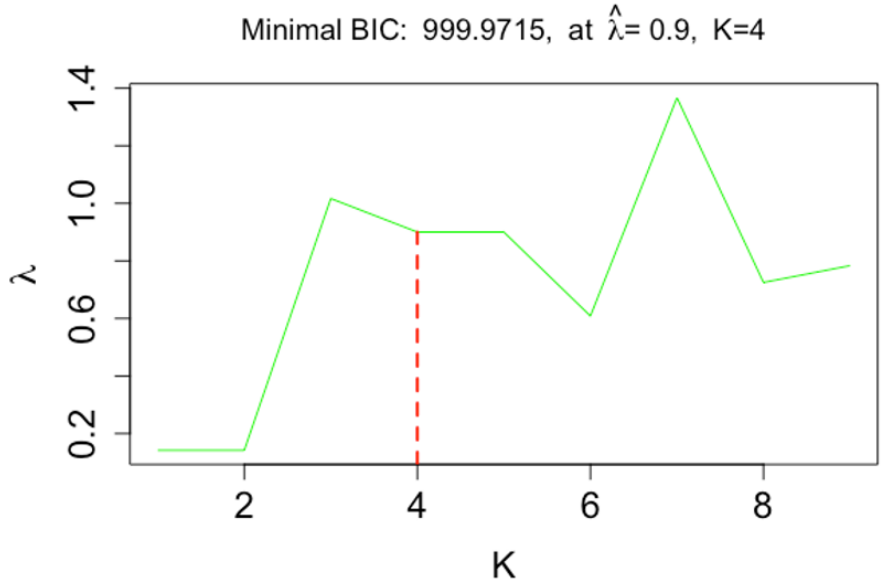

Figure 3: $\hat{\lambda}$ as a function of K for modelling the `WWWusage` data. The minimal BIC value at K=4 is highlighted through a vertical dashed line.

From Table 5, one finds that, for the untransformed data, the disparity settles at $K = 8$, but AIC and BIC attain their minimum already at $K = 1$, with both criteria thereafter monotonically increasing until $K = 7$. AIC and BIC values of the model after applying the response transformation are also shown in Table 5 and the minimal AIC value (172.93) occurred at $K = 9$ with $\hat{\lambda} = -3$, while the minimal BIC value (186.05) occurs at $K = 1$ with $\hat{\lambda} = 0.1$.

The first row of Table 5, for $K = 1$, corresponds just to a fixed effect model. That is, the value $\hat{\lambda} = 0.1$ given for $K = 1$ in the right hand part of the table is just the 'usual' Box-Cox estimate for λ under the model

$$y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{5.2}$$

This suggests that, while both the fixed and random effect model benefit from trans-

formation, there is only some ambiguous evidence for using a random effect model at all, with AIC favoring a model with $K = 9$, and BIC favoring the fixed effects model ($K = 1$).

What can be said about the relative merits of the AIC and BIC solutions for the transformed model? The BIC solution involves a relatively mild transformation, which is reasonably stable over a quite wide range of K values (Figure 4). In contrast, the AIC solution delivers a high number of K, and a very extreme setting of λ, which is furthermore quite instable over neighboring values of K. It is noted that the AIC solution $\hat{\lambda} = -3$ sits at the boundary of the considered range of transformation parameters, but as indicated in Section 2.4 we would advise against using such extreme settings of λ if it can be avoided — and here it can clearly be avoided using the BIC solution. It is furthermore noted that the AIC solution at $K = 9$ involves only five clearly distinct mass points and has a random effect standard deviation of $\hat{\sigma} \approx 0.0001$, indicating that the NPPML routine falls trap to some type of likelihood spike at this instance (Aitkin et al. , 2009). We also see that for $K = 10$, estimation of the transformed model deteriorates considerably.

From the right hand part of Table 5, there appears to be good evidence that a suitable response transformation of the fabric data will be in the region $\lambda \approx 0$, which suggests a log-transformation. Indeed, given that the 'number of faults' are count data, this is what many practitioners would have intuitively considered. Also, as mentioned, models fitted to these data in the literature are usually Poisson, which carry a log-link. So, it is of interest to compare to this model too, and the results are presented in Table 6. We see that, while for the Poisson model there is evidence for heterogeneity, for the logarithmically transformed model there is not. Interestingly, a log-transformation for

		$\lambda = 1$				$\lambda = \hat{\lambda}$		
K	`tol`	$-2\ell_P(\lambda)$	AIC	BIC	$\hat{\lambda}$	$-2\ell_P(\lambda)$	AIC	BIC
1	–	194.28	**198.28**	**201.21**	0.1	175.65	181.65	**186.05**
2	1.5	192.21	200.21	206.07	-0.3	171.88	181.88	189.20
3	1.5	192.21	204.21	213.01	-0.3	171.88	185.88	196.14
4	1.5	192.21	208.21	219.94	-0.3	171.88	189.88	203.07
5	1.4	192.21	212.21	226.87	-0.3	171.88	193.88	210.00
6	0.1	192.21	216.21	233.80	-0.4	164.93	190.94	209.99
7	0.1	192.21	220.21	240.73	-0.4	162.31	192.31	214.29
8	0.1	**181.20**	213.20	236.65	-2.8	142.58	176.58	201.50
9	0.1	181.20	217.20	243.58	-3	**134.93**	**172.93**	200.78
10	1.5	181.20	221.20	250.51	-1.6	158.16	200.16	230.94

Table 5: Comparison of results from the untransformed and transformed `fabric` data using K from 1 to 10. Minimal values for each column given in bold face.

$K = 1$ is preferable, under both selection criteria, to the Poisson log-linear model for $K = 2$, giving some justification to the use of a transformation model for these data. Note finally that the only reason why the AIC and BIC values of the logarithmic model in Table 6 are smaller than those of the full transformation model from Table 5 is that the former involves one less degree of freedom for the estimation of the transformation parameter; arguably this is not quite fair since even for the logarithmic model the data analyst needs to 'decide' on using that transformation which could be considered as a process 'costing' 1 df as well.

In practical applications, it is usually a good idea to look beyond simple model selection criteria, and investigate properties of the fitted model in more detail before making a final judgement. For instance, control charts are a helpful tool to assess

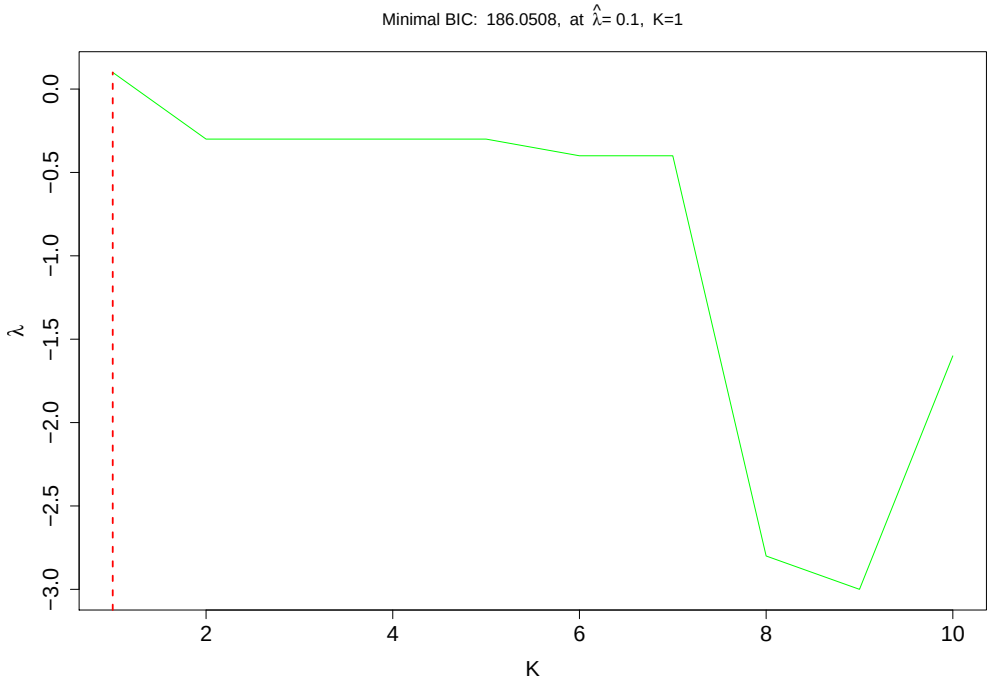

Figure 4: $\hat{\lambda}$ as a function of K for modelling the `fabric` data. The minimal BIC value at K=1 is highlighted through a vertical dashed line.

the normality of data and/or the homogeneity of variance. An application of control charts to models fitted to the fabric data is given in Appendix B.

5.3 Heights of boys in Oxford data

Next we consider a data set available as part of the **R** package **nlme** (Pinheiro et al , 2016), which consists of measurements of `height` (in cm) and `age` for 26 boys in Oxford. The variable `age` is reported on a standardized and dimensionless scale with nine possible values, yielding a total of 234 observations. We fitted a variance component model

$$E(y_{ij}|z_i) = \texttt{age}_j + z_i$$

	$\lambda = 0$			Poisson model		
K	$-2\ell_P(\lambda)$	AIC	BIC	$-2\ell_P(\lambda)$	AIC	BIC
1	175.98	**179.98**	**182.91**	187.84	191.84	194.77
2	173.29	181.29	187.15	**172.66**	**180.66**	**186.52**
3	**173.28**	185.28	194.07	172.67	184.67	193.46
4	173.28	189.28	201.00	172.66	188.66	200.39

Table 6: Comparison of results for $\lambda = 0$ and for the Poisson log-linear model, for the `fabric` data using K from 1 to 4. Minimal values for each column given in bold face (results for $K \geq 5$ do not bring further improvements and are hence omitted). The models on the left hand side are fitted using function `np.boxcoxmix` in R package **boxcoxmix**, which executes steps 1—3 in Section 2.4; i.e. it fits transformation models for fixed $\lambda = 0$. The models on the right hand side are fitted using `alldist` in R package **npmlreg**.

where z_i is boy-specific random effect and $\texttt{age}_j$ is the j-th standardized age measurement, $j = 1, \ldots, 9$, which is equal for all boys for fixed j.

The results before and after applying the response transformation are summarized in Table 7. We see that, for the untransformed data, BIC suggests $K = 9$ mass points. After transformation, an 8-component model is the best choice. These values of K still appear quite high, given that they describe the heterogeneity between only 26 boys on the upper level, but they concur with results reported previously for this data set in the literature (Einbeck et al. (2007); Aitkin et al. (2009)). We see again a subtle interplay between K and λ: For the fixed effects model with $K = 1$, there is no strong evidence that a transformation is required, but once we go to three or more classes we see that the selected transformation parameter oscillates around a log-transformation ($\lambda = 0$). This is also illustrated, for the specified range of K from

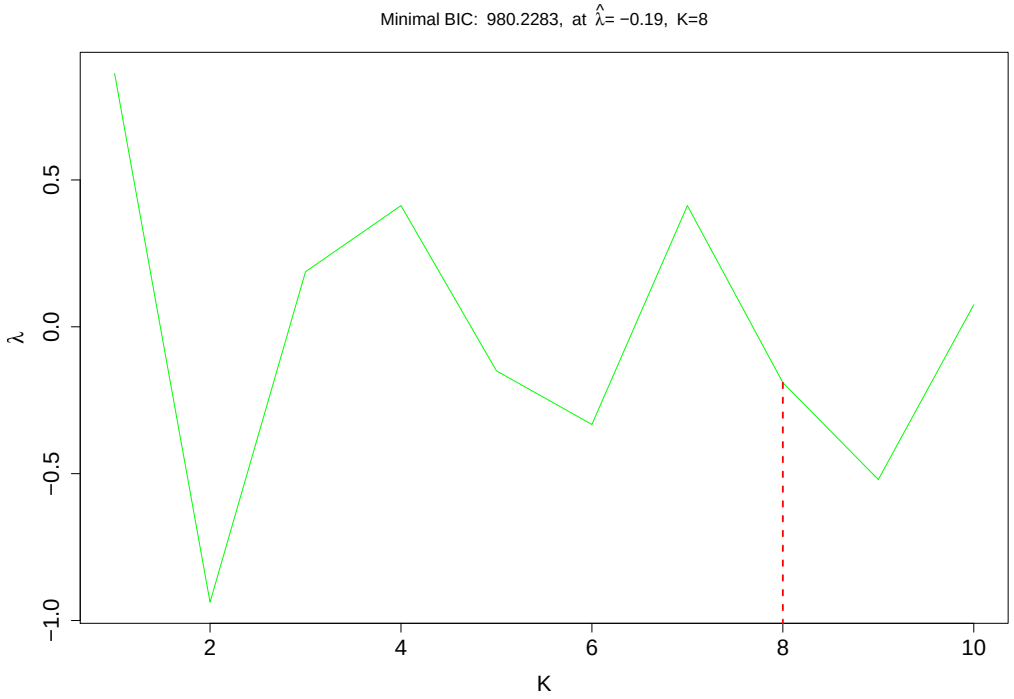

Figure 5: $\hat{\lambda}$ as a function of K for the `Oxboys` data. The minimal BIC value at K=8 is highlighted through a vertical dashed line.

1 to 10, in Figure 5. In direct comparison of BIC values for each fixed $K = 1, ..., 10$, respectively, the transformed version is superior for all $K \geq 5$, giving some evidence that the transformation leads to better fitting models than the original data.

For the practicing data analyst, it will of course not be practicable to produce tables such as Table 7, 4 or 5 for every data set considered. Hence, if heterogeneity is suspected, we suggest to follow the guidance formulated at the end of Section 2.4, i.e. consider firstly the BIC values up to $K = 3$ and then proceed further if required.

		$\lambda = 1$			$\lambda = \hat{\lambda}$			
K	`tol`	$-2\ell_P(\lambda)$	AIC	BIC	$\hat{\lambda}$	$-2\ell_P(\lambda)$	AIC	BIC
1	–	1641.93	1645.93	1652.84	0.86	1641.88	1647.88	1658.25
2	1.5	1466.76	1474.76	1488.58	-0.938	1457.86	1467.86	1485.14
3	1.2	1320.88	1332.88	1353.61	0.188	1318.47	1332.47	1356.66
4	0.2	1212.66	1228.66	1256.30	0.41	1211.35	1229.35	1260.44
5	0.8	1132.85	1152.85	1187.40	-0.150	1121.09	1143.09	1181.10
6	1.1	1048.27	1072.27	1113.73	-0.333	1025.25	1051.25	1096.17
7	0.5	1017.27	1045.27	1093.64	0.41	1002.91	1032.91	1084.74
8	0.5	931.38	963.38	1018.66	-0.190	887.49	921.49	**980.23**
9	0.5	916.09	952.09	**1014.29**	-0.520	878.56	916.56	982.21
10	0.3	**908.00**	**948.00**	1017.11	0.075	**866.76**	**908.76**	981.32

Table 7: Comparison of results from the untransformed and transformed `Oxboys` data using K from 1 to 10. Minimal values for each column given in bold face.

6 Discussion

It is common to normalize non-normal data via a normalizing transformation prior to analysis. In order to select an appropriate transformation parameter for the linear model with random effects of unspecified distribution, we have developed methodology for simultaneous response transformation and estimation of regression parameters. This is achieved by extending the "Nonparametric Maximum Likelihood" towards a "Nonparametric Profile Maximum Likelihood" technique. The methodology is implemented in **R** package **boxcoxmix** (Almohaimeed and Einbeck , 2020) which is available on CRAN. This package features further variants and capabilities which have not been introduced here, such as a version for logistic mixed effect models.

To assess the performance of the proposed approach, we conducted two simulation studies. The first simulation was for the Box-Cox transformed random effects model. We have seen that the method is able to identify well the true value of both λ and β, with, however, some outlying estimates of the former having a potentially severe impact on the estimates of the latter. The second study concerned the Box-Cox transformed variance component model. To some extent, the results of this simulation differ from those of the random effects models: There was less variability in the estimates of the transformation and regression parameters of the variance components model, except for a few, relatively symmetrically distributed outliers, and also the approximation of EM-based to empirical standard errors was closer for this scenario.

The simulation results also showed a high precision and accuracy of all parameter estimates (including the transformation parameter) when the log–transformation is the most appropriate transformation; see the left column of Figures 1 and 2. These results appear to concur with Asar et al. (2017) who proposed different approaches to estimate the Box-Cox power transformation parameter and carried out simulation studies to compare their effectiveness. The results indicated that all of the methods, including the one that was not preferred to estimate λ, performed well at $\lambda = 0$ regardless of what design is used to generate the data. So, it appears to be right to say that there is something 'special' about the log-transformation. Keene (1995) devoted the title of his paper to this observation, and went on to give several reasons why this is the case. One of these is the fact that the log transformation is the only member of the Box-Cox family which can produce a genuinely normally distributed transform of a positive variable. A second reason, which we already touched upon the introduction, is its variance-stabilizing property. Note from equation (1.2) for $\lambda = 0$,

that

$$\mathrm{Var}(\log(y_i)) \approx \frac{1}{E(y_i)^2}\mathrm{Var}(y_i),$$

which will be approximately constant whenever the ratio between variance and mean is quadratic, which is the case for the Gamma distribution but which is also compatible (up to an additional linear term) with many overdispersed Poisson distributions such as the negative binomial (type II) distribution. It is worth noting that a quadratic marginal variance is also obtained when including a normal random effect into the linear predictor of a Poisson log-linear model. Further discussion on the interplay of the Box-Cox transformation and variance heterogeneity, under the general scenario $\mathrm{Var}(y_i) = \sigma^2 E(y_i)^\delta$ for some known or unknown constant δ, was given by Sakia (1992).

However, some more sceptical views about log-transformations have also been expressed in the literature. Changyong et al. (2014) showed that the log transformation does not necessarily make data conform more closely to the normal distribution. Gurka (2004) considered the possibility of using a small value of λ that is close to zero for transforming the response instead of the log-transformation when $\lambda = 0$ is selected as the optimum.

As in the fixed-effect case, the Box-Cox transformation under random effects does not guarantee that the assumption of normality of the response distribution in the random effects model is met after applying the transformation, however, it provides data for which the normality assumption is more reasonable than not applying the transformation at all. The examples and simulations that we have presented showed that substantial improvements in terms of the AIC and BIC criteria can be achieved through transformation; noting that our transformed model fits are 'conservative'

since we chose not to optimize the tuning parameter `tol` for each different setting of λ.

It is not possible to report a simple likelihood-based confidence interval for $\hat{\lambda}$, the reason being that the likelihood in the considered model class is highly non–concave. Hence, when faced with the decision on whether or not needing to transform the response, not only the value of $\hat{\lambda}$ but also the relevant model selection criteria such as AIC and BIC should be taken into account. It is then essential that these are always based on likelihoods which are reported on the original response scale, as in the models (2.2) and (3.3); of course, this is the case for the values $-2\ell_P(\lambda)$, AIC and BIC provided herein.

In Example 5.1, $\hat{\lambda}$ for the fixed effect model was much further away from $\lambda = 1$ than for the random effects model, therefore, it is beneficial to test the need for a transformation of the response of a random effect model even if the fixed effect model does need transformation! Hence, the proposed method can help to judge whether the data really needs to be transformed or only the right number of components needs to be found in order to adequately reflect the heterogeneity in the response distribution. This ties in with other work recognizing that mixtures can also be used to model skewed data (Pearson , 1895; McLachlan and Peel , 2004); in the latter monograph it was also noted that "the choice between the log normal and normal mixture model is of much interest". It appears that there is a trade-off between transformation and mixed-effect models; both of them change the nature of the variance explained by the model.

This trade-off appears to manifest itself in different ways for different examples. By inverting the model equation, such as (5.1), to take the shape $z_i = y_i^{(\lambda)} - \beta_1 x_i - \epsilon_i$,

it feels plausible to think that a normalization of the response also has a normalizing effect on the random effect distribution. This should be reflected in a lower number of classes required for the transformation model. This effect was slightly visible in Example 5.3, but not clearly visible in Examples 5.1 and 5.2. However, in these two examples we found that there was either evidence for transformation, or for the use of a random effect model, but not for both at the same time, which could be considered another facet of the same effect. In this connection, it is worth noting that components of the mixture do not necessarily correspond to clusters of participants in the population, especially if the mixture is used to account for skewness (Bonate , 2011). In the context of factor mixture models, Lubke and Muthén (2005) raised the question of whether an extra class can provide a useful information about the heterogeneity.

Concluding, it is clear that skewness and heterogeneity are related concepts, and that statistical methods which tackle one of these problems will also implicitly address the other to some extent. However, the precise nature of this interplay is often not so clear. While the proposed approach can help us to handle this trade-off from a modelling point of view, the connection between normal mixture models and transformations to achieve homogeneous variance deserves further attention.

Appendix

A. Additional simulation results

In Tables 8, 9, and 10, we provide additional simulation results for the cases $K = 1$, $K = 2$, and $K = 8$, respectively, complementing the results for $K = 4$ presented in Table 2 in the main text. Data were simulated using Design (4.1), and λ and β were estimated using known K.

We further provide additional results for simulation under design (4.2), for the cases $J = 20$, $n_j = 10$, as well as $J = 40$, $n_j = 5$. In both cases $K = 4$ was used for simulation and estimation. Results are provided in Figure 6 and Table 11.

B. Control charts

The content of this section is made available as part of the supplementary material.

Supplementary materials

Supplementary materials, including R code reproducing some of the analyses in this paper, as well as Appendix B, are available from http://www.statmod.org/smij/archive.html.

$K=1$	$n=100$				$n=200$			
	$\lambda=0$	$\lambda=0.5$	$\lambda=1$	$\lambda=2$	$\lambda=0$	$\lambda=0.5$	$\lambda=1$	$\lambda=2$
Mean($\hat{\lambda}$)	0.0002	0.4947	0.9898	1.9374	0	0.5055	1.0105	2.0039
Median($\hat{\lambda}$)	0	0.5	1	2	0	0.5	1	2
β_1	3	3	3	3	3	3	3	3
Mean($\hat{\beta}_1$)	3.1514	5.2816	6.7338	5.5577	2.9968	4.1776	4.8383	5.0099
Median($\hat{\beta}_1$)	2.9998	2.9824	2.9824	2.9602	2.9966	2.9957	2.9943	2.9911
RESD($\hat{\beta}_1$)	0.0893	2.5648	4.4355	4.8507	0.0644	2.2491	2.9084	3.6661
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.0917	0.1526	0.1941	0.1614	0.0616	0.0857	0.0992	0.1028
Median($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.0871	0.0860	0.0850	0.0845	0.0616	0.0615	0.0625	0.0616
β_2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mean($\hat{\beta}_2$)	0.5245	0.8739	1.1125	0.9233	0.5010	0.6966	0.8059	0.8350
Median($\hat{\beta}_2$)	0.5008	0.4963	0.4920	0.4825	0.5012	0.5018	0.5033	0.5003
RESD($\hat{\beta}_2$)	0.0311	0.4518	0.7903	0.8164	0.0196	0.3788	0.4936	0.6189
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0305	0.0507	0.0645	0.0536	0.0205	0.0286	0.0331	0.0343
Median($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0290	0.0287	0.0282	0.0282	0.0205	0.0206	0.0209	0.0208

Table 8: Summary of simulation results for fixed effects model.

Acknowledgements

The first author is grateful to Qassim University for financial support. The second author was partly supported by CRoNoS COST Action IC1408.

$K = 2$	$n = 100$				$n = 200$			
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
Mean($\hat{\lambda}$)	0	0.4977	0.9957	1.9886	0	0.5	0.9980	1.9954
Median($\hat{\lambda}$)	0	0.5	1	2	0	0.5	1	2
β_1	3	3	3	3	3	3	3	3
Mean($\hat{\beta}_1$)	2.9992	3.0820	3.1068	3.1264	3.0030	3.0591	3.0563	3.0748
Median($\hat{\beta}_1$)	2.9972	2.9965	2.9902	2.9669	3.0034	3.0041	3.0013	3.0018
RESD($\hat{\beta}_1$)	0.0914	0.1329	1.3171	0.9462	0.0616	0.0724	0.1247	0.8728
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.0860	0.0811	0.0888	0.0893	0.0610	0.0620	0.0619	0.0622
Median($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.0861	0.0852	0.0845	0.0821	0.0609	0.0608	0.0607	0.0601
β_2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mean($\hat{\beta}_2$)	0.4982	0.5117	0.5161	0.5193	0.5002	0.5094	0.5088	0.5117
Median($\hat{\beta}_2$)	0.4989	0.4974	0.4964	0.4900	0.5008	0.5007	0.4996	0.4978
RESD($\hat{\beta}_2$)	0.0291	0.0435	0.1939	0.1681	0.0204	0.0245	0.0422	0.1392
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0287	0.0294	0.0296	0.0298	0.0203	0.0207	0.0207	0.0208
Median($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0286	0.0285	0.0282	0.0278	0.0203	0.0203	0.0203	0.0201

Table 9: Summary of simulation results for random effects model with $K = 2$.

References

References

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004) Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, **47**, 639–653.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

$K = 8$	$n = 100$				$n = 200$			
	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
Mean($\hat{\lambda}$)	0	0.4817	0.9693	1.8067	0	0.4454	0.8336	1.3028
Median($\hat{\lambda}$)	0	0.4	0.9	1.8	0	0.4	0.7	1.2
β_1	3	3	3	3	3	3	3	3
Mean($\hat{\beta}_1$)	2.9993	6.6694	9.1456	3.4388	2.9989	30.3473	25.6251	2.6043
Median($\hat{\beta}_1$)	3.0017	2.2609	2.1694	1.9714	2.9983	1.5774	1.0086	0.5492
RESD($\hat{\beta}_1$)	0.3297	2.6062	2.4508	1.9403	0.1872	3.5756	2.0646	2.0503
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.2743	0.3507	0.4527	0.2148	0.1552	0.9388	0.8610	0.1008
Median($\widehat{\mathtt{SE}}(\hat{\beta}_1)$)	0.2522	0.1487	0.1447	0.1224	0.1568	0.0543	0.0428	0.0249
β_2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mean($\hat{\beta}_2$)	0.5061	1.1029	1.5366	0.5709	0.5066	5.0861	4.3184	0.4311
Median($\hat{\beta}_2$)	0.5049	0.3897	0.3748	0.3377	0.5050	0.2631	0.1801	0.0935
RESD($\hat{\beta}_2$)	0.1097	0.3968	0.4025	0.3192	0.0663	0.5785	0.3636	0.3344
Mean($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0916	0.1173	0.1500	0.0719	0.0519	0.3129	0.2875	0.0336
Median($\widehat{\mathtt{SE}}(\hat{\beta}_2)$)	0.0835	0.0510	0.0481	0.0409	0.0525	0.0180	0.0141	0.0084

Table 10: Summary of simulation results for random effects model with $K = 8$.

Aitkin, M. A., Francis, B., Hinde, J., and Darnell, R. (2009) Statistical Modelling in R. *Oxford, University Press Oxford*.

Almohaimeed, A. and Einbeck, J. (2020) boxcoxmix: Box-Cox-Type Transformations for Linear and Logistic Models with Random Effects. *R package version 0.28*.

Akaike, H. (1998) Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike*, *Springer,* 199–213.

Asar, Ö., Ilk, O., and Dag, O. (2017) Estimating Box-Cox power transformation

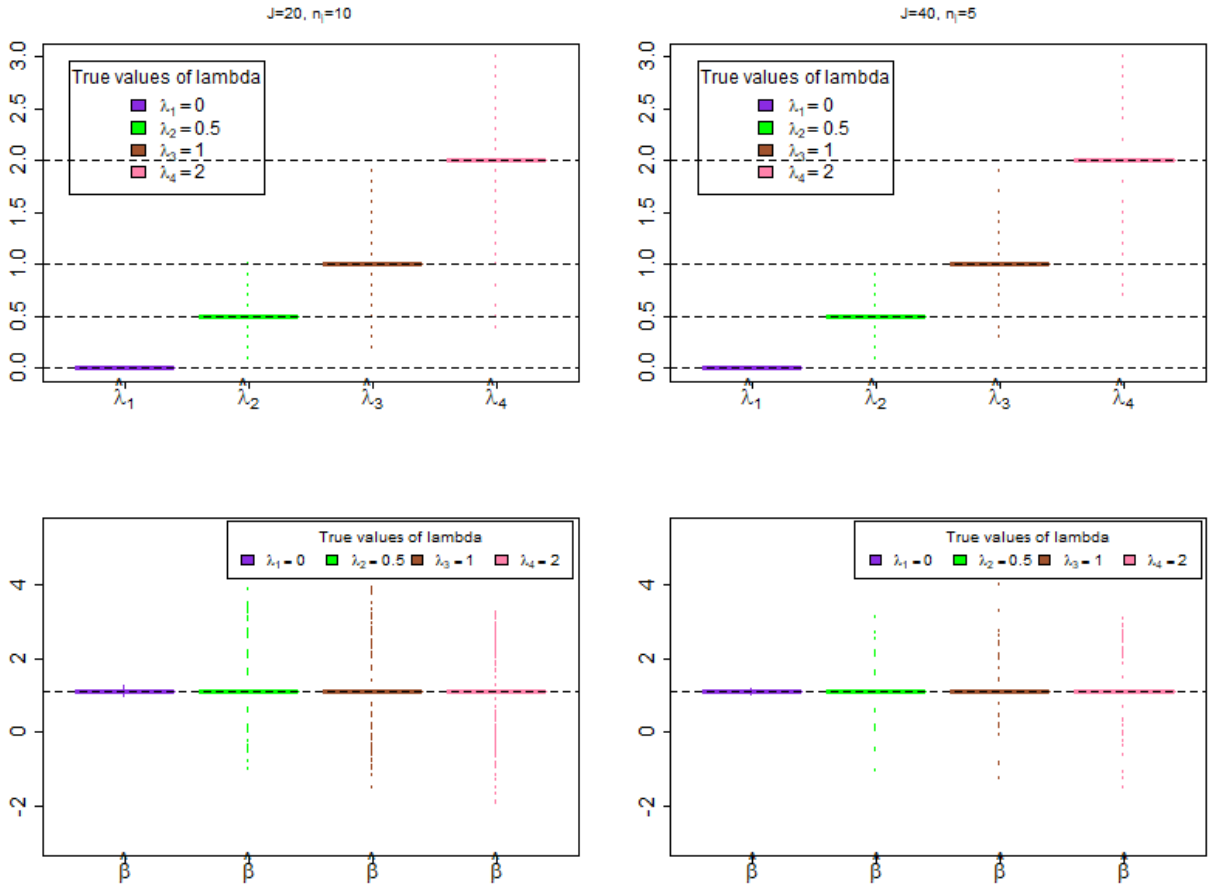

Figure 6: Simulation results for the variance component model with $K = 4$, for $J = 20$, $n_j = 10$ (left) and $J = 40$, $n_j = 5$ (right): Estimates $\hat{\lambda}$ (top) and $\hat{\beta}$ (bottom; logarithmic scale), in each plot for true $\lambda_\ell = 0, 0.5, 1, 2$ (from left to right). Horizontal lines indicate true values.

parameter via goodness–of–fit tests. *Communications in Statistics — Simulation and Computation*, **46**, 91–105.

Bhat, H. S. and Kumar, N. (2010) On the derivation of the Bayesian information criterion. *School of Natural Sciences, University of California*.

Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM Algorithm. *Psychometrika*, **46**, 443–459.

	$J = 20, n_j = 10$				$J = 40, n_j = 5$			
True values	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$
β	3	3	3	3	3	3	3	3
Mean($\hat{\lambda}$)	0	0.5074	1.0143	2.0122	0.0000	0.5024	1.0054	2.0055
Median($\hat{\lambda}$)	0	0.5	1	2	0	0.5	1	2
Mean($\hat{\beta}$)	3.0020	3.6028	3.8705	3.7526	2.9994	3.1261	3.2140	3.1785
Median($\hat{\beta}$)	3.0003	3.0003	3.0003	3.0002	3	3.0001	3.0001	3.0001
RESD($\hat{\beta}$)	0.0054	0.0055	0.0055	0.0055	0.0044	0.0044	0.0044	0.0044
Mean($\widehat{\mathtt{SE}}(\hat{\beta})$)	0.0146	0.0232	0.0282	0.0265	0.0067	0.0083	0.0100	0.0094
Median($\widehat{\mathtt{SE}}(\hat{\beta})$)	0.0043	0.0043	0.0043	0.0043	0.0042	0.0042	0.0042	0.0042

Table 11: Summary of simulation results for variance component model, for $K = 4$ and different sample size configarations totalling to $n = 200$.

Bonate, Peter L. (2011). Pharmacokinetic-Pharmacodynamic Modeling and Simulation. 2nd Edition, Springer.

Box, G. E. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252.

Böhning, D., Kuhnert, R., and Rattanasiri, S. (2008). Meta-analysis of Binary Data using Profile Likelihood. Chapman & Hall/CRC.

Böhning, D. (2000). Computer-assisted Analysis of Mixtures and Applications. Meta-analysis, Disease Mapping and others. London: Chapman & Hall.

Butler, S. M. and Louis, T. A. (1992) Random effects models with non-parametric priors. *Statistics in Medicine*, **11**, 1981–2000.

Carroll, R. J. (1982) Prediction and power transformations when the choice of power

is restricted to a finite set. *Journal of the American Statistical Association*, **77**, 908–915.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. (2018). Model Selection for Mixture Models – Perspectives and Strategies. Handbook of Mixture Analysis, CRC Press.

Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H., Ying, L., et al. (2014) Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, **26**, 105–109.

Clark, T. S. and Linzer, D. A. (2015) Should I use fixed or random effects? *Political Science Research and Methods*, **3**, 399–408.

Davies, R. (1987) Mass point methods for dealing with nuisance parameters in longitudinal studies. In: Crouchley, R. (Ed.) *Longitudinal Data Analysis*. Aldershot, Hants: Avebury, p. 88–109.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, **39**, 1–38.

Einbeck, J. and Hinde, J. (2006) A note on NPML estimation for exponential family regression models with unspecified dispersion parameter. *Austrian Journal of Statistics*, **35**, 233–243.

Einbeck, J., Hinde, J., and Darnell, R. (2007) A new package for fitting random effect models: The npmlreg package. *R News*, **7**, 26–30.

Gurka, M. J. (2004) The Box-Cox transformation in the general linear mixed model for longitudinal data. PhD thesis, *The University of North Carolina*.

Gurka, M. J., Edwards, L. J., Muller, K. E., and Kupper, L. L. (2006) Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 273–288.

Heckman, J. and Singer, B. (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration aata. *Econometrica: Journal of the Econometric Society*, **52**, 271–320.

Hinde, J. and Demetrio, C. (1998) Overdispersion: Models and Estimation. *Short Course for SINAPE 1998. Associate Brazilian Statistics Bureau.*

Keene, O. N. (1995) The log transformation is special. *Statistics in Medicine*, **14**, 811–819.

Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–811.

Leroux, B. G. and Puterman, M. L. (1992) Maximum-penalized-likelihood estimation for independent and Markov–dependent mixture models. **48**, *Biometrics*, 545–558.

Lindsay, B. G. (1983) The geometry of mixture likelihoods: a General Theory. *The Annals of Statistics*, **11**, 86–94.

Lubke, G. H. and Muthén, B. (2005) Investigating population heterogeneity with factor mixture models. *Psychological Methods*, **10**, 21-39.

Lukoĉienė, O. (2010) Latent class models for categorical data with a multilevel structure. PhD thesis, *Tilburg University.*

Lukoĉienė, O. and Vermunt, J. K. (2009) Determining the number of components in mixture models for hierarchical data. *Advances in Data Analysis, Data Handling*

and Business Intelligence - Proceedings of the 32nd Annual Conference of the GfKl e.V., Hamburg, July 16-18, 2008, p. 241-249.

Maruo, K., Yamaguchi, Y., Noma, H., and Gosho, M. (2017) Interpretable inference on the mixed effect model with the Box-Cox transformation. *Statistics in Medicine*, **36**, 2420–2434.

Prasad A. Naik, Peide Shi, and Chih-Ling Tsai (2007) Extending the Akaike Information Criterion to mixture regression models, *Journal of the American Statistical Association*, **102**, 244–254.

McLachlan, G. and Peel, D. (2004) Finite Mixture Models. *John Wiley & Sons*.

Osborne, J. W. (2010) Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, **15**(12).

Osborne, J. W. (2013) Best Practices in Data Cleaning: a complete guide to everything you need to do before and after collecting your data. *SAGE*.

Pearson, K. (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, **186**, 343–414.

Piepho, H.-P. and McCulloch, C. E. (2004) Transformations in mixed models: Application to risk analysis for a multienvironment trial. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 123–137.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2016) nlme: Linear and Nonlinear Mixed Effects Models. *R package version 3.1-128*.

Qarmalah, N. M., Einbeck, J., and Coolen, F. P. A. (2018) k–Boxplots for mixture data. *Statistical Papers*, **59**, 513–528.

R Core Team: R (2016) A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Sakia, R. (1992) The Box-Cox transformation technique: a review. *The Statistician*, **41**, 169–178.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Steele, R.J. and Raftery, A. (2010) Performance of Bayesian model selection criteria for Gaussian mixture models. In: Chen, M.-H. et al. (Eds.) *Frontiers of Statistical Decision Making and Bayesian Analysis*, p. 113-–130. Springer.

Wang, L. (2004) Parameter estimation for mixtures of generalized linear mixed-effects models. PhD thesis, *The University of Georgia.*

Wang, P., Tsai, G.-f., and Qu, A. (2012) Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association*, **107**, 725–736.