Nonparametric estimation of the mixing distribution in mixed models with random intercepts and slopes

by

Rabih Saab
B.Sc., York University, 2005
M.A., York University, 2006

A DISSERTATION Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Mathematics and Statistics

Nonparametric estimation of the mixing distribution in mixed models with random intercepts and slopes

by

Rabih Saab
B.Sc., York University, 2005
M.A., York University, 2006

Supervisory Committee

Dr. Mary L. Lesperance, Supervisor
(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Departmental Member
(Department of Mathematics and Statistics)

Dr. David Giles, Outside Member
(Department of Economics)

Supervisory Committee

Dr. Mary L. Lesperance, Supervisor
(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Departmental Member
(Department of Mathematics and Statistics)

Dr. David Giles, Outside Member
(Department of Economics)

ABSTRACT

Generalized linear mixture models (GLMM) are widely used in statistical applications to model count and binary data. We consider the problem of nonparametric likelihood estimation of mixing distributions in GLMM's with multiple random effects. The log-likelihood to be maximized has the general form

$$l(G) = \sum_i \log \int f(y_i, \boldsymbol{\gamma}) dG(\boldsymbol{\gamma})$$

where $f(., \boldsymbol{\gamma})$ is a parametric family of component densities, y_i is the i^{th} observed response dependent variable, and G is a mixing distribution function of the random effects vector $\boldsymbol{\gamma}$ defined on Ω.

The literature presents many algorithms for maximum likelihood estimation (MLE) of G in the univariate random effect case such as the EM algorithm (Laird, 1978), the intra-simplex direction method, ISDM (Lesperance and Kalbfleish, 1992), and vertex exchange method, VEM (Böhning, 1985). In this dissertation, the constrained Newton method (CNM) in Wang (2007), which fits GLMM's with random intercepts only, is extended to fit clustered datasets with multiple random effects. Owing to the general equivalence theorem from the geometry of mixture likelihoods (see Lindsay, 1995), many NPMLE algorithms including CNM and ISDM maximize the directional derivative of the log-likelihood to add potential support points to the mixing distribution G. Our method, *Direct Search Directional Derivative* (DSDD), uses a directional

search method to find local maxima of the multi-dimensional directional derivative function. The DSDD's performance is investigated in GLMM where f is a Bernoulli or Poisson distribution function. The algorithm is also extended to cover GLMM's with zero-inflated data.

Goodness-of-fit (GOF) and selection methods for mixed models have been developed in the literature, however their application in models with nonparametric random effects distributions is vague and ad-hoc. Some popular measures such as the Deviance Information Criteria (DIC), conditional Akaike Information Criteria (cAIC) and R^2 statistics are potentially useful in this context. Additionally, some cross-validation goodness-of-fit methods popular in Bayesian applications, such as the conditional predictive ordinate (CPO) and numerical posterior predictive checks, can be applied with some minor modifications to suit the non-Bayesian approach.

Contents

Supervisory Committee ii

Abstract iii

Table of Contents v

List of Tables viii

List of Figures x

Acknowledgements xii

Dedication xiii

1 Introduction 1

2 Nonparametric likelihood estimation: Theory and algorithm 6

2.1 The geometry of mixture likelihoods 6

2.2 Identifiability 10

2.3 Algorithm 11

2.3.1 DSDD Algorithm 12

2.4 Direct search methods 14

2.4.1 Convergence Analysis 17

2.5 Computing the Mixing Proportions 19

3 Binary and Poisson models: Simulations and a case study 22

3.1 Bernoulli model 22

3.1.1 Two sample simulations 23

3.1.2 1000 dataset binary data simulations 28

3.2 Poisson model 35
3.2.1 Two sample simulations 36
3.2.2 1000 Sample Simulations 39
3.3 National Basketball Association (NBA) case study 41
3.3.1 Data description 43
3.3.2 Analysis 43

4 Zero-inflated data **46**
4.1 Zero-inflated Poisson (ZIP) models 47
4.2 Zero-inflated Binomial (ZIB) models 48
4.2.1 ZIB Simulation 48
4.3 Hurdle models 49
4.3.1 Hurdle model simulation 51
4.3.2 Pharmaceutical Data 51

5 Fitted values and random effects predictions **55**
5.1 Random effects predictors 56
5.1.1 Random effects posterior mean 56
5.1.2 Random effect modes 61
5.2 *Posterior* means for μ 66
5.3 Result comparison 67

6 Model Selection and fit **72**
6.1 Penalized information criteria for model selection 73
6.1.1 Deviance Information Criterion 74
6.1.2 Conditional Akaike Information 80
6.2 Cross-validation and predictive based goodness-of-fit methods 84
6.2.1 Conditional Predictive Ordinate 84
6.2.2 Numerical posterior predictive checks 88
6.3 Coefficient of determination for generalized linear mixed models . . . 92

7 Discussion and topics for further research **96**
7.1 Algorithm performance 97
7.2 Goodness-of-fit Summary 98

A NBA Dataset summary **100**

Bibliography **102**

List of Tables

Table 3.1 Specification for binary simulations 24
Table 3.2 Binary Sample 1 simulation results 25
Table 3.3 Binary Sample 2 simulation results 25
Table 3.4 Bernoulli model: 1000 sample simulation results 34
Table 3.5 Specification for Poisson simulations 37
Table 3.6 Poisson Sample 1 simulation results 37
Table 3.7 Poisson Sample 2 simulation results 38
Table 3.8 DSDD and glmer results for the NBA dataset. 44

Table 4.1 ZIB model simulation description and DSDD estimations 49
Table 4.2 Description of the hurdle model simulation 52
Table 4.3 DSDD results of the hurdle simulation 52
Table 4.4 Pharmaceutical case study: EM and DSDD comparison 54

Table 5.1 Estimated and true random effect comparison: Squared difference summaries for the simulated datasets 58
Table 5.2 Comparing posterior modes of the DSDD and glmer algorithms 66
Table 5.3 Brier and RSB scores of three methods for computing fitted values 68
Table 5.4 Comparing DSDD and glmer Brier scores 69
Table 5.5 Pharmaceutical case study fitted values 71

Table 6.1 Comparison of DSDD and glmer models using DIC_1 76
Table 6.2 comparison of DSDD and glmer models using DIC_2 78
Table 6.3 comparison of DSDD and glmer models using DIC_μ 79
Table 6.4 DIC's for Sample 1 Bernoulli simulation with and without random slopes 79
Table 6.5 Comparing the LPML measures of DSDD and glmer 86
Table 6.6 Numerical posterior predictive checks: discrepancy summaries . 90

Table 6.7 Coefficients of determination for the simulated datasets and NBA case study 95

Table A.1 NBA data, summary statistics and estimates 101

List of Figures

Figure 2.1 Plot of a normal mixture convex hull for two normal observations y_1 and y_2 9
Figure 2.2 Direct Search method description 15

Figure 3.1 Gradient surfaces evaluated at the MLE, $\hat{G}$, for two Bernoulli samples 27
Figure 3.2 Binary Sample 1 simulation: average probabilities of success per cluster estimated by the DSDD and glmer routines compared to the true probabilities. 29
Figure 3.3 Binary Sample 2 simulation: average probabilities of success per cluster estimated by the DSDD and glmer routines compared to the true probabilities 30
Figure 3.4 Bernoulli model: two sample simulation results 31
Figure 3.5 Bernoulli Sample 1: 1000 simulation results 32
Figure 3.6 Bernoulli Sample 2: 1000 simulation results 33
Figure 3.7 Poisson model: two sample simulation results 40
Figure 3.8 Poisson model: 1000 simulation results 42

Figure 5.1 Bernoulli Sample 1: within cluster averages of DSDD fitted values using random effects posterior means 59
Figure 5.2 Poisson Sample 1: within cluster averages of DSDD fitted values using random effects posterior means 60
Figure 5.3 Bernoulli Sample 1: within cluster averages of DSDD fitted values using random effects posterior modes 64
Figure 5.4 Poisson Sample 1: within cluster averages of DSDD fitted values using random effects posterior modes 65
Figure 5.5 NBA data: DSDD and glmer fitted values 70

Figure 6.1 Comparing the CPO index of the DSDD and glmer fitted models 87

ACKNOWLEDGEMENTS

I would like to thank:

my supervisor, Dr. M. Lesperance, for her encouragement, patience, and mentorship. Without her guidance and persistent help this dissertation would not have been possible.

The University of Victoria and NSERC, for the fellowships and scholarships.

Last but not least, I would like to express my sincere gratitude to my aunt and uncle, Lina and Wadid Saab, for their support, kindness and generosity.

DEDICATION

For my parents, who offered me unconditional love and support throughout the course of this dissertation.

Chapter 1

Introduction

Generalized linear mixed models (GLMM's) are an extension of generalized linear models that introduce random effects to the linear predictor. The presence of packages that fit GLMM's in standard software such as SAS, STATA and R is indicative of the extensive use of these models in modern research. For example the lme4 package in R (Bates et al., 2011) fits GLMM's with multivariate normal random effects with the function glmer, and the package glmmML (Broström et al., 2011) fits the mixed model using a maximum likelihood approach. Mixture models are used when modelling independent observations, $y_1, \ldots, y_n$ arising from an assumed parametric distribution $f(y_i, \boldsymbol{\gamma}_i)$ where the parameter $\boldsymbol{\gamma}_i$ varies according to a mixing distribution $G(\boldsymbol{\gamma})$ to accommodate the heterogeneity that is often exhibited in natural populations.

The use of mixture distributions started with Pearson (1894) who used the method of moments to estimate a normal mixture with two support points. Theoretical justifications were later provided by Kiefer and Wolfowitz (1956) who showed that under certain conditions the nonparametric maximum likelihood estimate (NPMLE) of the mixing distribution converges almost surely to the true mixture, which lead to an extensive literature on properties of NPMLE's and their computational challenges.

A GLMM is typically specified as a conditional distribution of the data $\mathbf{y}$ given the random effects vector, $\boldsymbol{\gamma}$, the random effect covariate vector $\mathbf{x}$ and the fixed effect covariate vector $\mathbf{z}$. GLMM's are useful for analysing repeated measurements or clustered observations and dealing with overdispersion issues that may be observed among outcomes with binomial or Poisson distributions. Parametric mixture models assume that the mixing distribution belongs to a specific parametric family, and parameter estimation is performed using maximum likelihood estimation, which requires numerical integration techniques for the score equations and information matrices (Brillinger and Preisler, 1986; Crouch and Spiegelman, 1999; Hinde, 1982). As an alternative, nonparametric maximum likelihood estimation provides a discrete estimate of the mixing distribution when there are no parametric assumption placed on it (Laird, 1978).

In a clustered data setting, the n_i observations of the i^{th} cluster, $y_{i1}, \ldots, y_{in_i}$, are conditionally independent and modelled as

$$y_{i1}, \ldots, y_{in_i} | \boldsymbol{\gamma}_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i} \mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i} \sim \prod_{j}^{n_i} f(y_{ij} | \boldsymbol{\gamma}_i, \mathbf{z}_{ij}, \mathbf{x}_{ij}) \tag{1.1}$$

where $\mathbf{z}_{ij}$ and $\mathbf{x}_{ij}$ are the fixed and random effects covariate vectors respectively characterizing observation y_{ij} for $i = 1, \ldots, n$. A transformation of the mean $\mu_{ij} = E[y_{ij} | \boldsymbol{\gamma}_i, \mathbf{z}_{ij}, \mathbf{x}_{ij}]$ is typically used to linearise the model,

$$\Psi(\mu_{ij}) = \mathbf{z}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \boldsymbol{\gamma}_i. \tag{1.2}$$

We suppose that the random effects $\boldsymbol{\gamma}_i$ are independent random variables with distribution function $G_{\boldsymbol{\gamma}}$. Parametric mixture models often assume $G_{\boldsymbol{\gamma}}$ is normal as in Stiratelli, Laird and Ware (1984) and McCulloch et al. (2008). Thus the majority of

estimation methods that have been developed are based on the normality assumption of random effects (Breslow and Clayton, 1993). The estimation of the fixed effects are still relatively robust when G is misspecified. However, such misspecifications can compromise predictions and inferences made on the random effects (Tao et al., 1999; Verbeke and Lesaffre, 1996), and this problem can be crucial in longitudinal studies where interest lies in cluster or subject specific effects. Thus the need to relax these parametric assumptions may be more desirable in order to make robust inferences.

The marginal density of the response variable, $\mathbf{y}$, has a general form:

$$f(y, G_{\boldsymbol{\gamma}}) = \int_{\Omega} f(y, \boldsymbol{\gamma}) dG(\boldsymbol{\gamma}) \tag{1.3}$$

where $f(y, \boldsymbol{\gamma})$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \Omega \subset \mathbb{R}$, is the component density. Given a random sample $y_1 \ldots y_n$, the log-likelihood resulting from density (1.3) takes the form

$$l(G_{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \log \left\{ \int_{\Omega} f(y_i, \boldsymbol{\gamma}) dG_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \right\}. \tag{1.4}$$

For a discrete $G_{\boldsymbol{\gamma}}$, we can write $G_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \sum_{k=1}^{m} \pi_k \delta_{\boldsymbol{\gamma}_k}$, where $\boldsymbol{\gamma}_k \in \Omega$ is the k^{th} random effect component, δ_k puts mass 1 at $\boldsymbol{\gamma}_k$ and π_k is the weight of the k^{th} component such that $\sum \pi_k = 1$ and $\pi_k > 0$. Subsequently we can write the log-likelihood in (1.4) as

$$l(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{m} \pi_k f(y_i, \boldsymbol{\gamma}_k) \right\}. \tag{1.5}$$

Our goal is to find the MLE for $G_{\boldsymbol{\gamma}}$ that will maximize the log-likelihood (1.5). Not knowing m in advance prevents us from using the traditional optimization techniques. Earlier methods available in the literature handle univariate random effects and include the expectation-maximization (EM) algorithm (Laird, 1978), which was later

improved by combining it with a gradient step to achieve global convergence when m is known (Böhning, 2003), the vertex direction method (VDM) (Fedorov, 1972; Wu, 1978a and b), the vertex exchange method (VEM) (Böhning, 1985) and the intra-simplex direction method (ISDM) (Lesperance and Kalbfleisch, 1992). Wang (2007) proposed the Constrained Newton Method (CNM) which is a modification of Atwood's (1976) quadratic method. He used a linear regression formulation to solve the quadratic programming sub-problem for estimating the mixing weights for which Atwood did not offer a detailed solution. Wang also showed that his method converges at a faster rate than previously published algorithms.

We present a new algorithm that computes nonparametric maximum likelihood estimates of a mixing distribution for generalized linear mixed models containing multiple random effects with Poisson and binomial response variables. The algorithm, which we refer to as the *Direct Search Directional Derivative* (DSDD), incorporates the quadratically convergent CNM to estimate the mixing proportions and uses a direct search method to identify maxima of the gradient function to include as mixing distribution support points.

Once the mixing distribution and regression parameters have been estimated we proceed to computing fitted values. Subsequently, we explore model selection and goodness-of-fit methods for the various mixed models studied in the dissertation. From the several techniques available in the literature, we investigate likelihood based penalized information criteria methods such as the deviance information criteria (Spiegelhalter, 2002) and the conditional Akaike information criteria (Vaida and Blanshard, 2005). Additionally, model adequacy methods that emphasize the observable response variable rather than the parameters in the model are investigated. Such methods include the conditional predictive coordinate (CPO) by Geisser (1979),

numerical posterior predictive checks and variations of R^2 measures to account for the random effects in the models.

The dissertation is organized as follows: in Chapter 2 a brief review of the geometry of mixture likelihood of Lindsay (1983) is provided to introduce the DSDD algorithm, which is presented in Section 2.3. Chapter 3 investigates the performance of the algorithm with different simulation studies for Bernoulli and Poisson models. We also analyse a case study with data from the National Basketball Association (NBA) using the DSDD algorithm and the glmer routine of the lme4 library in R (Bates, 2011). In Chapter 4 the DSDD is extended to handle zero-inflated data and its performance is tested with simulated samples and a pharmaceutical case study presented in Min and Agresti (2005). Chapter 5 presents different methods to calculate the fitted values for clustered datasets with random effects. Lastly, Chapter 6 explores possible goodness-of-fit tests for mixture models.

Chapter 2

Nonparametric likelihood estimation: Theory and algorithm

2.1 The geometry of mixture likelihoods

The task of maximizing the log-likelihood in (1.5) over the set of all possible distributions, $\mathbb{M}$, can be computationally expensive. Several suggestions have been provided in the literature to compute the nonparametric MLE of the distribution $G_{\boldsymbol{\gamma}}$. The geometry of mixture likelihoods shown in Lindsay (1983) provides the framework for such estimation. In what follows, we give a brief review of his work.

For known values of fixed coefficients $\boldsymbol{\beta}$, we suppress the dependence of the log-likelihood on $\boldsymbol{\beta}$. Suppose we have a sample of n observations $y_1, \ldots, y_n$, let $\mathbf{L}_{\boldsymbol{\gamma}} = \{L_1(\boldsymbol{\gamma}), \ldots, L_T(\boldsymbol{\gamma})\}$ represent their T respective distinct likelihoods. The log-likelihood for a given mixing distribution $G_{\boldsymbol{\gamma}}$ can be written as:

$$l(G_{\boldsymbol{\gamma}}) = \sum_{t=1}^{T} r_t \log \int L_t(\boldsymbol{\gamma}) dG_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \tag{2.1}$$

where r_t is the multiplicity of $L_t(\boldsymbol{\gamma})$, for example if the observed response vector is $2, 2, 4, 5$ then $r_1 = 2, r_2 = 1, r_3 = 1$ and $T = 3$. Let $G_{\boldsymbol{\gamma}} = \delta_{\boldsymbol{\gamma}}$ be the mixing distribution placing mass 1 on the random component $\boldsymbol{\gamma}$, we denote by $\Gamma = \{\mathbf{L}_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \Omega\}$, where Ω is the domain of $\boldsymbol{\gamma}$, as the set of all likelihood components with mixing distribution $G_{\boldsymbol{\gamma}}$. Note that Γ is a set in $\mathbb{R}^T$. The convex hull of Γ, denoted by $conv(\Gamma)$ is the set of all convex combinations of elements in Γ:

$$conv(\Gamma) = (\sum_k \pi_k \mathbf{L}_{\boldsymbol{\gamma}_k} : \mathbf{L}_{\boldsymbol{\gamma}_k} \in \Gamma, \pi_k \geq 0, k = 1, \ldots, K, \sum \pi_k = 1, K = 1, 2, \ldots).$$

If $conv(\Gamma)$ is compact then the stronger result that the convex hull of Γ is the set of all the mixture vectors holds:

$$conv(\Gamma) = \left\{L_{G_{\boldsymbol{\gamma}}} : G_{\boldsymbol{\gamma}} \in \mathbb{M}\right\}$$

where $\mathbb{M}$ is the set of all distribution functions. As a result, maximizing (2.1) over $G_{\boldsymbol{\gamma}}$ is equivalent to maximizing the concave function

$$l(\hat{G}_{\boldsymbol{\gamma}}) = \sup_{G_{\boldsymbol{\gamma}} \in \mathbb{M}} \sum_t r_t \log \int L_t(\gamma) dG_{\boldsymbol{\gamma}}(\gamma) \tag{2.2}$$

over the convex region conv(Γ). Morevoer, if $conv(\Gamma)$ is compact, the strict concavity of (2.2) allows us to use Caratheodory's theorem to show that there exists a unique vector $\hat{L}$ maximizing $l(\hat{G}_{\boldsymbol{\gamma}})$, and the point $\hat{L}$ can be expressed as a mixture, $L_{\hat{G}_{\boldsymbol{\gamma}}}$, where $\hat{G}_{\boldsymbol{\gamma}}$ has at most $T+1$ support points. References include Lindsay (1983, theorem 3.1), Roberts and Varberg (1973).

The importance of the geometric interpretation above is that it allows the maximization problem to be solved over a set in $\mathbb{R}^{T+1}$ instead of the infinite dimensional

space. For given distributions $G_{\boldsymbol{\gamma}_0}$ and $G_{\boldsymbol{\gamma}_1}$, the directional derivative of the log-likelihood from the point $\mathbf{L}_{G_{\boldsymbol{\gamma}_0}}$ to $\mathbf{L}_{G_{\boldsymbol{\gamma}_1}}$ is given by

$$\begin{aligned} D(\mathbf{L}_{G_{\boldsymbol{\gamma}_1}}; \mathbf{L}_{G_{\boldsymbol{\gamma}_0}}) &= \lim_{h \to 0} \frac{l\left\{(1-h)G_{\boldsymbol{\gamma}_0} + hG_{\boldsymbol{\gamma}_1}\right\} - l(G_{\boldsymbol{\gamma}_0})}{h} \\ &= \sum_{t=1}^{T} \frac{r_t \left\{L_t(G_{\boldsymbol{\gamma}_1}) - L_t(G_{\boldsymbol{\gamma}_0})\right\}}{L_t(G_{\boldsymbol{\gamma}_0})}. \end{aligned} \tag{2.3}$$

Let δ_γ be a mixing distribution that places mass one on γ. We write the directional derivative from a point $L_{G_{\boldsymbol{\gamma}}} \in conv(\Gamma)$ to a point $L_\gamma = L_{\delta_\gamma}$ as

$$D(\boldsymbol{\gamma}, G_{\boldsymbol{\gamma}}) = D(L_{\boldsymbol{\gamma}}; L_{G_{\boldsymbol{\gamma}}}).$$

If $\hat{G}_{\boldsymbol{\gamma}}$ maximizes $l(G_{\boldsymbol{\gamma}})$, it must be equivalently characterized by the following three conditions (see theorem 4.1 from Lindsay, 1983):

1. $\hat{G}_{\boldsymbol{\gamma}}$ maximizes $l(G_{\boldsymbol{\gamma}})$
2. $\hat{G}_{\boldsymbol{\gamma}}$ minimizes $\sup_{\gamma \in \Omega} D(\boldsymbol{\gamma}, G_{\boldsymbol{\gamma}})$
3. $\sup_{\gamma \in \Omega} D(\boldsymbol{\gamma}, \hat{G}_{\boldsymbol{\gamma}})$=0.

The characterization of directional derivatives is illustrated in Figure 2.1, which shows the curve $\Gamma = [\{\phi(1-\gamma), \phi(4-\gamma)\}; \gamma \in \mathbb{R}]$ for two normal densities of mean γ and variances 1 with two observed points from the mixture $y_1 = 1$ and $y_2 = 4$, where ϕ is the standard normal probability density function. The convex hull of Γ is then defined by $conv(\Gamma) = \left\{(p_1, p_2) \in \mathbb{R}^2; p_1 = \int \phi(1-\gamma) dG(\gamma), p_2 = \int \phi(4-\gamma) dG(\gamma), G \in \mathbb{M}\right\}$, while the log-likelihood is $l(G) = \log(p_1) + \log(p_2)$. The maximum likelihood estimator, denoted by L_3, has directional derivatives that are negative or zero towards all points L_γ of $conv(\Gamma)$. The points L_1 and L_2 are not optimal because their directional

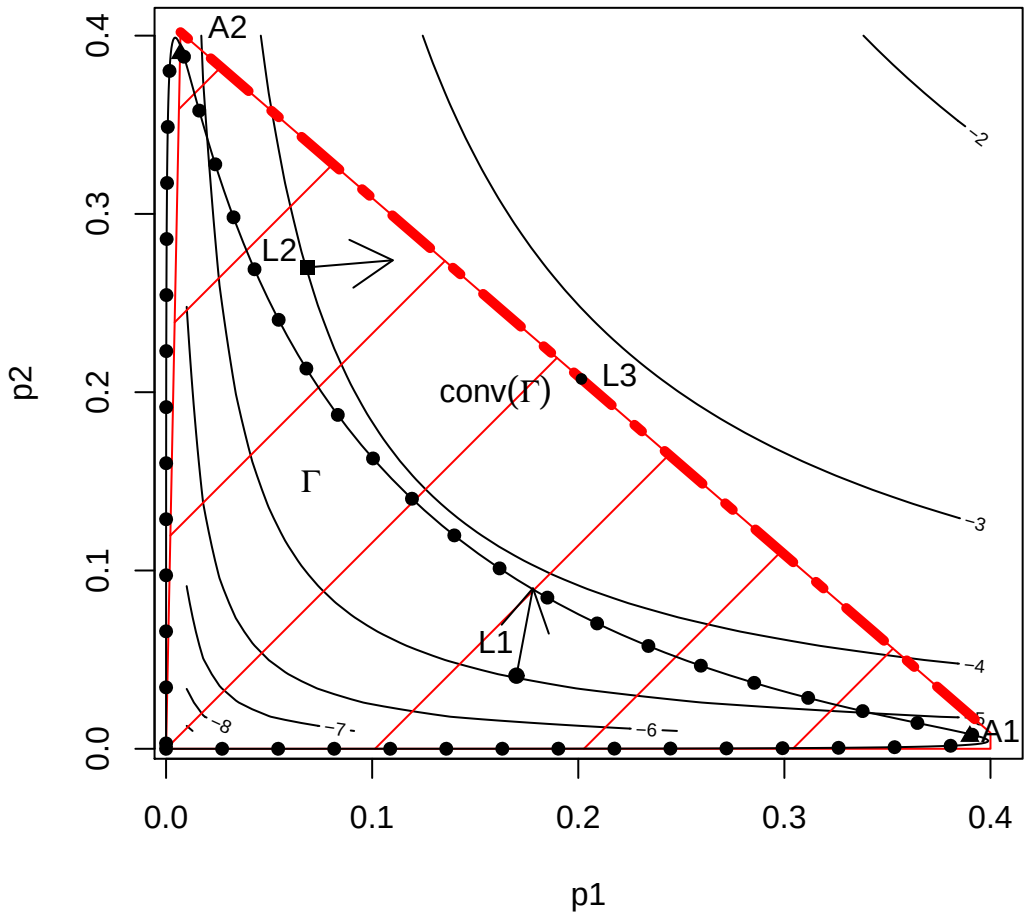

Figure 2.1: Plot of a normal mixture convex hull showing the curve Γ _._._, $conv(\Gamma)$ / / /, contours of the log-likelihood for two normal observations y_1 and y_2 ——. The plot shows the convex hull $conv(\Gamma) = \{(p_1, p_2) \in \mathbb{R}^2; p_1 = \int \phi(1-\gamma)dG(\gamma), p_2 = \int \phi(4-\gamma)dG(\gamma), G \in \mathbb{M}\}$ and the maximum likelihood estimator L_3, The points L_1 and L_2 are not optimal because their likelihood can be increased in the direction of A_1 and A_2.

derivatives are positive and non-zero towards the points A_1 and A_2, which shows that the likelihood can be increased in these directions.

2.2 Identifiability

We have just shown the existence of a unique maximizing vector $\hat{\mathbf{L}}$, however that does not address the question of whether the corresponding MLE of the mixing distribution, $\hat{G}_{\boldsymbol{\gamma}}$, is unique. The uniqueness of $\hat{G}_{\boldsymbol{\gamma}}$ depends on the identifiability of the true mixing distribution $G_{\boldsymbol{\gamma}_0}$ and the form of the kernel distribution, $f(y|\boldsymbol{\gamma})$, (Teicher, 1961)

Definition A mixing distribution $G_{\boldsymbol{\gamma}_0}$ is identifiable if $L_{G_{\boldsymbol{\gamma}_1}} = L_{G_{\boldsymbol{\gamma}_0}}$ implies $G_{\boldsymbol{\gamma}_1} = G_{\boldsymbol{\gamma}_0}$

Teicher (1961) established the identifiability of a wide range of families of parametric densities. For example, location and scale models $f(x-\theta)$ and $f(x/\sigma)$ for $\sigma > 0$ respectively), and the additively closed families of kernel distributions, i.e. any family $F(x,\theta)$ with the property that the sum of two random variables with distributions $F(x,\theta_1)$ and $F(x,\theta_2)$ has a distribution $F(x,\theta_1+\theta_2)$.

Maritz (1970, Chapter 2) also established the identifiability of certain mixing distributions. For example if a mixing distribution, $G_{\boldsymbol{\gamma}}$, with a kernel density $f(\mathbf{y}|\theta)$, has finite and distinct support points $\theta_1,\ldots,\theta_k$ with unknown point masses $\pi_1,\ldots,\pi_k$ then $G_{\boldsymbol{\gamma}}$ is identifiable if the k linear equations, $\sum_{i=1}^{k}\pi_i f(d_j|\theta_i)$ where $d_1,\ldots,d_k$ are k distinct values of y, have a unique solution in the k unknown probability masses.

Follmann and Lambert (1991) discussed the identifiability of a logistic regression model with non-random covariate coefficients β and a random intercept, a, having an unknown finite mixing distribution G_a. They provide sufficient conditions to show

that (β, G_a) is identifiable if either the number of trials in the binomial distributions is large enough or the number of covariate vectors that differ in only one coordinate is large enough.

2.3 Algorithm

As a background to the DSDD algorithm, we review previous NPMLE methods in the literature, more importantly the ones based on directional derivatives. All such algorithms are modifications of the Vertex Direction Method (VDM), which was discussed by Wynn (1970) and Fedorov (1972) but got named by Wu (1978a) in the context of optimal designs. The algorithm can be briefly described in the following steps:

1. Starting from an initial estimate G_0 set $j = 0$
2. find $\gamma^* := \arg\max_\gamma D(\gamma, G_j)$
3. find $\epsilon_0 := \arg\max_\epsilon l\{(1-\epsilon)G_j + \epsilon\delta_{\gamma^*}\}$ such that $\epsilon \in [0, 1]$
4. $G_{j+1} := (1-\epsilon_0)G_j + \epsilon_0\delta_{\gamma^*}$
5. set j=j+1 and go to step 2.

The iterations stop when $\arg\max_\gamma D(\gamma, G_j) = 0$. Fedorov (1972), Wu (1978a) and Lindsay (1983) have all shown in different contexts that the VDM converges to the maximum likehood estimate.

Many variants of the generally slow convergent VDM algorithm have been proposed in the literature such as the Vertex Exchange Method (VEM) proposed by Böhning (1985) and the Intra Simplex Direction Method (Lesperance and Kalbfleish,

1992). The DSDD is based on the constrained Newton's method (CNM) developped by Wang (2007), which obtains estimates of mixing distributions in a GLMM with a random intercept and fixed covariates. The CNM algorithm iteratively adds new important points to the support by maximizing the directional derivative function, then it computes the corresponding mixing proportion $\boldsymbol{\pi}$ at a quadratic order of convergence. The algorithm is in itself an extension of Atwood's quadratic method (Atwood, 1976), which was originated in the context of optimal design of experiments, however the CNM converges at a faster rate than other methods in the literature because it adds multiple support points (as in the ISDM algorithm) and discards points with small masses at each iteration, whereas Atwood's quadrature method adds only one support point at a time while collapsing similar ones. Moreover, Wang establishes proof of the method's convergence that depends on weaker conditions than required by Atwood's algorithm.

Wang's CNM algorithm can fit GLMM's with one random intercept and fixed coefficients. We extend Wang's work by including random slopes to the model and computing the NPMLE of the random intercept and slopes' joint distribution as well as the fixed effects parameter $\boldsymbol{\beta}$. The addition of random slopes adds complexity to the problem of computing local maxima of the directional derivative, which is typically a very bumpy function of the intercept and slopes. A direct search method to find the local maxima is used due to its simple implementation and guaranteed convergence.

2.3.1 DSDD Algorithm

We begin with an initial estimate of the mixing distribution $\hat{G}_0$ with finite support and an initial estimate for the non-random coefficients, $\hat{\boldsymbol{\beta}}_0$.

Main Program, Direct Search Directional Derivative Method:

1. $r = 1$; Set $\epsilon > 0$, the convergence criterion; $\hat{G}_r = \arg\max_G l(G, \hat{\boldsymbol{\beta}}_{r-1})$; $\hat{\boldsymbol{\beta}}_r = \arg\max_\beta l(\hat{G}_r, \boldsymbol{\beta})$.

2. **while** $(|l(\hat{G}_{r-1}, \hat{\boldsymbol{\beta}}_{r-1}) - l(\hat{G}_r, \hat{\boldsymbol{\beta}}_r)| > \epsilon$ and $r \leq$ max_number_iterations_β) {

   $\hat{G}_{r+1} = \arg\max_G l(G, \hat{\boldsymbol{\beta}}_r)$ (MAXG)

   $\hat{\boldsymbol{\beta}}_{r+1} = \arg\max_\beta l(\hat{G}_{r+1}, \boldsymbol{\beta})$ (MAXβ)

   $r = r + 1$

   }

Maximizing over G, MAXG:

1. $s = 0$; Initialize $\boldsymbol{\gamma_s}$, starting values for the direct search; the lattice, C on which to perform the direct search; $\hat{G}_{r,s} = \hat{G}_r$, the starting value for G; $\hat{\boldsymbol{\beta}}_r$ takes its value from the main program; Set gradient *tolerance*; Set *precision* levels for removing/combining support points.

2. **For** s in 1 to max_number_iterations_G {

   (a) Use direct search to find maximal gradient using lattice C and starting at $\boldsymbol{\gamma_{s-1}}$; if maximal gradient $<$ *tolerance*, **break** (see Section 3.2).

   (b) Let $\boldsymbol{\gamma_s}$ be the support points of $\hat{G}_{r,s-1} \cup$ the maximal points computed in (a).

   (c) Update the weights $\boldsymbol{\pi}$ for support points $\boldsymbol{\gamma_s}$ using Wang's (2007) method. (See Section 3.3).

   (d) Remove support points with weight less than *precisionR*. Merge support points that are within *precisionM* of each other. The result is $\hat{G}_{r,s}$. }

Maximizing over β, MAXβ: Use a gradient based method to maximize $l(\hat{G}_{r+1}, \boldsymbol{\beta})$ over β.

2.4 Direct search methods

We use a direct search method in step 2(a) of the MAXG algorithm discussed above. This method does not require an explicit use of derivatives and usually works well in practice (Torczon, 1991).

Direct search methods fell out of favour because of their slower rate of convergence compared to other gradient based methods, the lack of a proof of convergence and the emergence of new software that ease the use of more sophisticated numerical techniques. However, direct search methods work well for certain non-linear optimization problems that prove difficult for more sophisticated approaches. Torczon (1991, 1997) revived interest in these methods as effective optimization tools by providing proofs of their convergence. A general description of the method is provided in Hooke and Jeeves (1961):

> We use the phrase "direct search" to describe sequential examination of trial solutions involving comparison of each trial solution with the "best" obtained up to that time together with a strategy for determining (as a function of earlier results) what the next trial solution will be.

The *coordinate search* method is a subclass of direct search methods called *pattern search methods*. Pattern search methods were generally defined in Torczon (1997) in the form of lattices such that iteration steps all lie on the scaled lattice. The lattice used determines the pattern search method and it is independent of the objective function. Torczon introduced a generating matrix $C \in \mathbb{Z}^{N \times p}$, for $p > 2N$ and N the

dimension of the problem, that governs the possible directions that an iteration s can take in order to move to iteration $s+1$. For the case of a random intercept and one random slope ($N = 2$), we used a coordinate search method with constant generating matrix $C_s = C$ with $p = 3^2$ columns that contain all the possible pairs of $\{-1, 0, 1\}$. A trial step at iteration s is defined as any vector of the form $t_s^i = \Delta_s \boldsymbol{c}_i$, where $\boldsymbol{c}_i$ is the i^{th} column of C. In this case the step size is governed by Δ_s and its direction is governed by $\boldsymbol{c}_i$.

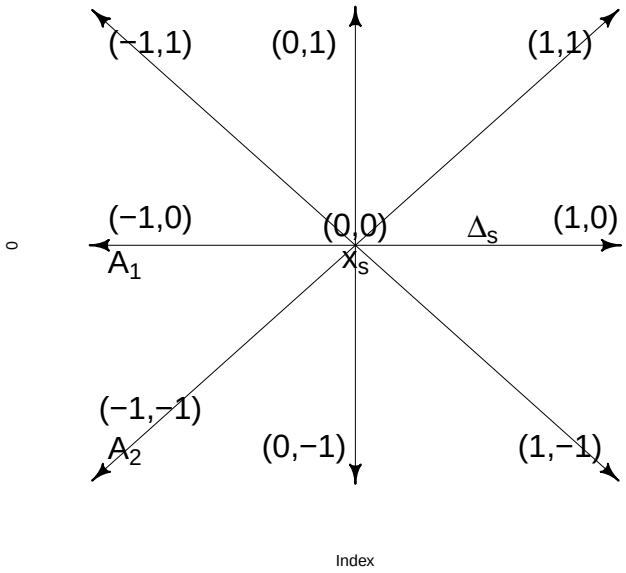

Figure 2.2: Coordinate search for N=2 starting from the point x_s with a step length Δ_s. The columns of the generating matrix C generate the arrows in the above illustration.

Figure 2.2 presents an example of the coordinate search method for $N = 2$ and the investigation of a trial move between two iterations x_s and x_{s+1}. In this case the

generating matrix is given by:

$$C = \begin{bmatrix} 1 & 0 & -1 & 0 & 1 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 1 & -1 & -1 & 1 & 0 \end{bmatrix}.$$

Hence, the column in C with coordinates $(-1, 0)$ indicates a move from the origin towards the point A_1 with length Δ_s. Similarly, the column $(-1, -1)$ generates a move towards the point A_2 and $(0, 0)$ indicates that the next iteration stays at x_s. Given the current iteration x_s, current step size Δ_s, the generating matrix C, and let $max = f(x_s)$ where f is the objective function to maximize, x_{s+1} is computed as:

For $i = 1$ to 9 **do**

- $x_s^i = x_s + \Delta_s \boldsymbol{c}_i$.
- If $f(x_s^i) > max$ then set $x_{s+1} = x_s^i$ and $max = f(x_s^i)$.

The *for* loop is repeated starting at x_{s+1}, and iteration proceeds until no movement occurs at which point the step size, Δ_s, is updated as $\Delta_s = \Delta_s \xi$ for $0 < \xi < 1$.

We use the coordinate search method to search for the maxima of the directional derivative from a given mixing distribution $\hat{G}_r$. We take the initial step size control parameter $\Delta_0 = 0.5$. The choice of Δ_0 should be reasonable relative to the range of the random effects, and we set $\xi = 0.5$ as it is a common practice in direct search methods. If an additional random slope is added to the model, the search directions become three dimensional and the generating matrix C would contain 27 columns instead of 9.

The convergence of the procedure is certain (Torczon, 1997) and its implementation is straightforward. It avoids first and second order derivatives, which are potentially problematic if a Newton's method is used. Although its execution time

per iteration increases with the addition of random effects to the model, we found that the method is numerically stable in practice.

2.4.1 Convergence Analysis

The direct search method described above, more generally the *pattern search method*, has the important feature that the moves at each iteration are decided by considering a rational lattice of points and investigating the behaviour of the objective function at these points. The moves from point to point are made by a constant step only reduced when it is certain that no change in one parameter improves the fit. This feature is important in convergence analysis presented in the literature.

Isolated convergence results have been published on pattern search methods though they all lacked a general theory. Polak (1971), referring to this technique by the *method of local variations*, provided a strong result:

Theorem 2.4.1. *If $\{x_k\}$ is a sequence constructed by the method of local variations, then any accumulation point x' of $\{x_k\}$ satisfies $\nabla f(x') = 0$ (By assumption, $f(x)$ is at least once continuously differentiable).*

Polak notes that the *method of local variations* "cannot jam up at one point" and hence their structure is sufficient to support a global convergence result. Céa (1971) also provided a proof of global convergence of the pattern search method of Hooke and Jeeves (1961) by adding stronger assumptions, mainly that f is strictly convex and $\lim_{||x||\to+\infty} f(x) = +\infty$, to Polak's assumption. Nevertheless, both results establish the global convergence of sequence of iterates produced by Hooke and Jeeves' method.

The importance of both publications is that they prove convergence of an algorithm that does not use the directional derivative. Both Polak's method of local

variations and Hooke and Jeeves' pattern search method analysed by Céa do not allow the step size Δ to be reduced until the condition $f(x_k) \leq f(x_k \pm \Delta e_i)$ (in the case of a minimization problem), where $i = \{1, \ldots, n\}$ and e_i is the i^{th} unit coordinate vector, is satisfied. It is worth noting that from a non-stationary point x_k of f we can always find a descent direction (in the case of a minimization problem) out of the $2n$ possible moves defined by e_i.

More recent results have been published on the convergence of pattern search methods, most notable is the work of Torczon (1997). She restricts the manner of change of the step size; Céa and Polak divide the step size Δ by 2 when there is no improvement in the objective function but Torczon allows a more general approach by making the step size $\Delta_{k+1} = \tau^{\omega}\Delta_k$ for $\tau > 1$ and ω being an integer that is often negative in order to achieve a reduction in step size.

Torczon derives a general theory of global convergence by providing a proof to the following theorem, which guarantees the convergence of at least one sequence of iterates to a stationary point.

Theorem 2.4.2. *Assume that $L(x_0) = \{x | f(x) \leq f(x_0)\}$ is compact and that f is continuously differentiable on a neighbourhood of $L(x_0)$. Then the sequence of iterates $\{x_k\}$ produced by a generalized pattern search algorithm,*

$$\liminf_{k \to +\infty} ||\nabla f(x_k)|| = 0.$$

Note that $L(x_0) = \{x | f(x) \geq f(x_0)\}$ in the case of a maximization problem.

2.5 Computing the Mixing Proportions

For a finite mixing distribution, $G_{\boldsymbol{\gamma}} = \sum_{k=1}^{m} \pi_k \delta_{\boldsymbol{\gamma}_{(k)}}$. The log-likelihood function (2.1) is written as

$$\begin{aligned} l(\boldsymbol{\pi}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \log L_i(G_{\boldsymbol{\gamma}}) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{m} \pi_k L_i(\boldsymbol{\gamma}_{(k)}), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\gamma}_{(k)}$, $k = 1, \ldots, m$ are the m support points each with mass π_k. For fixed $\boldsymbol{\gamma}$, the algorithm updates $\boldsymbol{\pi}$ in the log-likelihood function (2.4) iteratively at a quadratic order of convergence using the method of Wang (2007) reviewed here.

Let $\boldsymbol{\pi}'$ be an updating vector of $\boldsymbol{\pi}$. A quadratic approximation to $l(\boldsymbol{\pi}, \boldsymbol{\gamma}) - l(\boldsymbol{\pi}', \boldsymbol{\gamma})$ using a Taylor's expansion about $\boldsymbol{\pi}$, is

$$\begin{aligned} Q(\boldsymbol{\pi}'|\boldsymbol{\pi}, \boldsymbol{\gamma}) &\equiv -\mathbf{1}^T S(\boldsymbol{\pi}' - \boldsymbol{\pi}) + \frac{1}{2}(\boldsymbol{\pi}' - \boldsymbol{\pi})^T S^T S(\boldsymbol{\pi}' - \boldsymbol{\pi}) \\ &= \frac{1}{2}||S\boldsymbol{\pi}' - 2\mathbf{1}||^2 - \frac{n}{2}, \end{aligned} \tag{2.5}$$

where $\mathbf{1} = (1, ..., 1)^T$, $||.||$ is the L_2-norm and where the i^{th} row of S is $\nabla \log L_i(G_{\boldsymbol{\gamma}})$ with respect to $\boldsymbol{\pi}$ so that $\nabla l(\boldsymbol{\pi}, \boldsymbol{\gamma}) = S^T \mathbf{1}$. Maximizing the log-likelihood with respect to $\boldsymbol{\pi}'$ in the neighbourhood of $\boldsymbol{\pi}$ is summarized by the following constrained optimization problem:

$$\begin{aligned} &\min_{\pi'} ||S\boldsymbol{\pi}' - 2\mathbf{1}||^2 \\ &\text{subject to } \sum_{k=1}^{m} \pi'_k = 1, \text{ and } \pi'_k \geq 0 \text{ for } k = 1, \ldots, m \end{aligned} \tag{2.6}$$

To solve the optimization problem in (2.6), Wang used the theoretical results

from Haskell and Hanson (1981) which handle small values of $\boldsymbol{\pi}'$ in the solution yielding sufficient numerical stability. Haskell and Hanson state that the following least squares problem with non-negative constraint,

$$\min_{\boldsymbol{\pi}'} |\mathbf{1}^T\boldsymbol{\pi}' - 1|^2 + \psi||S\boldsymbol{\pi}' - 2\mathbf{1}||^2,$$
$$\text{subject to } \pi'_k \geq 0 \text{ for } k = 1, \ldots, m$$

has a solution that converges to the solution of (2.6), as $\psi \to 0+$. Wang used this result because it allows the use of an NNLS (Non-Negative Least-Squares) optimization algorithm without the need to transform the equality constraint in (2.6) to an inequality constraint through variable elimination. This avoids small rounding errors that are truncated to zero which can result in failure of convergence of his NPMLE algorithm.

To ensure an adequate monotonic increase of the log-likelihood function, Wang used a backtracking line search method given below.

- Let $0 < \alpha < 0.5$, $0 < \lambda < 1$ and t=0
- While $l(\boldsymbol{\pi} + \lambda^t\mathbf{d}_t, \boldsymbol{\gamma}) < l(\boldsymbol{\pi}, \boldsymbol{\gamma}) + \alpha\lambda^t\nabla l(\boldsymbol{\pi}, \boldsymbol{\gamma})^T\mathbf{d}_t$
  do $\quad t = t + 1.$

The resulting $\boldsymbol{\pi}' = \boldsymbol{\pi} + \lambda^t\mathbf{d}_t$ where $\mathbf{d}_t$ is a search direction is the new $\boldsymbol{\pi}$. In the case of Newton-type methods for minimization problems $\mathbf{d}_t = -B_t\nabla l(\boldsymbol{\pi}', \boldsymbol{\gamma})$ where B is a symmetric and non-singular matrix, for example the Hessian matrix. The choice of α is governed by the Armijo (or sufficient decrease) rule. It is usually set to be between 0 and 1 but for the line search $1/2 \leq \alpha \leq 1$ is excluded because such values are quite large and risk missing the optimal solution especially when the log-likelihood is well

approximated quadratically. The step halving strategy used in Wang's computations sets $\lambda = 0.5$.

Wang referred to the above algorithm for computing $\boldsymbol{\pi}'$ together with the backtracking line search method as the constrained Newton (CN) method. He then improved this algorithm by adding many support points and discarding redundant ones (i.e. support points with point mass 0) at each iteration. He referred to the resulting modified algorithm as the CNM method where the M stands for multiple support points added at each iteration.

Chapter 3

Binary and Poisson models: Simulations and a case study

3.1 Bernoulli model

Clustered binary data arise frequently in clinical studies where repeated measurements are gathered on experimental units with binary outcomes, for example, passing or failing a test, the presence or absence of a certain anomaly, etc. In family studies, where data from all family members about the presence of a certain attribute is observed, each family would constitute a natural cluster. One approach to modelling such data is to include random effects for the clusters in the linear predictor to account for correlation among observations within clusters (Stiratelli et al, 1984; Anderson and Aitken, 1985). We use the standard logistic model in which the j^{th} observed response of cluster i, y_{ij}, has a Bernoulli(p_{ij}) distribution. Hence, (1.2) becomes

$$\text{logit}\,(\mu_{ij}) = \mathbf{z}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ij}^T\boldsymbol{\gamma}_i \tag{3.1}$$

where $\mu_{ij} = p_{ij}$, $\boldsymbol{\gamma}_i$ is a q-dimensional random effect vector pertaining to $\mathbf{x}_{ij}$ and $\boldsymbol{\beta}$ is the fixed coefficient vector. We assume that the random effect vector $\boldsymbol{\gamma}_i$ has a joint distribution $G(\boldsymbol{\gamma})$.

3.1.1 Two sample simulations

We perform simulation studies to evaluate the behaviour of the algorithm. All simulations reported in this section have direct search step size parameters Δ_0 and ξ, both set to 0.5. The parameters max_number_iterations_β, max_number_iterations_G, *precisionR*, *precisionM*, and *tolerance* of Section 2.3.1 are set to 20, 50, 0.0001, 0.05 and 0.001 respectively.

For each sample, we simulate 100 clusters with five binary observations per cluster, resulting in 500 observations per sample. The simulations were also fitted using the "glmer" routine of the lme4 package in R (Bates et al., 2011). The results of both routines are compared in later chapters of the dissertation. Binary data were simulated according to two logistic regression models. The Sample 1 simulation model includes a random intercept, a, and two slopes, b and w, generated from a discrete multivariate mixture with equal mass on the two (a, b, w) points $(-1, -1, -1)$ and $(1, 1, 1)$. The random between cluster covariate, x_b, takes on values -0.5 for the first 50 clusters and 0.5 for the remaining 50 clusters. The random within cluster covariate, x_w, takes on values $-1, -0.5, 0, 0.5, 1$ for the five items within each cluster. Sample 2 simulation includes a random intercept, a, and a slope, b, generated from the discrete bivariate distribution that places equal mass on the (a,b) pairs $(0.5, -1)$ and $(0, 1)$. The random covariate vector, $\mathbf{x}$, is equal to -2 for the first 50 clusters and 2 for the remaining 50 clusters. The linear model for Sample 2 also includes a fixed covariate vector, $\mathbf{z}$, with values $-1, -0.5, 0, 0.5, 1$ for the five items within each cluster

Sample 1		
Generating Mixture	$\boldsymbol{\gamma} = (a, b, w)$ $\boldsymbol{\pi}$	(-1,-1,-1),(1,1,1) 0.5, 0.5
Covariates	$\mathbf{x}_b$ $\mathbf{x}_{w_i}$	-0.5 for the first 50 clusters and 0.5 for the rest $(-1, -0.5, 0, 0.5, 1)$ for a cluster i
Sample 2		
Generating Mixture	$\boldsymbol{\gamma} = (a, b)$ $\boldsymbol{\pi}$	(0.5,-1),(0,1) 0.5, 0.5
Covariates	$\mathbf{x}$ $\mathbf{z}_i$	-2 for the first 50 clusters and 2 for the rest $(-1, -0.5, 0, 0.5, 1)$ for a cluster i
Fixed coefficient	β	1

Table 3.1: Summary of the random effects true distribution, the fixed coefficients and covariates used in generating the binary response for two binary samples

and a corresponding fixed coefficient $\beta = 1$. Table 3.1 summarizes the parameters and covariates used to generate both samples.

Alternatively, the glmer routine assumes that $G(\boldsymbol{\gamma})$ is multivariate normal with mean 0. In order to have a better suited comparison with the DSDD outputs, the linear link functions for samples 1 and 2 become respectively

$$\text{logit}(p_{ij}) = a_i + B_a + (b_i + B_b)x_{b_{ij}} + (w_i + B_w)x_{w_{ij}}, \tag{3.2}$$

$$\text{logit}(p_{ij}) = a_i + B_a + (b_i + B_b)x_{ij} + \beta z_{ij} \tag{3.3}$$

where B_a, β, B_b, and B_w are fixed coefficients. Tables 3.2 and 3.3 display estimation results from the DSDD and glmer routines. The DSDD achieved a maximum gradient at the final iteration of -2.56e-05 and 1.1e-04 for Samples 1 and 2, respectively.

The estimate of the fixed effect coefficient β in Sample 2 is the maximum likelihood estimator. The standard error of β was estimated by drawing 1000 bootstrap samples each one containing 100 clusters of 5 observations. Each bootstrap sample is drawn

Method	Component	$\hat{\gamma}$ a	b	w	$\hat{\pi}$
TRUE	1	-1	-1	-1	0.5
	2	1	1	1	0.5
DSDD	1	-1	-1	-1	0.47
	2	-0.67	-1	0	0.10
	3	-0.18	-2.01	0	0.12
	4	1.03	1.04	0.98	0.31
	mean	-0.23	-0.48	-0.16	
	$\hat{\sigma}$	0.88	1.07	0.86	
glmer		$\hat{B}_a = -0.23(0.14)$	$\hat{B}_b = -0.49(0.27)$	$\hat{B}_w = -0.23(0.17)$	
		$\hat{\sigma}_a = 0.98$	$\hat{\sigma}_b = 0.92$	$\hat{\sigma}_w = 0.88$	

Table 3.2: Binary Sample 1 simulation results: including the DSDD estimated distribution of the random effects with their standard errors and weighted means and the distribution of the results obtained from the glmer routine of the lme4 library. () denotes the standard errors corresponding to the fixed coefficients $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ while $\hat{\sigma}$ denote the estimated standard deviation of the random effect.

Method	Component	$\hat{\gamma}$ a	b	$\hat{\pi}$	$\hat{\beta}$ $(s.e.)$
TRUE	1	0.5	-1	0.5	1
	2	0	1	0.5	
DSDD	1	-0.01	0.99	0.41	1.23(0.23)
	2	0.67	-0.96	0.44	
	3	0.77	-0.99	0.13	
	4	3	-2.13	0.02	
	mean	0.48	-0.21		
	$\hat{\sigma}$	0.5	1		
glmer		$\hat{B}_a = 0.59(0.30)$	$\hat{B}_b = -0.29(0.14)$		1.15(0.18)
		$\hat{\sigma}_a = 1.38$	$\hat{\sigma}_b = 1.13$		

Table 3.3: Binary Sample 2 simulation results: including the DSDD estimated distribution of the random effects with their standard errors and weighted means and the results obtained from the glmer routine of the lme4 library. () and denote the estimated standard errors of the fixed effect, while $\hat{\sigma}$ denotes the estimated standard deviation of the random coefficients.

using the following 2-level bootstrapping technique;

1. Draw $\mathbf{y}_{ij}^{boot} \sim \text{Bernoulli}(p_{ij}^{boot})$ by calculating the bootstrapped probability for the j^{th} observation in the i^{th} cluster according to

$$\text{logit}(p_{ij}^{boot}) = \ddot{a} + \ddot{b}x_{ij} + \hat{\beta}z_{ij} \tag{3.4}$$

   where $(\ddot{a}_i, \ddot{b}_i)$ are posterior modes of the random effects (see Section 5.1.2 for details) and $\hat{\beta} = \hat{\beta}(\mathbf{y})$ is the MLE of β given the original response $\mathbf{y}$.

2. The vector $\mathbf{y}^{boot}$ is refitted to obtain the bootstrap estimate, $\hat{\beta}^{boot}$, of the fixed coefficient.

As a result, the estimated standard error reported in Table 3.3 is obtained by calculating the standard error of the 1000 $\hat{\beta}^{boot}$ estimates. The method is similar to the semi-parametric 2-level technique described in Chambers et al. (2011).

The first plot of Figure 3.1 shows the resulting surface of the gradient for Sample 1 computed from $L_{\hat{G}}$ to the point mass distribution of the random slopes where the random slope w is fixed at -1. The second plot in the figure shows a three dimensional gradient surface computed from $L_{\hat{G}}$, where $\hat{G}$ is the MLE of G, toward the point (a, b). In the first plot the gradient surface reaches 0 at a point close to $a = -1$ and $b = -1$ while the modes in the second plot coincide with the resulting random effect mass points shown in Table 3.3.

Figures 3.2 and 3.3 show the cluster averages of the estimated probabilities using the mixed effect model from the DSDD and glmer algorithms compared to the true average probabilities by cluster for the two samples. The prediction of the probabilities are done using the technique described in Section 5.1.2. The graphs demonstrate the ability of the DSDD algorithm to successfully estimate the heterogeneity between

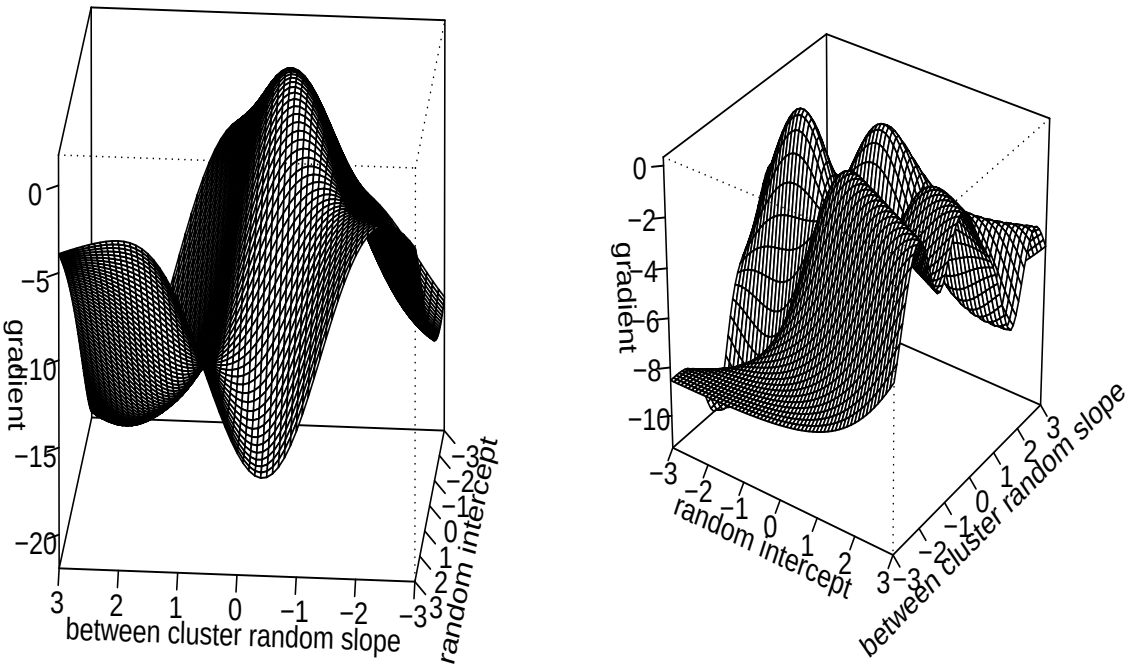

Figure 3.1: Gradient graphs of mixtures resulting from two simulated samples of binary responses. Plot 1 shows a three dimensional gradient surface of the first simulated dataset, $D(\boldsymbol{\gamma} = (a, b, -1), \hat{G}\boldsymbol{\gamma})$, from the MLE $\hat{G}$ to the points $\boldsymbol{\gamma} = (a, b, -1)$. Similarly, plot 2 shows the resulting three dimensional gradient plot of the second simulated data from the MLE of the mixing distribution to the points $\boldsymbol{\gamma} = (a, b)$.

clusters.

Figure 3.4 compares the distribution of the estimated DSDD and glmer mixtures to the true mixture. The displayed frequencies corresponding to DSDD estimates in the histogram are obtained by multiplying the corresponding probability of each mass point by 100, the total number of clusters in the sample. The histograms also display the distributions of the glmer posterior modes calculated for each cluster by minimizing the penalized residuals sum of squares function (PRSS) of the posterior probability with respect to the random effects. This is done using a Cholesky decomposition of sparse positive semi-definite matrices representing the conditional model given the random effects (Bates et al., 2011, see Section 5.1.2 for more details). For each random effect, the glmer routine returns one mode per cluster resulting in 100 modes in total for our samples. Figure 3.4 plots the histogram of the 100 modes for each random effect. To more adequately compare the distributions of the random effects fitted by the two models, the constants $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ of (3.2) are added to the modes of **a**, **b** and **w** in Sample 1, respectively. Similarly, the estimates $\hat{B}_a$ and $\hat{B}_b$ of (3.3) are added to the modes of the random intercept and slope in Sample 2.

3.1.2 1000 dataset binary data simulations

The two sample simulations previously described were each repeated 1000 times to assess the performance of the DSDD. The algorithm parameters were unchanged with the exception of *tolerance*, the stopping criterion for maximizing over the mixing distribution, which was changed to 0.01 from 0.001 in Sample 2. This considerably speeds the convergence without a significant loss in estimation accuracy.

Figure 3.5 shows a summary of the results from simulating 1000 datasets according to Sample 1 specifications. The first three plots show the distributions of the

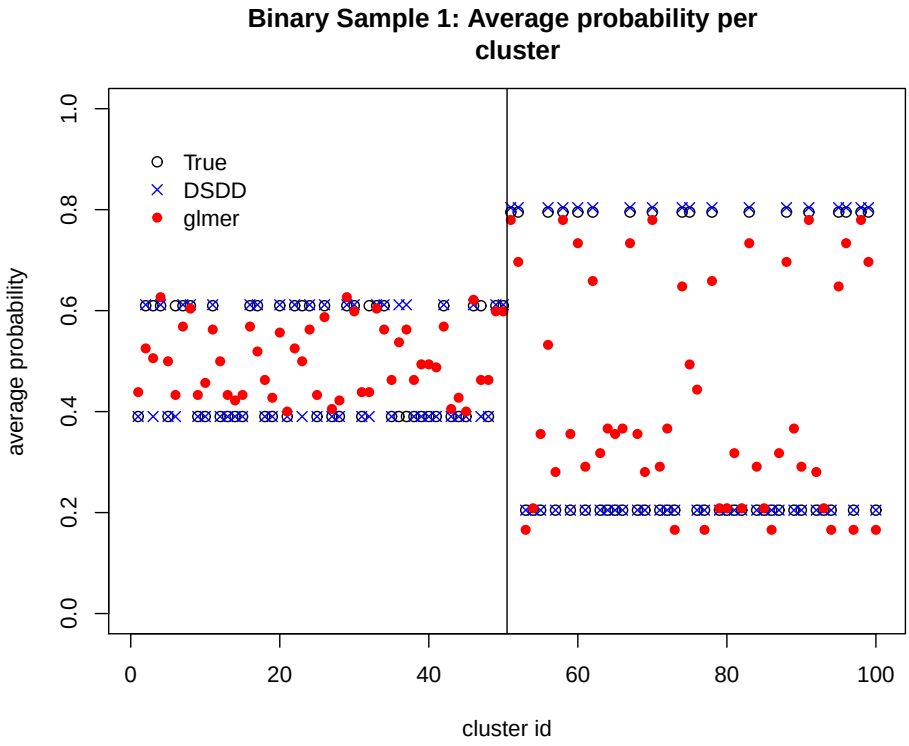

Figure 3.2: Binary Sample 1 simulation: average probabilities of success per cluster estimated by the DSDD and glmer routines compared to the true probabilities.

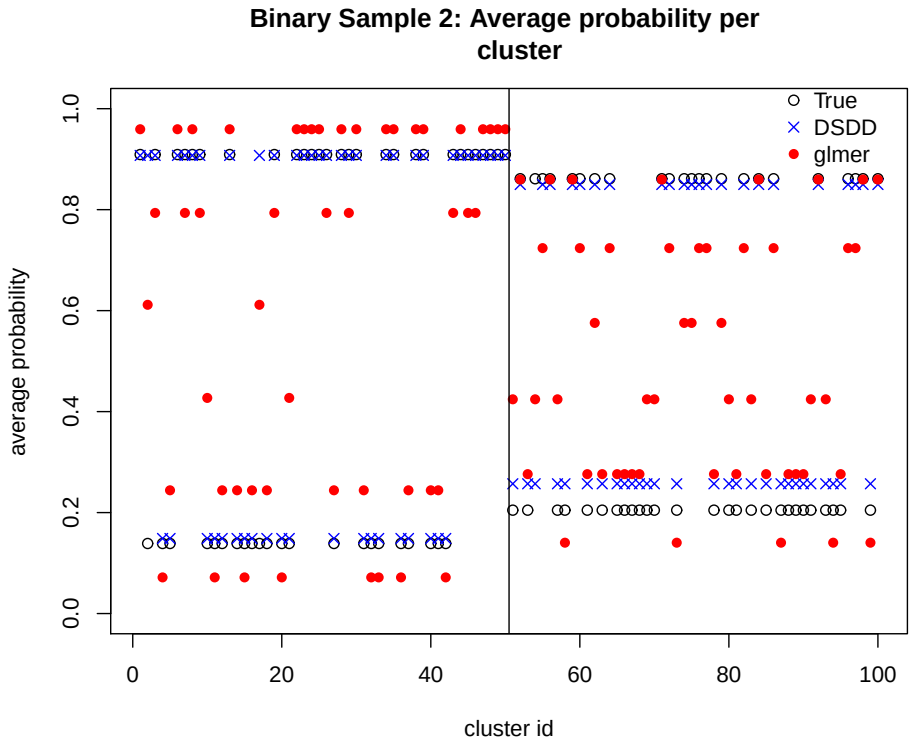

Figure 3.3: Binary Sample 2 simulation: average probabilities of success per cluster estimated by the DSDD and glmer routines compared to the true probabilities

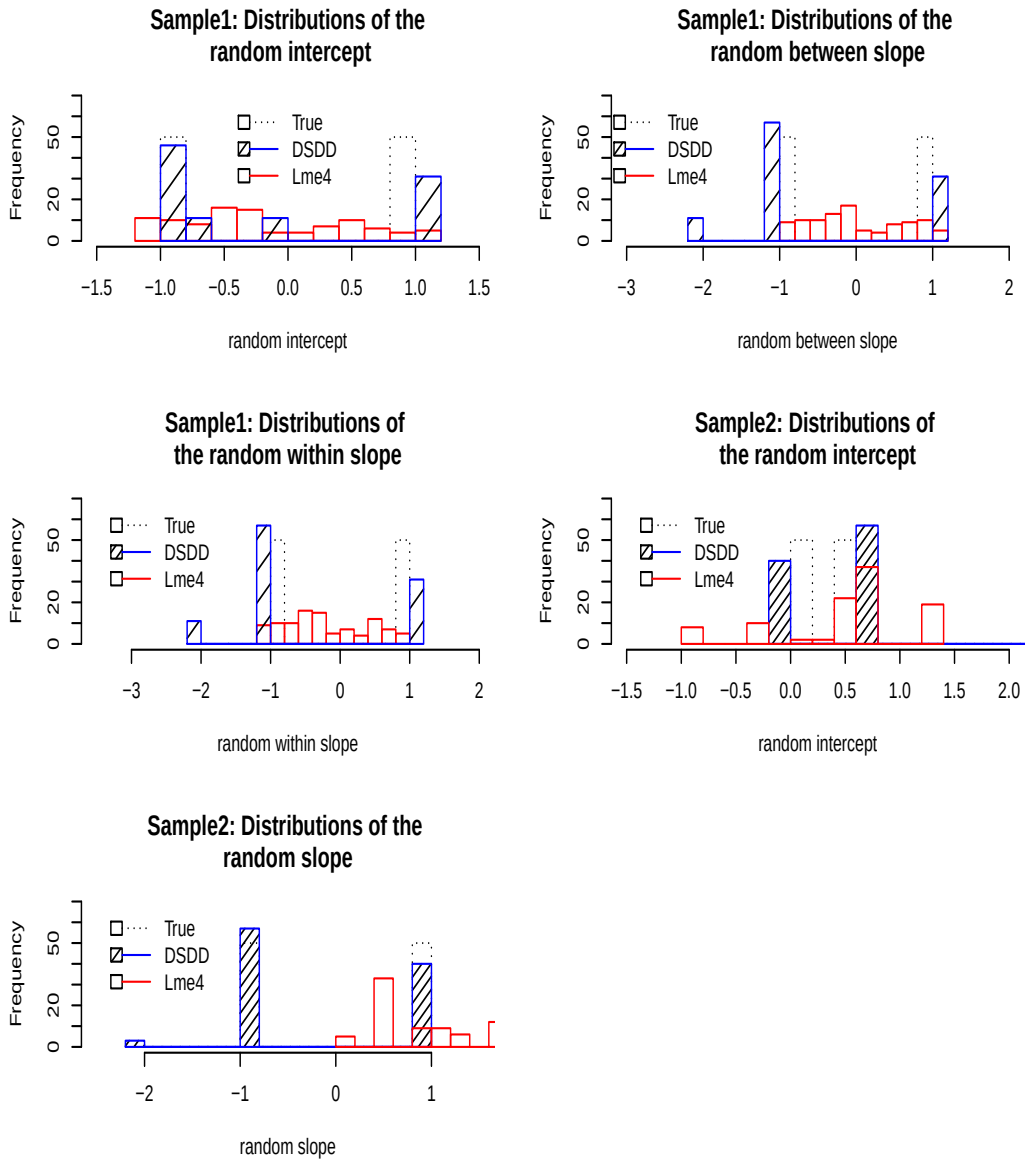

Figure 3.4: Plots comparing the mass point distributions of the binary simulations 1 and 2 obtained by the DSDD algorithm and the posterior modes generated by the glmer routine of the package lme4 in R. We add $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ estimated by the glmer routine to the modes of the random intercept, random between slope and within slope (in Sample 1 only), respectively, because the posterior distribution of the random effects fitted by the glmer routine are forced to have zero means.

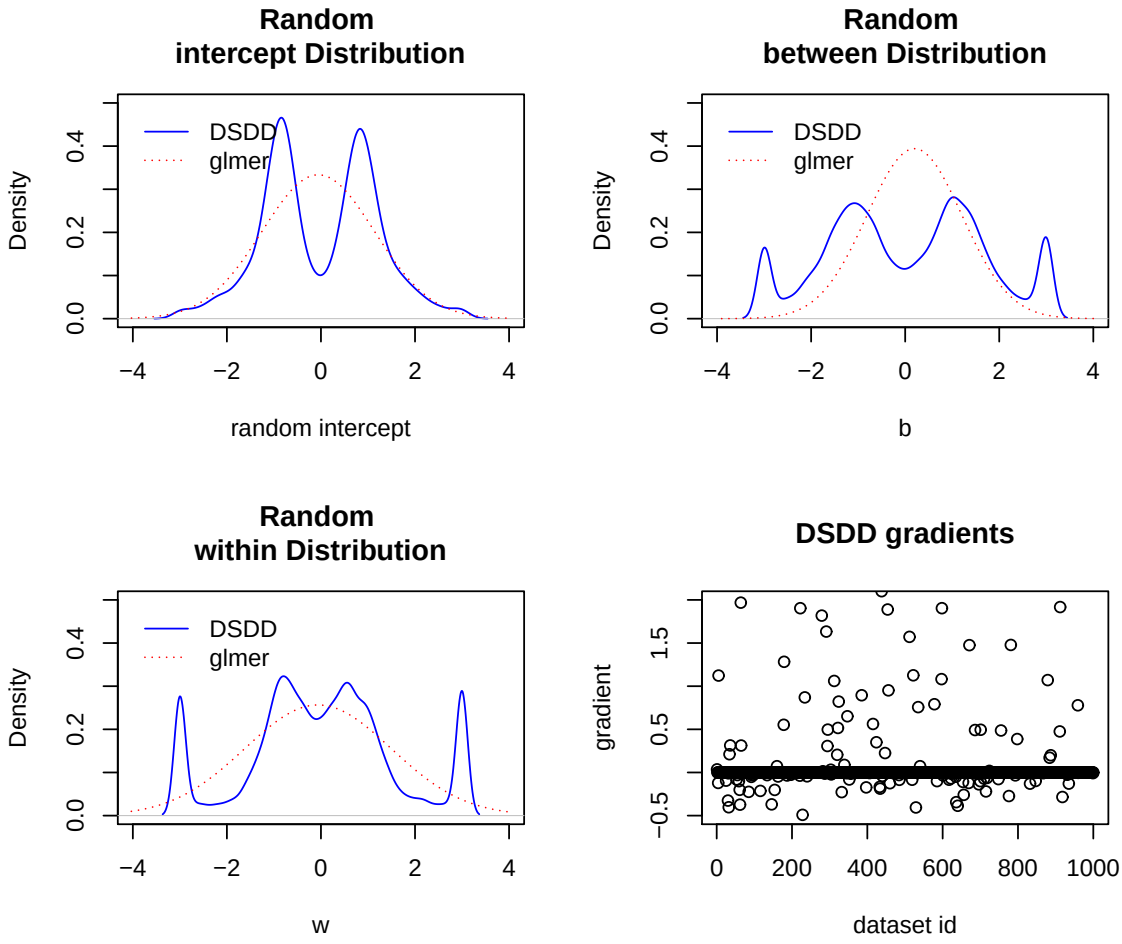

Figure 3.5: Summary of the results over 1000 binary Sample 1 datasets displaying the random effects distributions resulting from DSDD and glmer algorithms, and the maximum gradient achieved per dataset for the DSDD method.

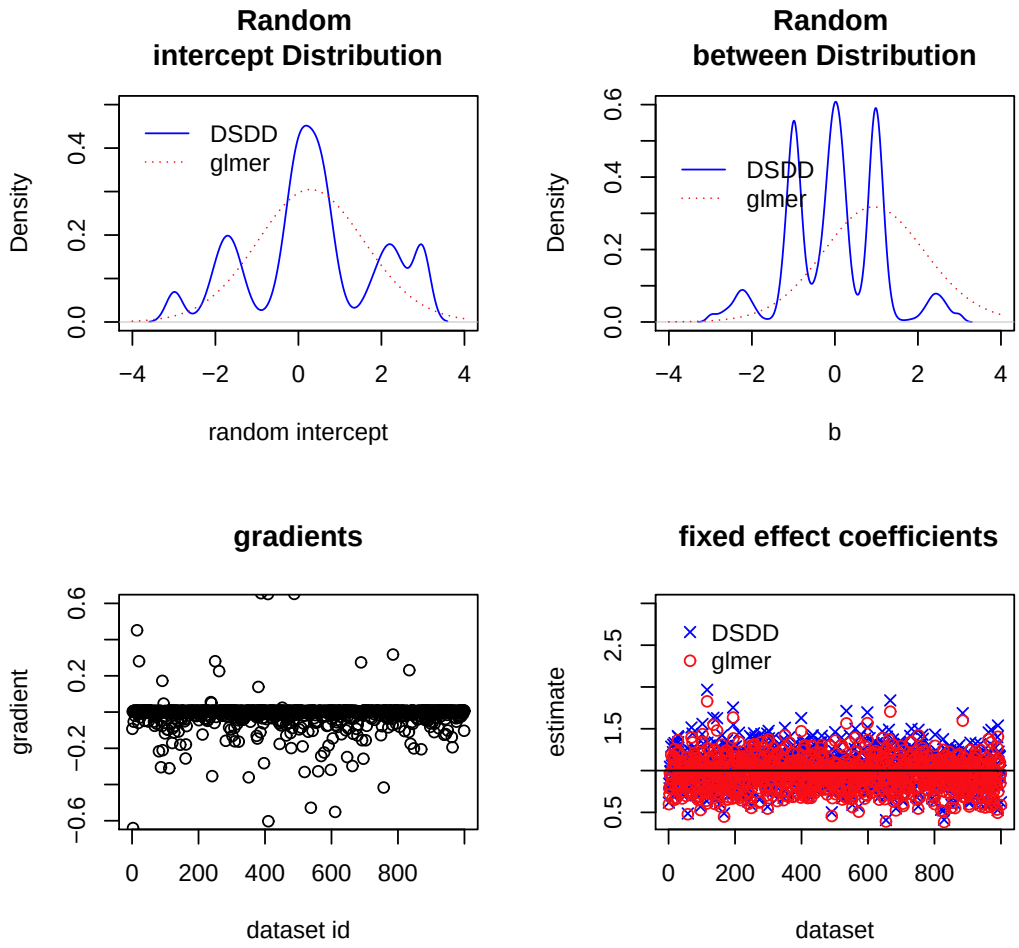

Figure 3.6: Summary of the results over 1000 binary Sample 2 dataset displaying the distributions of random intercept and slope estimates obtained by the glmer and DSDD algorithms, the maximum gradient achieved per dataset for DSDD, and fixed coefficients estimates of the glmer and DSDD algorithms.

Sample	Method	average(log-lik.(ll))	$se(ll)$	average($\hat{\beta}$)
1	DSDD	-310.76	8.08	na
	lme4	-318.66	7.24	na
2	DSDD	-247.25	12.82	1.03
	lme4	-258.16	11.88	0.96

Table 3.4: Summary of the 1000 dataset binary simulation from Sample 1 and 2 including the log-likelihoods, ll, with their standard errors (se) and the estimated fixed coefficient, $\hat{\beta}$.

estimated random effects of the DSDD and glmer methods. The distribution of the random effects produced by the DSDD algorithm was plotted by aggregating the resulting 1000 random effect vectors and plotting their density while accounting for the respective weight of each mass point. The aggregated glmer mixture distributions were obtained by the following method: for a one dimensional grid of equally spaced points centred around the average of the fixed covariate coefficient, we obtain 1000 probability density vectors (one from each estimated mixing distribution for 1000 datasets), then we average the densities for every point of the vector and plot the results across the grid. The fourth plot in Figure 3.5 shows the maximum gradient achieved at the end of each simulation for the Sample 1 set. The results were generally around zero and less than the tolerance level of 0.001. A few cases failed to converge, most likely due to their relatively flat gradient surfaces.

Results for Sample 2 simulations are displayed in Figure 3.6. The maximum gradient achieved at the end of each simulation was close to zero and less than the tolerance level of 0.01. Figure 3.6 shows that the estimates of the fixed coefficient were generally consistent between the DSDD and glmer results. The distributions of the random effects resulting from the DSDD and glmer routines are also shown. Summaries of the estimated log-likelihoods, ll, and fixed coefficients, $\hat{\beta}$, are displayed in Table 3.4.

Distributions of the random effects shown in Figures 3.5 and 3.6 are expectedly multi-modal for the nonparametric mixture estimated by the DSDD and uni-modal for the glmer estimated mixture. These results highlight the advantage of using a nonparametric mixture when making inference about the random effects in situations where the mixing distribution deviate from the normal assumption.

3.2 Poisson model

Count data are usually viewed as arising from a Poisson distribution. Population heterogeneity and unaccounted variations necessitate the inclusion of random effects in Poisson models. In a clustered data setting, let y_{ij} be the j^{th} observation of the i^{th} cluster in a random sample of count data. The mean, λ_{ij}, can be linearly modelled using the log-link:

$$log(\lambda_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\gamma}_i + \mathbf{z}_{ij}^T \boldsymbol{\beta} \tag{3.5}$$

where $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}$ are the random effect and fixed coefficient vectors respectively, while $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the random and fixed effects covariate vectors respectively. The probability mass function of the response vector of cluster i, $\mathbf{y}_i$ becomes

$$f(\mathbf{y}_i, G) = \int_{\Omega} \prod_{j=1}^{n_i} \frac{\mathrm{e}^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!} dG(\boldsymbol{\gamma}) \tag{3.6}$$

where $G(\boldsymbol{\gamma})$ is the mixing distribution of $\boldsymbol{\gamma}$.

Simar (1976) proposed using nonparametric estimation of G for Poisson mixtures when there is a lack of prior information about G. He also presented an algorithm to estimate the mixing distribution by maximizing the log-likelihood of the count dataset while proving that the NPMLE of G is unique and strongly consistent. Simar's

results were later extended to mixtures of other families of distributions (Jewell, 1982; Lindsay, 1983).

3.2.1 Two sample simulations

We performed simulation studies to assess the behaviour of the DSDD algorithm when dealing with count data. The algorithm's parameters are kept the same as in Section 3.1.1. Firstly we present two simulated samples according to a Poisson generalized linear model following (3.5). Sample 1 includes a random intercept, a, and two random slopes, b and w, generated from the discrete distribution that places equal probabilities on the mass points $(a, b, w) = \{(-1, -1, -1), (1, 1, 1)\}$. The random between cluster covariate, x_b, equals -1 for the first 50 clusters and 1 for the rest while the random within cluster covariate, x_w, takes on values $-0.5, -0.25, 0, 0.25, 0.5$ for the five observations within each cluster. Sample 2 was simulated according to a GLMM Poisson model including a random intercept and slope generated from a discrete distribution that places equal probabilities on the two mass points, $(intercept = a, slope = b) = \{(-1, -1), (1, 1)\}$. The random covariate x takes a value of -0.5 for the first 50 clusters and 0.5 for the remaining. While the fixed covariate z takes on values $-2, -1, 0, 1, 2$ for the five items within each cluster with a corresponding coefficient $\beta = 0.5$. Table 3.5 summarizes the parameters and covariates used to generate both samples.

The glmer routine in R fits Poisson GLMM models, where $G(\boldsymbol{\gamma})$ is assumed to be multivariate normal with mean 0 (Bates et al., 2011). Similarly to the Bernoulli two sample simulations, we fit the Poisson models using glmer with the following two

Sample 1		
Generating Mixture	$\boldsymbol{\gamma} = (a, b, w)$	(-1,-1,-1),(1,1,1)
	$\boldsymbol{\pi}$	0.5, 0.5
Covariates	$\mathbf{x}_b$	-1 for the first 50 clusters and 1 for the rest
	$\mathbf{x}_{w_i}$	$(-0.5, -0.25, 0, 0.25, 0.5)$ for a cluster i
Sample 2		
Generating Mixture	$\boldsymbol{\gamma} = (a, b)$	(-1,-1),(1,1)
	$\boldsymbol{\pi}$	0.5, 0.5
Covariates	$\mathbf{x}$	-0.5 for the first 50 clusters and 0.5 for the rest
	$\mathbf{z}_i$	$(-2, -1, 0, 1, 2)$ for a cluster i
Fixed coefficient	β	0.5

Table 3.5: Summary of the random effects true distribution, the fixed coefficients and covariates used in generating the Poisson response for two simulated count data.

		$\hat{\boldsymbol{\gamma}}$			
Method	Component	a	b	w	$\hat{\pi}$
TRUE	1	-1	-1	-1	0.5
	2	1	1	1	0.5
DSDD	1	-0.87	-1.02	-1.27	0.052
	2	-0.64	-0.78	-0.28	0.49
	3	0.94	1.08	0.94	0.46
	mean	0.075	0.065	0.23	
	$\hat{\sigma}$	0.80	0.94	0.69	
glmer		$\hat{B}_a = 0.012(0.14)$	$\hat{B}_b = 0.011(0.14)$	$\hat{B}_w = 0.24(0.11)$	
		$\hat{\sigma}_a = 0.92$	$\hat{\sigma}_b = 1.05$	$\hat{\sigma}_w = 0.68$	

Table 3.6: Poisson Sample 1 simulation results including the DSDD estimated distribution of the random effects with their standard deviations and weighted means. We also display the distribution of the random effects obtained by fitting the model using the glmer routine where $\hat{\sigma}$ denotes the estimated standard deviation of the random effects. () denotes the standard errors of the fixed coefficients.

Method	Component	$\hat{\gamma}$ a	b	$\hat{\pi}$	$\hat{\beta}$ $(s.e.)$
TRUE	1	-1	-1	0.5	0.5
	2	1	1	0.5	
DSDD	1	-1.06	-0.80	0.45	0.49(0.04)
	2	-0.56	-1.81	0.034	
	3	0.56	2.46	0.017	
	4	0.93	1.17	0.50	
	mean	-0.017	0.22		
	$\hat{\sigma}$	0.98	1.07		
glmer		$\hat{B}_a = -0.005(0.13)$	$\hat{B}_b = 0.19(0.25)$		0.49(0.025)
		$\hat{\sigma}_a = 0.99$	$\hat{\sigma}_b = 1.25$		

Table 3.7: Poisson Sample 2 simulation results including the DSDD estimated distribution of the random effects with their standard errors and weighted means and glmer estimates. () denotes the standard errors of the fixed coefficients.

linear link functions for Samples 1 and 2 respectively

$$\log(\lambda_{ij}) = a_i + B_a + (b_i + B_b)x_{b_{ij}} + (w_i + B_w)x_{w_{ij}}, \tag{3.7}$$

$$\log(\lambda_{ij}) = B_a + a_i + (b_i + B_b)x_{b_{ij}} + \beta z_{ij} \tag{3.8}$$

where B_a, β, B_b, and B_w are fixed coefficients. Estimated results from the glmer and DSDD algorithms are displayed in Tables 3.6 and 3.7. The DSDD algorithm achieved convergence at gradients -0.001 and 7.1e-05 for Samples 1 and 2, respectively.

Figure 3.7 compares the posterior mode distributions of the random effects obtained by fitting both samples using the glmer routine with the random effect distributions obtained from the DSDD algorithm. The displayed frequencies corresponding to DSDD estimates in the histogram are obtained by multiplying the corresponding probability of each mass point by 100, the total number of clusters in the sample. The histograms also display the distributions of the glmer posterior modes calculated

for each cluster. The modes are obtained by minimizing the penalized residuals sum of squares function (PRSS) of the posterior probability with respect to the random effects through a Cholesky decomposition of sparse positive semi-definite matrices representing the conditional model given the random effects (Bates et al., 2011. See Section 5.1.2 for more details).

For each random effect the glmer routine returns one mode per cluster resulting in 100 modes in total for our samples. Figure 3.7 plots the histogram of the 100 modes for each random effect. The constants $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ estimated in (3.7) are added to the modes of **a**, **b** and **w** in the Sample 1, respectively. Similarly, the estimates $\hat{B}_a$ and $\hat{B}_b$ in (3.8) to the modes of the random intercept and slope of Sample 2.

3.2.2 1000 Sample Simulations

We simulate 1000 datasets following the specifications of Sample 2 in Section 3.2.1. Figure 3.8 summarizes the densities of the random intercept and slope resulting from the DSDD algorithm and the glmer routine. The DSDD density curves were obtained by aggregating all the mass points resulting from the 1000 runs according to their respective probabilities. The glmer density curve is obtained by aggregating the posterior densities.

As in the Bernoulli 1000 simulations, the distribution of the random effects produced by the DSDD algorithm was plotted by aggregating the resulting 1000 random effect vectors and plotting their density while accounting for the respective weight of each mass point. The glmer mixture distributions were aggregated by obtaining 1000 probability density vectors (one from each estimated mixing distribution for 1000 datasets) for points equally spaced on a one dimensional grid centred around the average of the fixed coefficients of the random covariate. The densities were averaged

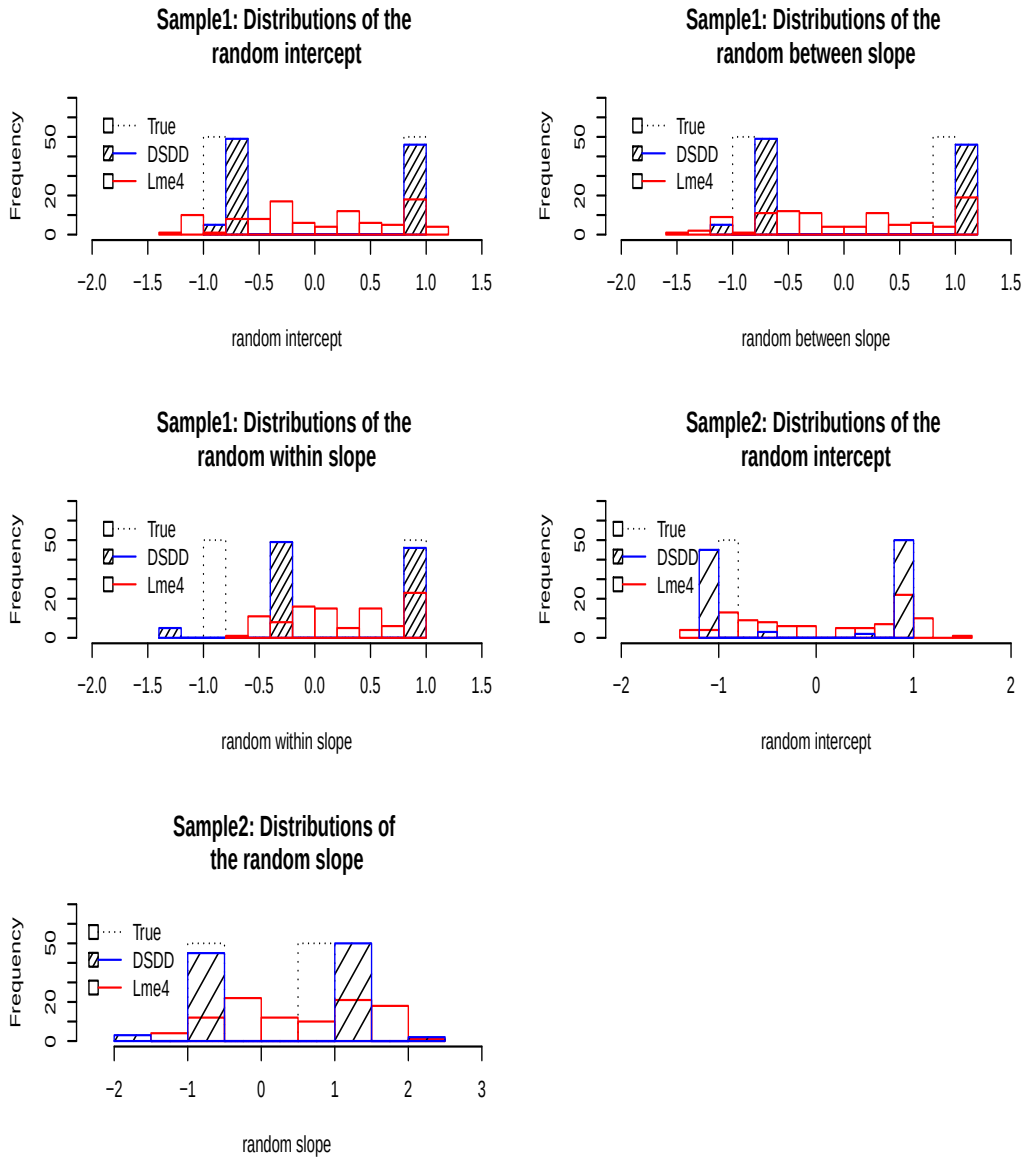

Figure 3.7: 5 plots comparing the mass point distributions of the Poisson sample simulations 1 and 2 obtained by the DSDD algorithm and the posterior modes generated by the glmer routine of the package lme4 in R. We add $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ estimated by the glmer routine to the modes of the random intercept, random between slope and random within slope (in the Sample 1 only), respectively, to accommodate the fact that the posterior distribution of the random effects fitted by the glmer routine have means zero.

for every point and the results are plotted across the grid. The first plot in Figure 3.8 shows that the DSDD estimated distribution is bimodal at a =-1 and 1. The second plot shows that the DSDD random slope density have four visible peaks with the two highest peaks at the true random slope mass points, $b = -1$ and $b = 1$. The glmer estimated distributions of the random intercept and slope are centred around the average of the estimates of $\hat{B}_a$ and $\hat{B}_b$, respectively. The gradient plot in Figure 3.8 shows that most samples achieved gradients around or below the tolerance level of 0.01. Lastly, the estimates for the fixed coefficients were close to the true value of 0.5 in most cases with means 0.504 and 0.499 and sample standard deviations of 0.069 and 0.084 for the DSDD and glmer algorithms, respectively.

Similarly to Figures 3.5 and 3.6, the multi-modal DSDD estimated random effect distributions in Figure 3.8 highlights the advantage of using nonparametric mixture distribution instead of the uni-modal normal distribution when making random effects inferences. Comparisons of the DSDD and glmer estimates are performed in Chapter 5.

3.3 National Basketball Association (NBA) case study

Eight of the top ten spending NBA (National Basketball Association) teams qualified to the postseason in the 2009-2010 season so that one might conjecture that spending is associated with a successful NBA franchise. The hypothesis is difficult to prove given the numerous factors that come into play during a season, for example, injuries, coaching staff, etc. Nevertheless we investigate this association through the analysis of data from the 1994-1995 to 2009-2010 seasons.

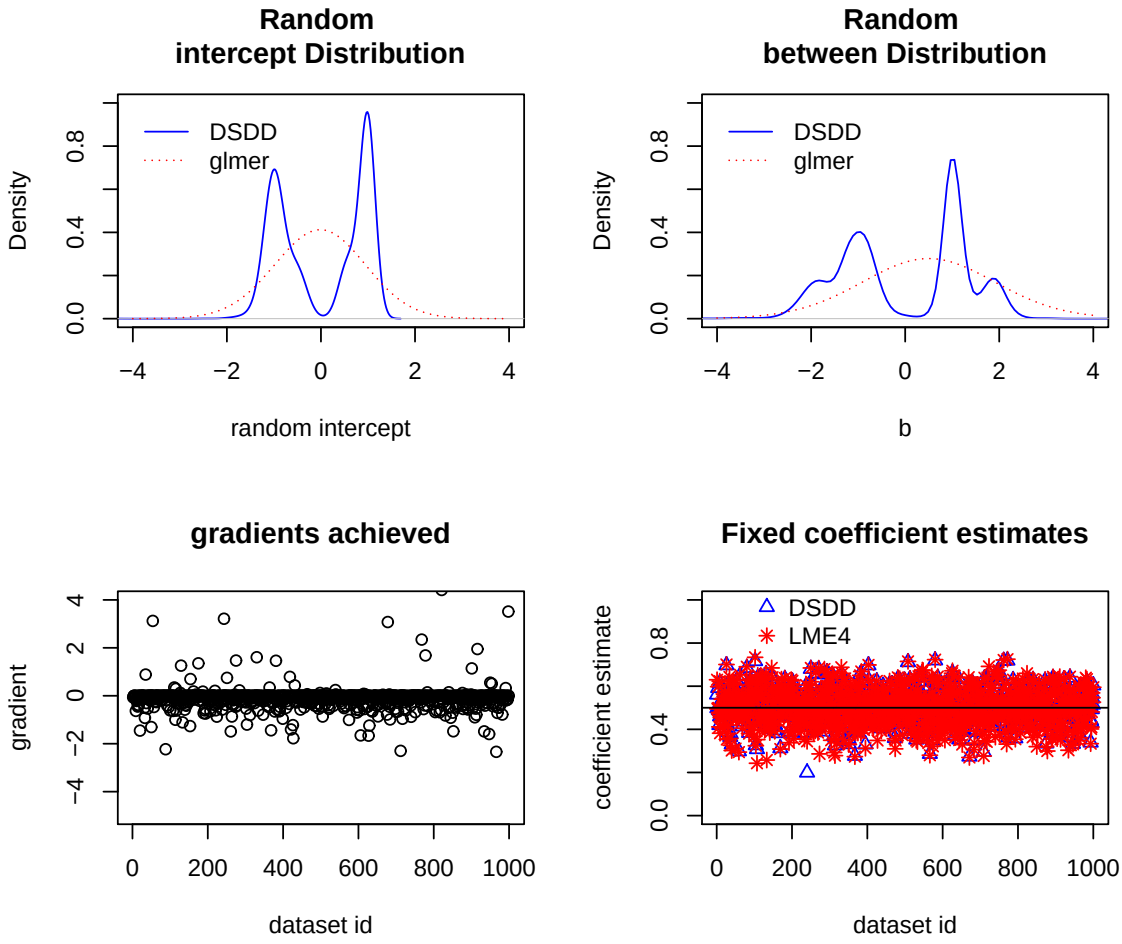

Figure 3.8: Summary plots for 1000 Poisson datasets simulated according to the specifications of Sample 2 in Section 3.2.1. The plots compare the distributions of the random intercept and slope obtained by the glmer and DSDD algorithms

3.3.1 Data description

Data were collected for 30 NBA teams over 16 seasons and consisted of their payroll, the mean age of the roster and whether the team made the playoffs (Bender, 2011). The response variable is a binary outcome $y_{ij} = 1$ denoting success of team i in making the playoffs during year j and $y_{ij} = 0$ denoting otherwise. The payroll variable is a standardized score within each season, calculated by subtracting the mean payroll of all NBA teams in a particular season and divided by their standard deviation. It is worth noting that certain NBA franchises are younger than 16 years and hence have shorter cluster sizes (the Vancouver/Memphis Grizzlies, the Charlotte Bobcats and the Toronto Raptors) and some franchises relocated during this period (Vancouver Grizzlies relocated to Memphis, the Hornets relocated from Charlotte to New Orleans and the Seattle Sonics moved to Oklahoma City) but were counted as the same franchise. The relative experience of a roster could also affect its winning total, hence the mean age of the team was standardized and added to the model as a fixed effect covariate. Refer to Table A.1 in the Appendix for a summary of the data.

3.3.2 Analysis

In the analysed model we treat the covariate Pay as a random effect due to the varying effect it could have in different markets. For example teams in relatively small markets might have to overpay to attract players, conversely productive players might choose big markets despite earning lower salaries. On the other hand, the age covariate is treated as a fixed effect. The suggested model is the following:

$$\text{logit}(p_{ij}) = a_i + b_i Pay_{ij} + age_{ij}\beta_{age} \tag{3.9}$$

Method	Component	$\hat{\gamma}$ a	b	$\hat{\pi}$	$\hat{\beta}_{age}$ $(s.e.)$
DSDD	1	-1.35	3	0.04	0.72(8.29)
	2	-0.12	-0.17	0.24	
	3	0.49	0.31	0.39	
	4	0.54	3	0.19	
	5	1.01	-0.19	0.05	
	6	2.07	2.19	0.09	
	mean	0.44	0.96		
	$\hat{\sigma}$	0.67	1.28		
glmer		$\hat{\beta}_0 = 0.42(0.15)$	$\hat{\beta}_{pay} = 0.72(0.22)$		0.74(0.42)
		$\hat{\sigma}_a = 0.50$	$\hat{\sigma}_b = 0.80$		
GLM		$\hat{\beta}_0 = 0.25(0.10)$	$\hat{\beta}_{pay} = 0.37(0.13)$		0.73(0.12)

Table 3.8: DSDD and glmer results for the NBA dataset.

where a and b are the random effects and β_{age} is the fixed coefficient pertaining to age. Alternatively, the logistic model fitted by glmer takes the form

$$\text{logit}(p_{ij}) = \beta_0 + a_i + \beta_{age} age_{ij} + (\beta_{pay} + b_i) Pay_{ij} \tag{3.10}$$

where a and b are random effects assumed to have a joint multivariate normal distribution with mean 0 while β_0 and β_{pay} are the fixed effects.

In addition to the aforementioned mixed models, we consider a standard generalized linear model (GLM) where the covariate Pay is treated as a fixed effect. The model takes the form

$$\text{logit}(p_{ij}) = \beta_0 + \beta_{age} age_{ij} + \beta_{pay} Pay_{ij} \tag{3.11}$$

The DSDD algorithm was set to stop when the maximum gradient among all support points in the estimated mixture was less than $tolerance = 0.001$ or the change in the log-likelihood was at most 0.001. Table 3.8 summarizes the results from

models 3.9, 3.10 and 3.11. The DSDD algorithm converged quickly with 25 iterations and the resulting estimated mixture produced a maximum gradient of 0.0005. The algorithm also yielded an estimate of 0.72 for the fixed covariate coefficient β_{age}, which is consistent with the glmer and GLM models. A comparison of the three model fits is presented in Section 6.2.1.

The values of the estimated random slopes $\hat{b}$ in the mixed model given in Table A.1 show that for most teams there exist a positive correlation between the probability of a playoff appearance and payroll, which agrees with the common perception that success costs more money. One glaring exception is a team like New York which had a poor playoffs proportion of 0.5 despite a relatively expensive payroll over the span of 16 years; this can be explained by the career threatening injuries to their top earning players. Other notable exceptions are teams that made the playoffs in multiple years despite their relatively modest payrolls (such as Toronto, New Orleans and New Jersey), which could be explained by the accumulation of talented young players on their rosters who command lower salaries than their veteran peers as stipulated in the NBA salary scale.

The model analysis presented above is oversimplified because it ignores many intangible factors that govern the success of an NBA franchise. However, including payroll as a random effect in the model allows for a flexible investigation of the relationship between spending and winning for each team without aggregating over the whole league.

Chapter 4

Zero-inflated data

Count data with excess zeros arise in many applications. For example, it is very common to observe a spike at zero in a study investigating the number of bear attacks on humans in a certain region. Analyzing such data by assuming a Poisson distribution ignores the zero-inflation and can give rise to overdispersion. The zero-inflated Poisson model (Lambert, 1992) is one way to allow for overdispersion, it assumes that the sample comes from a distribution that is a mixture of a count model and one degenerate at zero, in this case the zeros in the model come from the zero state and from the ordinary count distribution. The hurdle model (Arulampalam and Booth, 1997; Mullahy, 1986) is an alternative way to deal with zero-inflated data. It considers the zeros to be completely separate from the non-zeros, hence it consists of two-stages where the first stage models the binary variable that determines whether the response is zero and the second stage uses a truncated count distribution to model non-zero count data. Welsh et al. (1996) found the zero-inflated Poisson and hurdle models to fit equally well but recommended using the hurdle model because its results are easier to interpret. Other recommendations in the literature are situational and guided by model selection techniques or by prior assumptions from the researcher. For exam-

ple, the zero-inflated modelling framework is preferred if the study design leads to count endpoints with both structural and sample zeros, conversely the hurdle model is preferred if the endpoint of interest only exhibits sample zeros.

4.1 Zero-inflated Poisson (ZIP) models

Lambert (1992) introduced the zero-inflated Poisson (ZIP) regression, which models the population as a mixture of an ordinary count model (e.g. Poisson model) and a distribution degenerate at zero. Let Y be a count random variable and let $y_1, \ldots, y_n$ be an observed random sample. As a result, the ZIP regression model can be described as follows:

$$Y_i = \begin{cases} 0 & \text{with probability } p_i + (1-p_i)e^{-\lambda_i} \\ y_i & \text{with probability } (1-p_i)e^{-\lambda_i}(\lambda_i)^{y_i}/y_i! \text{ for } y = 1, 2, 3, \ldots \end{cases}$$

The parameters p_i and λ_i may be functions of a set of explanatory variables. A logistic link function is used to model the probability p_i, while the log-link is used to model the mean λ_i. When p and λ are not functionally related the log-likelihood can be expressed as

$$\begin{aligned} l(\lambda, p; \mathbf{y}) &= \sum_{y_i=0} \log\left[p_i/(1-p_i) + \exp(-\lambda_i)\right] \\ &+ \sum_{y_i>0} \left[y_i \log(\lambda_i) - \lambda_i\right] + \sum_{i=1}^{n} \log(1-p_i) \end{aligned}$$

4.2 Zero-inflated Binomial (ZIB) models

In certain situations the presence of an upper bound on count data necessitate the use of a binomial model (Hall, 2000). Thus, the ZIP regression model proposed by Lambert (1992) can be modified to the following zero-inflated binomial model (ZIB):

$$Y_i = \begin{cases} 0 & \text{with probability } p_i + (1-p_i)(1-\pi_i^*)^{n_i^*} \\ y_i & \text{with probability } (1-p_i)\binom{n_i^*}{y_i}\pi_i^{*y_i}(1-\pi_i^*)^{n_i^*-y_i} \text{ for } y_i = 1, 2, 3, \ldots, n_i^* \end{cases}$$

where π^* is a vector of the binomial probabilities for the count data, which can be modelled using a logit link, and n_i^* is the number of trials for observation i.

4.2.1 ZIB Simulation

The DSDD method can be used to fit a ZIB regression with random effects. We perform a simulation study to examine its performance for this model. The simulated dataset contains 100 clusters with 5 observations each. The j^{th} response from the i^{th} cluster follow the ZIB model provided above with $\text{logit}(p_{ij}) = a_{0_i} + \beta_0 z_{ij}$ and $\text{logit}(\pi_{ij}^*) = a_{1_i} + \beta_1 z_{ij}$, where a_0 and a_1 are two random intercepts for the zero and count models, respectively. Parameters β_0 and β_1 are the fixed coefficients corresponding to the fixed covariate z. A description of the simulated parameters along with the DSDD results are presented in Table 4.1. Note that 57.8% of the simulated response were zeros. The parameters Δ_0, ξ, max_number_iterations_β, max_number_iterations_G, *precisionR*, *precisionM*, and *tolerance* of Section 2.3.1 were set to 0.5, 0.5, 20, 50, 0.0001, 0.05 and 0.001 respectively. The estimated mixture and coefficients were close to the true values and convergence was achieved at a gradient of 1.41e-05.

ZIB Simulation Specifications		
No.of clusters	100	
Size of each cluster	5	
True mixture	$\gamma = (\alpha_0, \alpha_1)$	π
	(-1,0)	0.5
	(1,2)	0.5
Fixed Cov. (z)	$\mathbf{z}_i$=(0, 0.25, 0.50, 0.75, 1.00) for the i^{th} cluster	
n_i^*	$n_i^* = 15$	
β_1	0.5	
β_0	1	
DSDD Results		
Estimated mixture	$\hat{\gamma} = (\hat{\alpha}_0, \hat{\alpha}_1)$	$\hat{\pi}$
	(-0.92,-0.09)	0.47
	(0.97,1.83)	0.53
$\hat{\beta}_1$	0.67 (0.42)	
$\hat{\beta}_0$	0.73 (0.47)	

Table 4.1: ZIB model simulation description and DSDD estimations

4.3 Hurdle models

Hurdle models have the ability to handle both zero-inflated and deflated datasets (Min and Agresti 2005). They consist of two parts: the first part is a binary process that calculates the probability of the response variable being at or below a hurdle h (which is zero when modelling zero-inflated data), the second part is a truncated model for the remaining observations, i.e. it truncates a known distribution such as the Poisson or negative binomial by conditioning on observations above the hurdle. To illustrate this process, suppose we wish to model the data using the hurdle model with $h = 0$. The first part of the process is the probability mass function to model Y_i at 0 for $i = 1, \ldots, n$ and the second part is a truncated probability mass function to model $P(Y_i|Y_i > 0)$. A general expression of the hurdle model is:

$$P(Y_i = 0) = 1 - p_i \tag{4.1}$$

$$P(Y_i > 0) = p_i \frac{f(y)}{1 - f(0)} \quad y = 1, 2, \ldots \tag{4.2}$$

where $f(.)$ is a probability distribution function to model the count data (such as the Poisson distribution). For ZIP regression models, the logistic link and the log-link functions are used for p_i and λ_i respectively, where λ_i is the mean of $f(.)$,

$$\text{logit}(p_i) = \mathbf{z}_{0i}^T \boldsymbol{\beta}_0, \text{ and } \log\lambda_i = \mathbf{z}_{1i}^T \boldsymbol{\beta}_1,$$

$\mathbf{z}_{0i}^T$ and $\mathbf{z}_{1i}^T$ are the i^{th} rows of the covariate matrices $\mathbf{z}_0$ and $\mathbf{z}_1$, having $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ as their respective coefficient vectors.

We extended the DSDD algorithm to handle zero-inflated data by fitting a hurdle model with nonparametric random effects. Hence, for a dataset with n clusters each of size n_i, p_{ij} and $f(y)$ in (4.1) and (4.2), for $j = 1, \ldots, n_i$, become

$$\begin{aligned} \text{logit}(p_{ij}) &= \mathbf{z}_{0ij}^T \boldsymbol{\beta}_0 + \mathbf{x}_{0ij}^T \boldsymbol{\gamma}_{0_i}, \\ f(Y_{ij}) &\sim Poisson\left(\exp\left\{\mathbf{z}_{1ij}^T \boldsymbol{\beta}_1 + \mathbf{x}_{1ij}^T \boldsymbol{\gamma}_{1_i}\right\}\right), \end{aligned} \tag{4.3}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ denote the fixed coefficient vectors of the zero and count model respectively. Similarly, $\boldsymbol{\gamma}_{0_i}$ and $\boldsymbol{\gamma}_{1_i}$ represent the random effect coefficient vectors. For simplicity, let $\mathbf{z}_{0ij}^T = \mathbf{z}_{1ij}^T$ and $\mathbf{x}_{0ij}^T = \mathbf{x}_{1ij}^T$ throughout the remainder of this discussion. As a result, the corresponding likelihood function is

$$L(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) = \prod_{i=1}^n \int \left[\prod_{j=1}^{n_i} (1 - p_{ij})^{I_{y_{ij}=0}} \left(p_{ij} \frac{f(y_{ij}; \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_{0_i}, \boldsymbol{\gamma}_{1_i})}{1 - f(0; \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_{0_i}, \boldsymbol{\gamma}_{1_i})} \right)^{1 - I_{y_{ij}=0}} \right] dG(\gamma_0, \gamma_1) \tag{4.4}$$

where $I(.)$ is an indicator function.

4.3.1 Hurdle model simulation

We simulated a zero-inflated dataset composed of 100 clusters with 5 observations each. The simulated responses followed the hurdle model

$$\begin{aligned} P(Y_{ij} = 0) &= 1 - p_{ij}, \\ P(Y_{ij} > 0) &= p_{ij} \frac{e^{-\lambda_{ij}} {\lambda_{ij}}^{y_{ij}} / y_{ij}!}{1 - e^{\lambda_{ij}}}, \end{aligned} \tag{4.5}$$

where $\lambda_{ij} = \exp(\alpha_{1_i} + \beta_1 z_{ij})$ and $\text{logit}(p_{ij}) = \alpha_{0_i} + \beta_0 z_{ij}$.

The simulated parameters in (4.5) are presented in Table 4.2. The parameters, Δ_0, ξ, max_number_iterations_β, *precisionR*, *precisionM*, max_number_iterations_G, and *tolerance* of the DSDD algorithm are set to 0.5, 0.5, 20, 0.0001, 0.05, 50 and 0.001, respectively. The results displayed in Table 4.3 show that the DSDD converges with a maximum gradient approximately 0 and the estimated fixed coefficients are close to their true values. However, the DSDD results suggest a greater bias in estimating the mass points of the mixture distribution which can affect predictions from the proposed model. We discuss the assessment of model fit in greater details in Chapter 6.

4.3.2 Pharmaceutical Data

Min and Agresti (2005) presented a pharmaceutical dataset where the goal of the study was to compare the effect of two treatments, A (TREAT1) and B (TREAT2), on the number of episodes of two of the treatments' side effects. It was conducted on 118 patients evenly split between the two treatments and measurements were taken at each of six visits where time between visits (Time) is considered a covariate in the

Hurdle Simulation 1		
No.of clusters	100	
Size of each cluster	5	
True mixture	$\gamma = (\alpha_0, \alpha_1)$	π
	(0,0)	0.7
	(2,4)	0.3
Fixed Cov. (z)	$\mathbf{z}_i$=(1.00, 1.75, 2.50, 3.25, 4.00)	
Fixed coef. (β_0)	0.5	
Fixed coef. (β_1)	-0.5	

Table 4.2: Description of the hurdle model simulation

$\hat{\boldsymbol{\gamma}} = (\hat{\alpha}_0, \hat{\alpha}_1)$	$\hat{\pi}$	$\hat{\beta}_0$	$\hat{\beta}_1$	Max Gradient	Iterations
(0.54,-0.5)	0.40	0.56	-0.46	9.8e-05	415
(0.53,-0.5)	0.21				
(-1.24,0.07)	0.19				
(0.70,3.91)	0.16				
(1.70,4.16)	0.05				
mean$(\hat{\alpha}_0, \hat{\alpha}_1) = (0.29, 0.54)$					
$(\sigma_{\hat{\alpha}_0}, \sigma_{\hat{\alpha}_1}) = (0.77, 1.77)$					

Table 4.3: DSDD results of the hurdle simulation

model. 83% of the observations were zeros. Min et al. (2005) present formal tests for over-dispersion giving strong evidence of zero-inflation and showing that a Poisson GLMM is inadequate.

Min et al. (2005) fit a Poisson random effect hurdle model and obtain the NPMLE of the mixing distribution using the EM algorithm with three mass points. The model described in their paper includes random intercepts and treats the covariates B (TREAT2) and log(Time) as fixed effects, hence $\gamma_0 = \alpha_0$ and $\gamma_1 = \alpha_1$. Model (4.3) becomes

$$\begin{aligned} \text{logit}(p_{ij}|\alpha_{0i}) &= \beta_{00} + \alpha_{0i} + \beta_{01}TREAT2_i + \beta_{02}\log(TIME)_{ij} \\ f(Y_{ij}|\alpha_{1i}) &\sim Poisson(\exp\{\beta_{10} + \alpha_{1i} + \beta_{11}TREAT2_i + \beta_{12}\log(TIME)_{ij}\}) \end{aligned} \tag{4.6}$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12})$ and $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02})$ are the sets of fixed coefficients of the count and zero models, respectively, and the random intercepts are α_1 and α_0. We use j to index measurements of the i^{th} patient for $i = 1, \ldots, 118$.

Table 4.4 compares the DSDD and EM fitted results. The fixed coefficient estimates of both algorithms were similar with the exception of β_{00}, which could be attributed to the disparity in the random intercept estimates of the two algorithms.

		EM algorithm	DSDD
	α_0	-0.99 1.2 1.69	-2.34 -2.11 -1.81 -0.36 1.32 1.40
	α_1	0.34 2.18 -2.15	1.89 0.09 0.12 -1.56 1.49 0.93
	π	0.34 0.24 0.42	0.25 0.19 0.05 0.26 0.11 0.14
	$(\bar{\alpha}_0, \bar{\alpha}_1)$	(0.66,-0.26)	(-0.83,0.38)
intercept	β_{00}	-2.813	-2.032
	β_{10}	-2.880	-2.913
TREAT	β_{02}	0.022	0.022
	β_{12}	0.490	0.484
log(TIME)	β_{01}	0.958	0.927
	β_{11}	0.898	0.905
	-2loglik	809.30	808.60

Table 4.4: Pharmaceutical study: Comparing the results of the DSDD and EM algorithms (Min and Agresti, 2005).

Chapter 5

Fitted values and random effects predictions

In this chapter we investigate different methods to compute the fitted values of the observed response $\mathbf{y}$, denoted by $\hat{\mathbf{y}}$, for nonparametric mixed models. The fitted values can then be used in assessing the fit and making inference about a given model.

In a clustered data setting, $\hat{\mathbf{y}}$ is calculated using two main methods :

1. Calculate $\hat{\mathbf{y}}_i$ by computing random effect predictions for cluster i and substituting them into the expression for $E[Y_{ij}] = \mu_{ij}$. Two prediction methods of the random effects are considered:

   (a) The first method computes an estimate of the random effect by calculating its posterior mean given the observed $\mathbf{y}$

   (b) the second method uses the mode of the random effect, i.e. the value from the estimated mixture distribution $\hat{G}$ that maximizes the probability of the observed cluster data.

2. Given the estimated mixture distribution of $\boldsymbol{\gamma}$, $\hat{G}$, obtain $\hat{\mathbf{y}}_i$ using the conditional expectation $E[\mu_{ij}|\mathbf{y}]$. For the Bernoulli and Poisson models, μ is equivalent to the parameters p and λ respectively.

5.1 Random effects predictors

5.1.1 Random effects posterior mean

Predicting $\boldsymbol{\gamma}_i$ in mixed models becomes more important when the focus of the study is on individual clusters and their associated random effects. Best linear unbiased predictors (BLUP) of the random effects are often used in linear mixed models (Golderberger, 1962; Henderson, 1984, McCulloch et al., 2008). Robinson (1991) extensively discusses different derivations of BLUPs such as: "the joint maximum likelihood estimates" suggested by Henderson (1950), which maximizes the joint density of the response and the random effects with respect to the fixed and random effects, and a Bayesian derivation suggested by Goldberger (1962) where the BLUPs are derived from the posterior distribution of the random effects. Conversely, Kiefer et al. (1956) regard the random effects as identically and independently distributed chance variables following a common distributional function G. They prove the consistency of the fixed coefficients MLE's when the random effects are independent variables, and they focus on obtaining ML estimators of the distribution G instead of attempting to determine particular values of the random effects.

McCulloch et al. (2008, pp. 304 and 317-318) calculate the *best* predictors of the random effects in generalized mixed models by minimizing the mean square error. Thus, for a given cluster i the predictor of $\boldsymbol{\gamma}_i$, denoted by $\tilde{\boldsymbol{\gamma}}_i$, minimizes the mean

squared error of the prediction,

$$E[(\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)' \mathbf{A} (\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)] = \int \int (\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)' \mathbf{A} (\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) f(\mathbf{y}, \boldsymbol{\gamma}) d\boldsymbol{\gamma} d\mathbf{y} \tag{5.1}$$

where $\mathbf{A}$ is a positive definite symmetric matrix. Minimizing (5.1) yields the conditional mean of $\boldsymbol{\gamma}_i$ given the observed response in cluster i, $\mathbf{y}_i$, that is,

$$\tilde{\boldsymbol{\gamma}}_i = E[\boldsymbol{\gamma}_i | \mathbf{y}_i]. \tag{5.2}$$

McCulloch's definition of "best" is different from the common definition which refers to the predictor that minimizes the variance.

McCulloch et al. (2008, pp. 305) shows that the best predictor is unbiased for sampling over $\mathbf{y}$, i.e.

$$E_{\mathbf{y}} \left[E_{\boldsymbol{\gamma}|\mathbf{y}} (\boldsymbol{\gamma}|\mathbf{y}) \right] = E [\boldsymbol{\gamma}] . \tag{5.3}$$

Furthermore, the prediction errors $\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ have a variance-covariance matrix over sampling on $\mathbf{y}$

$$var(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = E_{\mathbf{y}} [var (\boldsymbol{\gamma}|\mathbf{y})] . \tag{5.4}$$

As an example, consider the Bernoulli mixture model of the first simulation of Section 3.1.1 which includes a random intercept a_i, a random between slope, b_i, and a random within slope, w_i,

$$\begin{aligned} & y_{ij}|\gamma_i \sim \text{Bernoulli}(p_{ij}) \\ & \text{logit}(p_{ij}) = a_i + b_i x_{b_{ij}} + w_i x_{w_{ij}} \\ & \boldsymbol{\gamma}_i = (a_i, b_i, w_i) \sim G(\boldsymbol{\gamma}) \\ & i = 1, ..., n; j = 1, ..., n_i. \end{aligned} \tag{5.5}$$

	Sample							
	Bernoulli 1		Poisson 1		Bernoulli 2		Poisson 2	
	average	SD	average	SD	average	SD	average	SD
$(\tilde{a}_i - a_i)^2$	0.64	0.46	0.50	0.46	0.059	0.06	0.52	0.49
$(\tilde{b}_i - b_i)^2$	0.75	0.84	0.47	0.47	0.37	0.40	0.50	0.44
$(\tilde{w}_i - w_i)^2$	0.65	0.35	0.58	0.52				

Table 5.1: Comparing the averages and standard deviations (SD) of the mean squared errors of the random coefficient posterior means for the Bernoulli and Poisson simulations from Sections 3.1.1 and 3.2.1, respectively.

The posterior mean for the random intercept of the i^{th} cluster becomes

$$E[a_i|\mathbf{y}] = \tilde{a}_i = \frac{\sum_k a_k \pi_k \exp\left\{\sum_j y_{ij}\eta_{ijk} - \log\left[1+\exp(\eta_{ijk})\right]\right\}}{\sum_k \pi_k \exp\left\{\sum_j y_{ij}\eta_{ijk} - \log\left[1+\exp(\eta_{ijk})\right]\right\}} \tag{5.6}$$

where $\eta_{ijk} = a_k + b_k x_{b_{ij}} + w_k x_{w_{ij}}$ and (a_k, b_k, w_k) take values over the estimated support set, $\hat{G}$. Similar calculations are used to obtain $\tilde{b}_i$ and $\tilde{w}_i$. Table 5.1 reports the average and the standard deviation (SD) of the mean squared errors of the random coefficient predictors $\tilde{\boldsymbol{\gamma}} = (\tilde{a}, \tilde{b}, \tilde{w})$. The mean squared error for the random intercept predictor is $\sum_{i=1}^{n} (\tilde{a}_i - a_i)^2 / n$. Similar calculation is applied for the between and within random slopes.

The DSDD fitted values for the Sample 1 Bernoulli and Poisson simulations of Sections 3.1.1 and 3.2.1 are respectively:

$$\hat{y}_{ij} = \tilde{p}_{ij} = \frac{\exp\left\{\tilde{a}_i + \tilde{b}_i x_{b_{ij}} + \tilde{w}_i x_{w_{ij}}\right\}}{1 + \exp\left\{\tilde{a}_i + \tilde{b}_i x_{b_{ij}} + \tilde{w}_i x_{w_{ij}}\right\}} \tag{5.7}$$

and

$$\hat{y}_{ij} = \tilde{\lambda}_{ij} = \exp\left\{\tilde{a}_i + \tilde{b}_i x_{b_{ij}} + \tilde{w}_i x_{w_{ij}}\right\}. \tag{5.8}$$

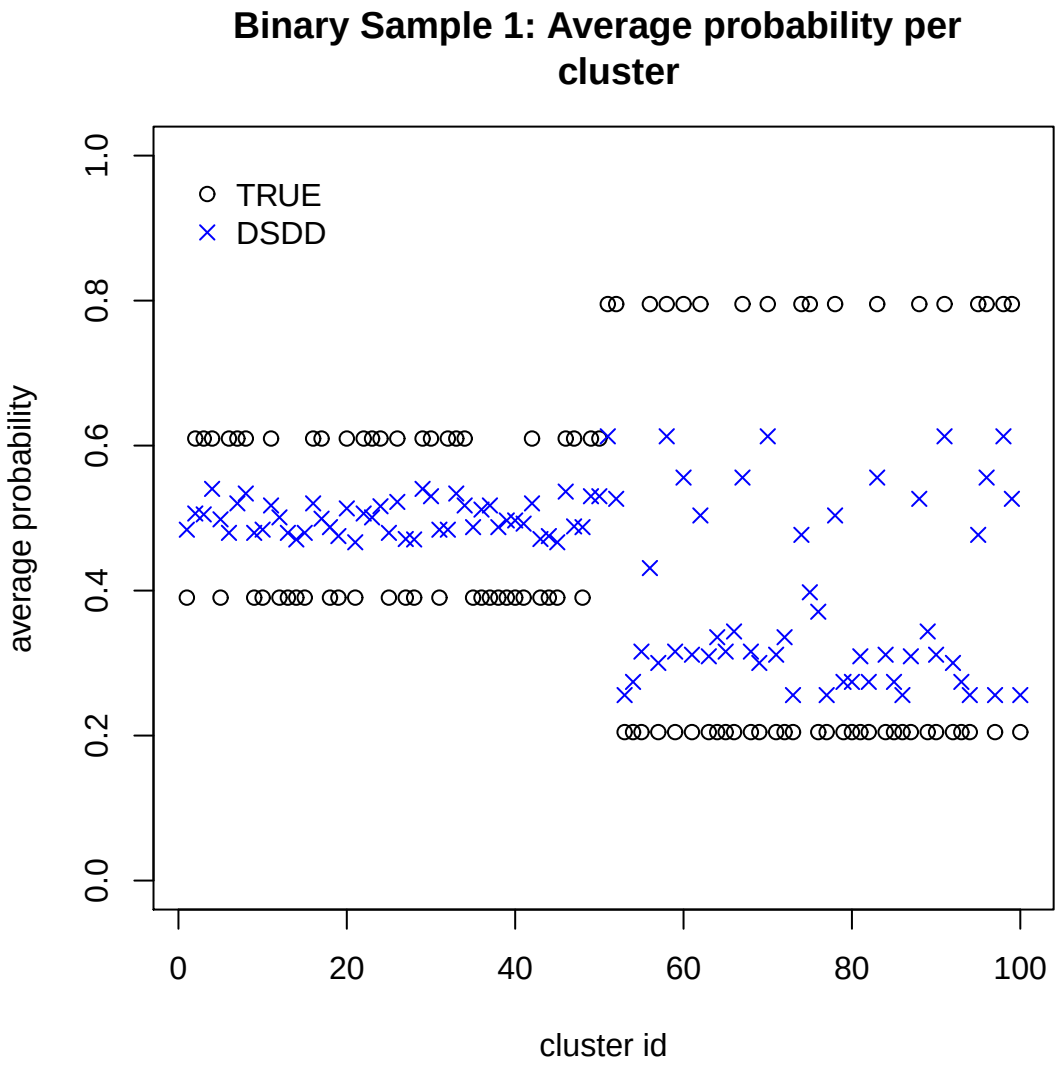

Figure 5.1: Bernoulli Sample 1 simulation: Comparison of the within cluster averages of the true probabilities p_{ij} and the fitted values $\hat{y}_{ij}$ obtained using (5.7), where the mixing distribution of the random effects is estimated by the DSDD algorithm.

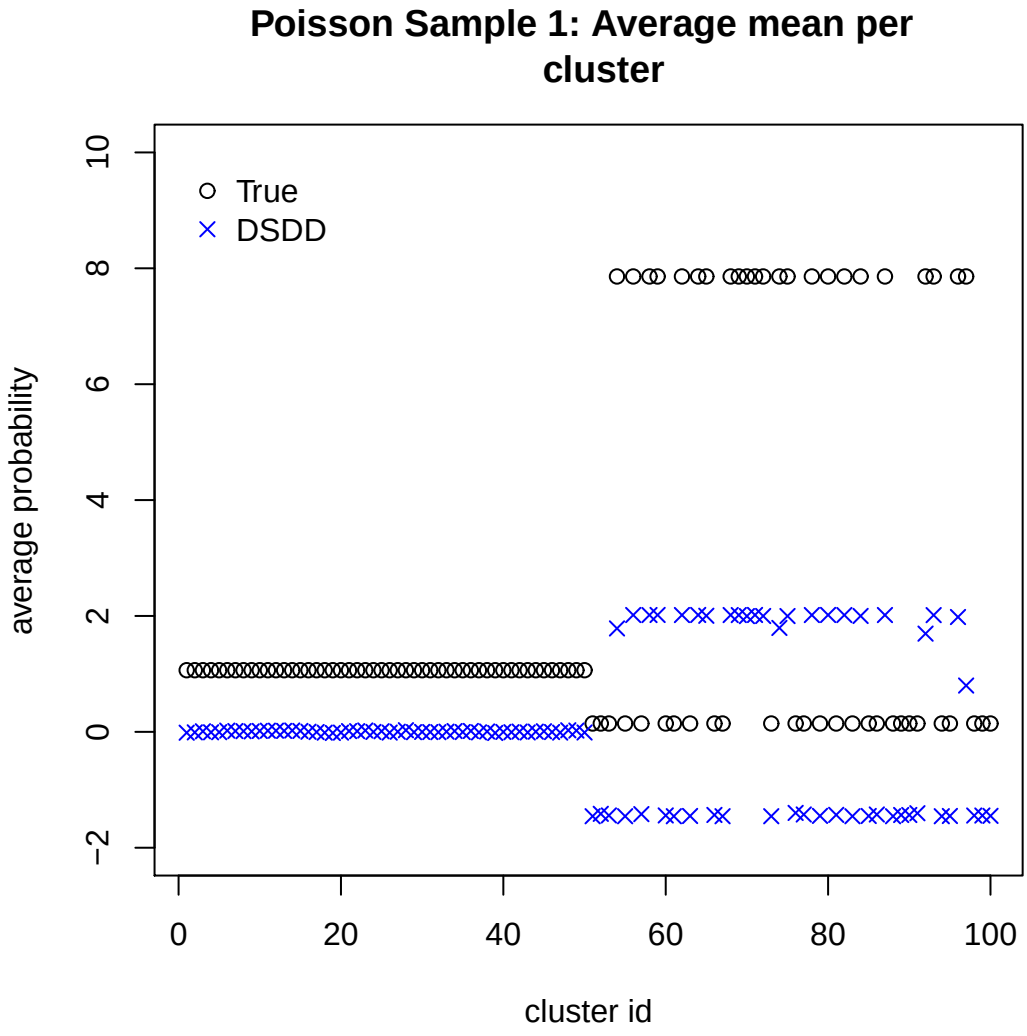

Figure 5.2: Poisson Sample 1 simulation: Comparison of the within cluster averages of the true parameter λ_{ij} and the fitted values $\hat{y}_{ij}$ obtained using (5.8) where the mixing distribution of the random effects is estimated by the DSDD algorithm.

A comparison of the within cluster averages of the true means, $\bar{p}_i = \frac{\sum_{j=1}^{n_i} p_{ij}}{n_i}$ and $\bar{\lambda}_i = \frac{\sum_{j=1}^{n_i} \lambda_{ij}}{n_i}$, and the fitted values in (5.7) and (5.8) are depicted in Figures 5.1 and 5.2 respectively.

5.1.2 Random effect modes

We can estimate the random effect coefficients using their respective modes from the estimated mixing distribution, $\hat{G}$. This method is often used in the literature, for instance Bates (2011) uses the conditional modes of the random effects for prediction purposes in nonlinear mixed effect models.

Bates (2011) computes modes in generalized mixed models for which the distribution of $Y|\boldsymbol{\gamma}$ is Gaussian,

$$Y|\boldsymbol{\gamma} \sim N(Z\boldsymbol{\beta} + X\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n). \tag{5.9}$$

Here, Y is the response variable, X and Z are considered the random and fixed effect matrices, respectively. I_n is the identity matrix and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \Sigma_\theta)$. The $q \times q$ positive definite variance-covariance matrix Σ depends on a variance component parameter vector θ.

Starting from the conditional distribution in (5.9), let $U = \Lambda_\theta \gamma$ where $\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^T$, which yields

$$\begin{aligned} Y = \mathbf{y}|\mathbf{U} = \mathbf{u} &\sim N(X\Lambda_\theta \mathbf{u} + Z\boldsymbol{\beta}, \sigma^2 I_q) \\ \mathbf{U} &\sim N(0, \sigma^2 I_q) \end{aligned} \tag{5.10}$$

Bates (2011) determines the conditional mode to be the following

$$\begin{aligned} \tilde{\mathbf{u}} &= arg\max_{\mathbf{u}} f_{U|Y}(\mathbf{u}|\mathbf{y}_{observed}) \\ &= arg\max_{\mathbf{u}} f_{Y|U}(\mathbf{y}_{observed}|\mathbf{u}) f_U(\mathbf{u}) \\ &= arg\max_{\mathbf{u}} h(\mathbf{u}). \end{aligned} \tag{5.11}$$

where

$$f_{Y|U}(\mathbf{y}_{observed}|\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{||\mathbf{y}_{obs} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}||^2}{2\sigma^2}\right) \tag{5.12}$$

and

$$f_U(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left(-\frac{||\mathbf{u}||^2}{2\sigma^2}\right). \tag{5.13}$$

Maximizing $-2\log[h(\mathbf{u})]$ leads to the following minimization problem

$$\tilde{\mathbf{u}} = arg\min_{\mathbf{u}} ||\mathbf{y}_{observed} - X\beta - Z\Lambda_\theta \mathbf{u}||^2 + ||\mathbf{u}||^2 \tag{5.14}$$

Bates states that this can be generalized to

$$\tilde{\mathbf{u}} = arg\min_{\mathbf{u}} ||\mathbf{y}_{observed} - \eta||^2 + ||\mathbf{u}||^2 \tag{5.15}$$

where $\eta = g(\mu)$ such that g(.) is the link function between η and the linear predictor $\mu = X\Lambda_\theta \mathbf{u} + Z\boldsymbol{\beta}$. The minimization of (5.15) is done through the iterative Gauss-Newton method for nonlinear least squares (Bates and Watts, 1988, Sect. 2.2.1). The lme4 package follows Bates' outline to obtain the posterior modes of the random effects. The fitted values are then calculated using the modes of the random effects and the estimates of the fixed coefficients.

For the estimated mixture distribution produced by the DSDD, $\hat{G}$, we calculate

the *posterior* modes of the random effects for cluster i by finding the value of the random effects component in $\hat{G}$ that yields the maximum *posterior* probability of cluster i given its observed data. Hence the *posterior* mode $\hat{\boldsymbol{\gamma}}$ is obtained as

$$\hat{\boldsymbol{\gamma}}_i = arg \max_{\boldsymbol{\gamma}_k} \pi_k \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\gamma}_k, \hat{\boldsymbol{\beta}}, \mathbf{x}_i, \mathbf{z}_i) \tag{5.16}$$

where f is the density function and $\hat{\boldsymbol{\beta}}$ is the fixed coefficient estimate obtained from the resulting mixture models and $\boldsymbol{\gamma}_k$ ranges over the support points of $\hat{G}$.

As an example, consider the Bernoulli Sample 2 simulation of Section 3.1.1. Estimates of p_{ij} under the mixed model are computed following the general model (3.1) and incorporating estimates of the random effects $\gamma_i = (a_i, b_i)$ and $\boldsymbol{\beta}$. The component of the estimated mixing distribution which yields the largest *posterior* probability of the random effects given the observed cluster data was used to estimate the random effects, a technique often used for clustered data (see Sofroniou (2006) and Li (2007)). For example, the vector of observations in the first cluster of Sample 2 is $\mathbf{y}_1 = (1, 1, 1, 1, 0)$. The largest *posterior* probability of the random effects given $\mathbf{y}_1$, is proportional to

$$\hat{\pi}_k \prod_{j=1}^{n_1} f(y_{1j}|\boldsymbol{\gamma}_{(k)}, \mathbf{z}_{1j}, \mathbf{x}_{1j}, \hat{\boldsymbol{\beta}})$$

and occurs for $k = 2$, hence $\hat{a} = 0.67$, $\hat{b} = -0.96$.

Table 5.2 reports the average and the standard deviation (SD) of the mean squared errors of the random coefficients modes $\hat{\boldsymbol{\gamma}}$. The mean squared error for the random intercept mode is $\sum_{i=1}^{n} (\hat{a}_i - a_i)^2 / n$, a similar calculation is applied for the between and within random slopes.

Similarly to (5.7) and (5.8), the fitted values for the first Bernoulli and Poisson

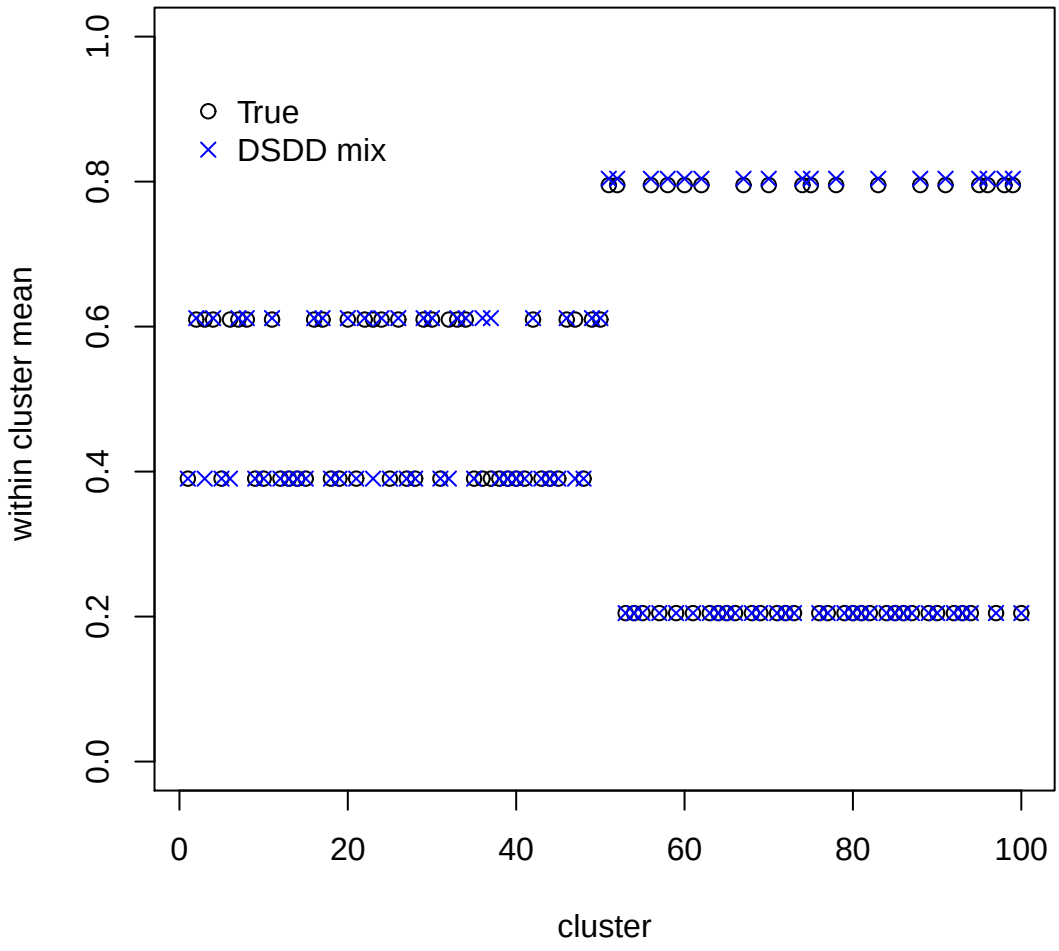

Figure 5.3: Bernoulli Sample 1 simulation: Comparison of the within cluster averages of the true probabilities p_{ij} and $\hat{y}_{ij}$ obtained using (5.17), when the mixing distribution of the random effects is estimated by the DSDD algorithm.

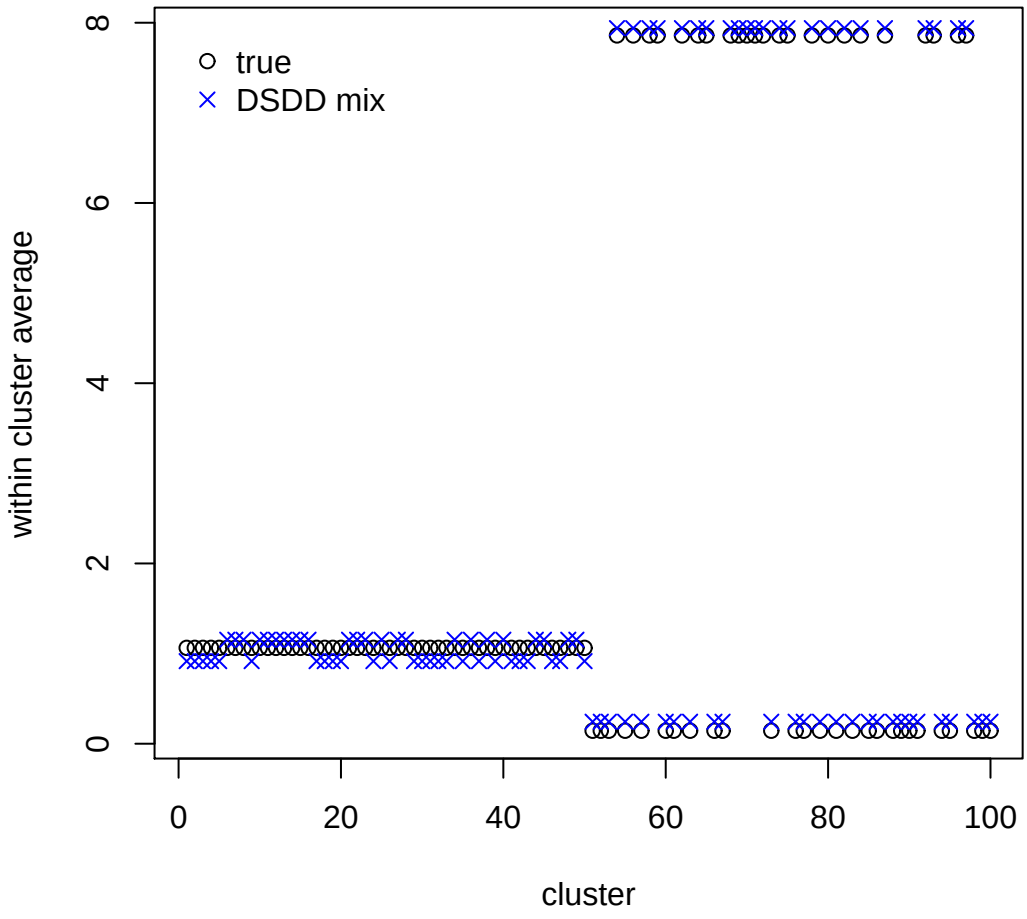

Figure 5.4: Poisson Sample 1 simulation: Comparison the within cluster averages of the true parameter λ_{ij} and $\hat{y}_i$ obtained using (5.18) when the mixing distribution of the random effects is estimated by the DSDD algorithm.

		Sample							
		Bernoulli 1		Poisson 1		Bernoulli 2		Poisson 2	
		average	SD	average	SD	average	SD	average	SD
DSDD	$(\hat{a}_i - a_i)^2$	0.28	1.03	0.46	1.04	0.03	0.07	0.17	0.81
	$(\hat{b}_i - b_i)^2$	0.28	1.04	0.49	1.22	0.11	0.65	0.18	0.71
	$(\hat{w}_i - w_i)^2$	0.27	1.02	0.54	0.86				
glmer	$(\hat{a}_i - a_i)^2$	1.42	1.40	0.45	0.61	0.33	0.54	0.19	0.30
	$(\hat{b}_i - b_i)^2$	1.45	1.56	0.44	0.67	0.25	0.29	0.18	0.31
	$(\hat{w}_i - w_i)^2$	1.33	1.25	0.51	0.50				

Table 5.2: Comparing the averages and standard deviations (SD) of the mean squared errors of the random coefficient posterior modes for the Bernoulli and Poisson simulations from Sections 3.1.1 and 3.2.1 respectively.

simulations of Sections 3.1.1 and 3.2.1 are obtained respectively,

$$\hat{y}_{ij} = \hat{p}_{ij} = \frac{\exp\left(\hat{a}_i + \hat{b}_i x_{b_{ij}} + \hat{w}_i x_{w_{ij}}\right)}{1 + \exp\left(\hat{a}_i + \hat{b}_i x_{b_{ij}} + \hat{w}_i x_{w_{ij}}\right)} \tag{5.17}$$

and

$$\hat{y}_{ij} = \hat{\lambda}_{ij} = \exp\left(\hat{a}_i + \hat{b}_i x_{b_{ij}} + \hat{w}_i x_{w_{ij}}\right) \tag{5.18}$$

where $\hat{a}_i$, $\hat{b}_i$ and $\hat{w}_i$ are the modes resulting from the DSDD estimated mixtures.

Figures 5.3 and 5.4 compare the within cluster averages of the true means, $\bar{p}_i$ and $\bar{\lambda}_i$, and the fitted values in (5.17) and (5.18) for the Bernoulli and Poisson Sample 1 simulations respectively.

5.2 *Posterior* means for μ

A canonical link can be the focus in the model (Spiegelhalter, 2002), in which case it may be more appropriate to obtain fitted values for $\mathbf{y}$ that are based on the *posterior*

means of p and λ for the Bernoulli and Poisson mixtures:

$$\begin{aligned} E[p_{ij}|y_{ij}] &= \frac{\int p_{ij}^{y_{ij}+1}(1-p_{ij})^{1-y_{ij}}d\hat{G}_{\boldsymbol{\gamma}}}{\int p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}d\hat{G}_{\boldsymbol{\gamma}}} \\ E[\lambda_{ij}|y_{ij}] &= \frac{\int \exp^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}+1}d\hat{G}_{\boldsymbol{\gamma}}}{\int \exp^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}d\hat{G}_{\boldsymbol{\gamma}}} \end{aligned} \tag{5.19}$$

where

$$\begin{aligned} \text{logit}\left[p_{ij}\left(\boldsymbol{\gamma},\hat{\boldsymbol{\beta}}\right)\right] &= \mathbf{x}_{ij}^T\boldsymbol{\gamma}+\mathbf{z}_{ij}\hat{\boldsymbol{\beta}} \\ \log\left[\lambda_{ij}\left(\boldsymbol{\gamma},\hat{\boldsymbol{\beta}}\right)\right] &= \mathbf{x}_{ij}^T\boldsymbol{\gamma}+\mathbf{z}_{ij}\hat{\boldsymbol{\beta}}, \end{aligned} \tag{5.20}$$

and $\hat{G}_{\boldsymbol{\gamma}}$ is the estimated distribution for $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\beta}}$ is the estimated fixed coefficient vector.

5.3 Result comparison

To compare fitted values of the response, y_{ij}, calculated by the three methods, we used Brier scores (Harrell, 2001)

$$\text{Brier} = \frac{1}{N}\sum_i\sum_j(\hat{y}_{ij}-y_{ij})^2 \tag{5.21}$$

and the average relative square bias (RSB)

$$RSB = \frac{1}{N}\sum_i\sum_j\frac{(\hat{y}_{ij}-\mu_{ij})^2}{\mu_{ij}} \tag{5.22}$$

where μ denotes the true parameters p and λ for the Bernoulli and Poisson models respectively. Table 5.3 lists the Brier and RSB scores obtained using the three different

Sample	$Brier_{BP}$	$Brier_{mode}$	$Brier_{par}$	RSB_{BP}	RSB_{mode}	RSB_{par}
Binary Sample 1	0.21	0.18	0.16	0.09	0.023	0.023
Binary Sample 2	0.13	0.11	0.07	0.24	0.13	0.31
Poisson Sample 1	2.40	2.08	1.92	0.14	0.09	0.29
Poisson Sample 2	3.93	1.90	1.51	0.40	0.06	0.55

Table 5.3: Comparing the Brier and RSB scores obtained by computing the fitted values of $\mathbf{y}$ using the methods described in Chapter 5.

methods for the simulated datasets described in Chapter 3. Note that the subscript "BP" indicates that the fitted values for y_{ij} were obtained using the posterior means of the random effects, "*mode*" refers to the method that uses the *posterior* modes for the random effects, and lastly the subscript "*par*" refers to the method using the *posterior* means of λ_{ij} and p_{ij}.

Lower Brier and RSB scores indicate a better agreement between the fitted and observed values. The results of Table 5.3 suggest that $Brier_{par}$'s are the smallest but comparable to $Brier_{mode}$. Nevertheless, RSB_{mode}'s are the smallest for all samples, therefore the mode is favoured as an estimate of the random effects compared with the alternative methods.

Table 5.4 compares the Brier scores of the DSDD and glmer results of the simulated datasets, the NBA and pharmaceutical case studies. The Brier scores for both algorithms were computed using the modes of the random effects calculated by the method shown in Section 5.1.2. For the NBA case study, the estimated fitted probability for team i in year j is

$$\hat{p}_{ij} = \frac{\exp\left(\hat{a}_i + \hat{b}_i Pay_{ij} + 0.72 age_{ij}\right)}{1 + \exp\left(\hat{a}_i + \hat{b}_i Pay_{ij} + 0.72 age_{ij}\right)}, \tag{5.23}$$

Data	DSDD	glmer
Binary Sample 1	0.18	0.18
Binary Sample 2	0.11	0.09
Poisson Sample 1	2.08	1.89
Poisson Sample 2	1.90	1.61
NBA case study	0.17	0.17
Pharmaceutical data	0.62	0.63

Table 5.4: Comparing the Brier scores for the glmer and DSDD results of the simulated datasets and two case studies.

where

$$\hat{\boldsymbol{\gamma}}_i = (\hat{a}_i, \hat{b}_i) = argmax_{\boldsymbol{\gamma}_{(k)}} \ \hat{\pi}_k \prod_{j=1}^{n_i} \frac{\left[\exp(a_{(k)} + b_{(k)} Pay_{ij} + 0.72 age_{ij})\right]^{y_{ij}}}{1 + \exp(a_{(k)} + b_{(k)} Pay_{ij} + 0.72 age_{ij})} \tag{5.24}$$

maximizes the posterior probability of the random effects given the data, the values $\boldsymbol{\gamma}_{(k)} = (\alpha_{(k)}, b_{(k)})$ and $\pi_{(k)}$ run over the set of estimates given in Table 3.8. Similarly, fitted values were obtained from the glmer algorithm by fitting model (3.10). Figure 5.5 shows the true average probability of playoffs appearance per team compared with the DSDD and glmer estimations. Overall, the plot shows that the DSDD model estimates were closer to the observed proportions.

Fitted values for the pharmaceutical case study are computed using

$$\hat{y}_{ij} = \left[\frac{\exp(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}_0 + \mathbf{x}_{ij}^T \boldsymbol{\gamma}_{0i}^{max})}{1 + \exp(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}_0 + \mathbf{x}_{ij}^T \boldsymbol{\gamma}_{0i}^{max})}\right] \left\{\frac{\exp(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}_1 + \mathbf{x}_{ij}^T \boldsymbol{\gamma}_{1i}^{max})}{1 - \exp\left[-\exp(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}_1 + \mathbf{x}_{ij}^T \boldsymbol{\gamma}_{1i}^{max})\right]}\right\}, \tag{5.25}$$

where $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1)$ are the estimated fixed coefficients for the zero and count parts respectively. Table 5.5 suggests that the fitted values of the DSDD and EM algorithms poorly describe the observed ones. Notably, there is a significant discrepancy at the upper tail and most values are packed around the observed mean.

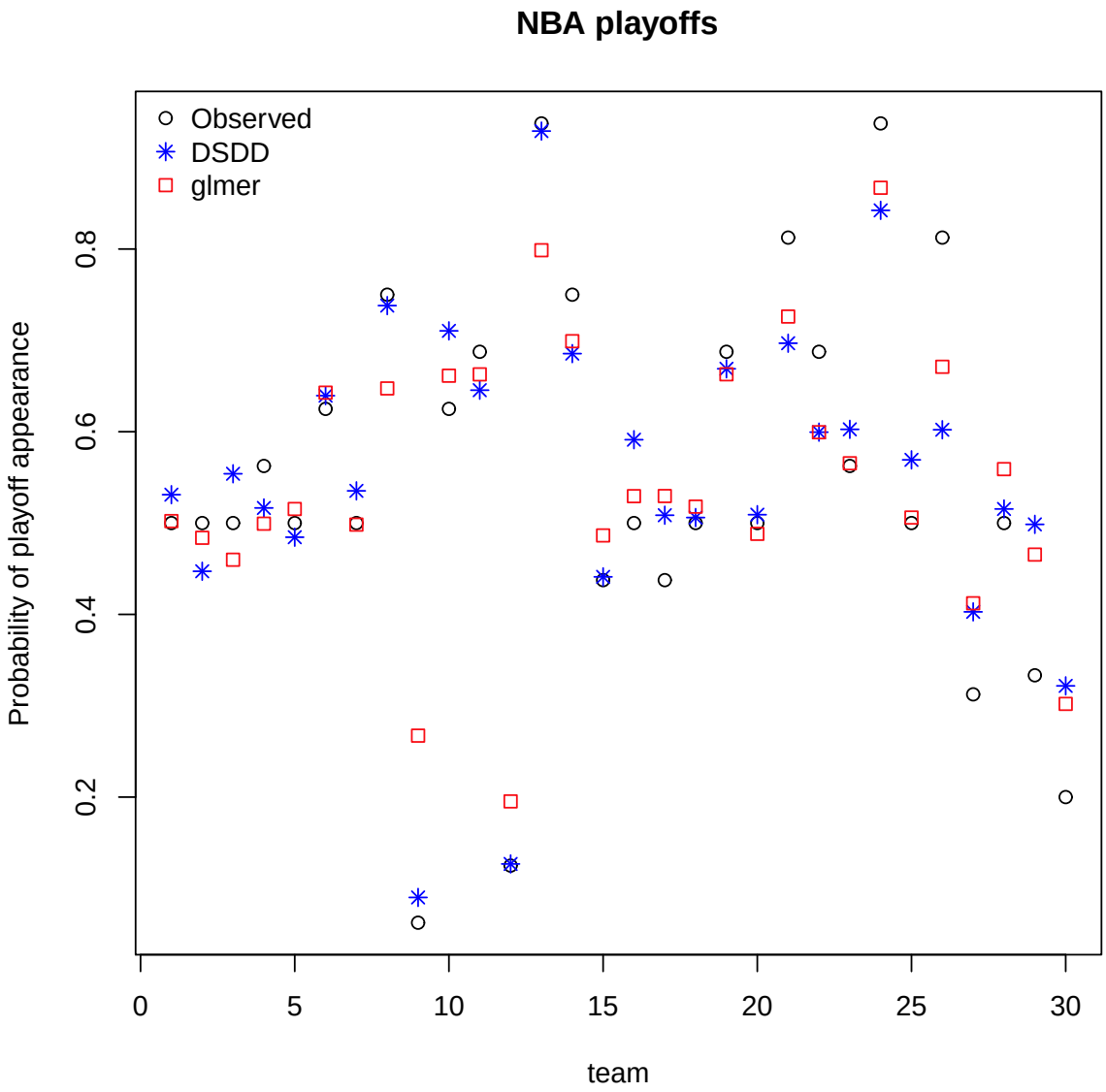

Figure 5.5: A comparison of the estimated playoff appearance probabilities from the mixed effect model fitted by the DSDD and glmer routines to the observed proportions for each NBA team.

fitted values	Min	1st Qu	Median	Mean	3rd Qu	Max
DSDD	0.23	0.38	0.45	0.49	0.64	0.86
EM	0.082	0.086	0.14	0.15	0.21	0.23
observed	0.00	0.00	0.00	0.29	0.00	6.00

Table 5.5: Pharmaceutical case study: summary statistics of the fitted values obtained from the DSDD and EM algorithms compared with the summary of the observed values.

Chapter 6

Model Selection and fit

Model determination consists of two components: goodness-of-fit and model choice or selection. Goodness-of-fit methods attempt to assess whether the model is adequate while model selection concerns the choice of the most plausible model within a collection of models under consideration. This chapter investigates different goodness-of-fit techniques, which are subsequently used to compare the DSDD estimated models with their counterparts (e.g. the glmer and EM estimates). Comparing fitted to the observed values as in Chapter 5 is one potential approach, however this does not take into account the complexity or the hierarchical structure of the models.

Classical goodness-of-fit tests used for models with only fixed effects are not appropriate for mixed models due to the complexity of their error structures. The assumption of normality of the random effects is often adopted in mixed models, and many goodness-of-fit methods available in the literature assess the validity of such distributional assumptions in GLMM. Lange and Ryan (1989) graphically check the normality assumption of the random effects using empirical Bayes estimates. Hwang and Wei (2006) transform correlated residuals in linear mixed models to uncorrelated residuals in order to apply classical tests of normality. Claeskens and Hart (2009)

proposed several formal tests of the random effects normality assumption. Nevertheless, the focus in this chapter is on methods that can be used to compare mixture models where the random effects do not necessarily follow the normality assumption.

We split the methods in the chapter into three groups: the first group is more suited for model selection and includes two likelihood-based penalized criteria, the DIC and cAIC. The second group focuses on cross-validation and predictive checking techniques. Lastly, we investigate modifications of the classical coefficient of determination, R^2, to accommodate mixed models.

6.1 Penalized information criteria for model selection

Penalized information criteria have the general form:

Model criteria=fit + Penalty for model complexity.

The adequacy of the model incorporates a value for the model fit, usually the log-likelihood or the deviance (-2 times the log-likelihood), plus a penalty that is indicative of the model complexity. Popular criteria include the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978). The AIC and BIC are defined as $-2L+2M$ and $-2L+Mlog(n)$ respectively, where L is the maximized log-likelihood of the associated model, M is the number of parameters included in the model and n is the number of observations. Although both criteria perform well in some cases, particularly in fixed effects models, their implementation is problematic when models differ in their hierarchical structure and include random effects (Spiegelhalter et al., 2002).

6.1.1 Deviance Information Criterion

Spiegelhalter et al. (1998) introduced a more general measure based on the posterior distribution of the deviance statistics, $D(\boldsymbol{\gamma}) = -2\log f(\mathbf{y}|\boldsymbol{\gamma}) + 2\log[h(\mathbf{y})]$ where $f(\mathbf{y}|\boldsymbol{\gamma})$ is the likelihood function of the data given the parameter vector $\boldsymbol{\gamma}$ and $h(\mathbf{y})$ is a normalizing term that is a function of the data alone. The calculation of $h(\mathbf{y})$ is often omitted when comparing different models because it is constant with respect to the parameter and cancels out in DIC comparisons for models with the same likelihood function. The DIC, defined as $2E_{\boldsymbol{\gamma}|\mathbf{y}}[D(\boldsymbol{\gamma})] - D[E(\gamma|\mathbf{y})]$, is considered a generalization of AIC, in fact the two measures are approximately equal for non-hierarchical models with fixed effects.

The complexity of the model is captured by the effective number of parameters, p_D, which is shown to be the difference of the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters (Spiegelhalter et al., 1998)

$$p_D = E_{\boldsymbol{\gamma}|\mathbf{y}}[D(\boldsymbol{\gamma})] - D[E(\boldsymbol{\gamma}|\mathbf{y})] = \bar{D}(\boldsymbol{\gamma}) - D(\bar{\boldsymbol{\gamma}}). \tag{6.1}$$

The DIC can then expressed as

$$DIC_1 = \bar{D} + p_D. \tag{6.2}$$

The complexity, p_D, can be thought of as the number of unconstrained parameters in the model. The contribution of a parameter to the penalty term counts closer to 1 if it has less prior information and constraints, and it is closer to zero if more constraints are imposed on the parameter or if most of the information stems from its prior distribution (Gelman et al., 2004 p: 182). As a result, each fixed effect coefficient estimated by the MLE in the models discussed in this dissertation contributes zero

parameters in the model since it has maximum constraint.

The DIC is a model comparison method where smaller values indicate a better model, however it is not intended as a method to identify the correct model. This measure has become a staple in Bayesian applications. Spiegelhalter et al. (2002) provided an asymptotic justification of the criterion particularly when the number of observations n increases with respect to the number of parameters and when the prior is completely specified. They further show that DIC for an assumed model is approximately the posterior expectation of a logarithmic loss function, $-2\log\left[p(\mathbf{y}|\bar{\boldsymbol{\gamma}})\right]$, i.e. the DIC describes the expected posterior loss when adopting a particular model. We derive DIC measurements for clustered datasets modelled with Poisson and binomial mixtures discussed in this dissertation. Full Bayesian analyses assign prior distributions to the random effects and obtain their posterior distributions. The DSDD estimates of the mixing distribution are not posterior distributions in a Bayesian sense; our analysis is closer to empirical Bayes methods for mixture models where the mixture distribution is estimated empirically using observed information, which is then used to obtain posterior probabilities of the parameters (Deely et al., 1981).

Following a suggestion from Gelman et al. (2004, p:181), we estimate the posterior expectation of the deviance, $\bar{D} = E_{\gamma|\mathbf{y}}\left[D(\boldsymbol{\gamma})\right]$, by simulating $\boldsymbol{\gamma}$ from the estimated mixing distribution, $\hat{G}$,

$$\hat{\bar{D}}(\boldsymbol{\gamma}) = \frac{1}{L}\sum_{l=1}^{L} D(\boldsymbol{\gamma}^l) \tag{6.3}$$

where $\boldsymbol{\gamma}^l$ is the l^{th} draw. Note that the glmer estimated mixture, $\hat{G}$, is multivariate normal with means $\hat{B}_a$, $\hat{B}_b$ and $\hat{B}_w$ for the three random effects. Table 6.1 displays the DIC and pD results of the glmer and DSDD estimated mixed models for the Bernoulli and Poisson simulations in Tables 3.2 and 3.3 and Tables 3.6 and 3.7. We used the mean of the estimated mixing distribution, $\hat{G}$, in place of $\bar{\boldsymbol{\gamma}}$ in $D(\bar{\boldsymbol{\gamma}})$ to calculate DIC

Sample	DIC_1^{DSDD} (pD)	DIC_1^{glmer} (pD)
Binary Sample 1	890.55 (101.15)	983.9 (147.45)
Binary Sample 2	753.53 (31.87)	1654.95 (477.31)
Poisson Sample 1	6194.99 (1329.66)	10626.18 (3461.05)
Poisson Sample 2	4445.41 (942.10)	5767.24 (1604.16)

Table 6.1: DIC_1 results for binary and Poisson simulations of Chapter 3

in the DSDD case.

One can justify using the posterior median or mode as an alternative to the posterior mean in place of $\bar{\boldsymbol{\gamma}}$ in (6.1), hence the DIC measure is dependent on the estimate of $\boldsymbol{\gamma}$ and on the parametrization of the model (Spiegelhalter, 2002). Furthermore, Spiegelhalter (2003) warns against naively applying it to a mixture model. Extensions of the DIC were provided by Celeux et al. (2006) in the setting of missing data models. Their work is mainly based on replacing $E(\boldsymbol{\gamma}|\mathbf{y})$ with other estimates and integrating the deviance with respect to missing data.

Motivated by Celeux's work, we present alternative DIC measures using the parameter estimates given in Section 5.1 for the central deviance $D(\bar{\gamma})$. Firstly, we consider the deviance at the mode, $\hat{\boldsymbol{\gamma}}_i$, for the i^{th} cluster as in Section 5.1.2. This version of the DIC (DIC_2 using Celeux's notation) is written as:

$$\begin{aligned} DIC_2 &= -4E_{\boldsymbol{\gamma}}\left[\log f(\mathbf{y}|\boldsymbol{\gamma})|\mathbf{y}\right] + 2\log f(\mathbf{y}|\hat{\boldsymbol{\gamma}}(\mathbf{y})) \\ &= 2\bar{D}(\gamma) - D(\hat{\boldsymbol{\gamma}}(\mathbf{y})) \end{aligned} \tag{6.4}$$

where $\hat{\boldsymbol{\gamma}}(\mathbf{y}) = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_n)$. Combining (6.4) with (6.3), The estimated average deviance in the Bernoulli case becomes

$$\hat{\bar{D}}(\boldsymbol{\gamma}) = \frac{-2}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i} -2\left[y_{ij}\eta_{ij}^l - \log(1+\exp\left\{\eta_{ij}^l\right\})\right] + 2\log(h(\mathbf{y})) \tag{6.5}$$

where $\eta_{ij}^l = \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\boldsymbol{\gamma}_i^l$ and $\boldsymbol{\gamma}_i^l = (\alpha_i^l, b_i^l, w_i^l)$ is the l^{th} simulated mass point from the estimated mixing distribution $\hat{G}(\boldsymbol{\gamma})$. Whereas, in the Poisson case, the estimated average deviance in (6.4) becomes:

$$\hat{\bar{D}}(\boldsymbol{\gamma}) \quad = \quad \frac{-2}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{-\lambda_{ij}^l + y_{ij}\log(\lambda_{ij}^l)\right\} + 2\log\left[h(y_{ij})\right] \tag{6.6}$$

where $\log(\lambda_{ij}^l) = \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\boldsymbol{\gamma}_i^l$.

The calculation for $D(\hat{\boldsymbol{\gamma}})$ uses the posterior mode of the random effect vectors, the Bernoulli and Poisson derivations are respectively the following:

$$D(\hat{\boldsymbol{\gamma}}) = -2\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\mathbf{y}\eta_{\hat{\boldsymbol{\gamma}}_i} - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\log\left[1+\exp\left(\eta_{\hat{\boldsymbol{\gamma}}_i}\right)\right]\right\} + 2\sum_{i=1}^{n}\sum_{j=1}^{n_i}h(y_{ij}) \tag{6.7}$$

where $\eta_{\hat{\boldsymbol{\gamma}}_i} = \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\hat{\boldsymbol{\gamma}}_i$ and

$$D(\hat{\boldsymbol{\gamma}}) = -2\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[-(\lambda_{\hat{\boldsymbol{\gamma}}_i} + y_{ij}\log(\lambda_{\hat{\boldsymbol{\gamma}}_i})\right] + 2\sum_{i=1}^{n}\sum_{j=1}^{n_i}h(y_{ij}) \tag{6.8}$$

where $\lambda_{\hat{\boldsymbol{\gamma}}_i} = \exp\left(\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\hat{\boldsymbol{\gamma}}_i\right)$.

Table 6.2 displays the DIC results of the glmer and DSDD estimated mixed models for the Bernoulli and Poisson simulations of Tables 3.2, 3.3, 3.6 and 3.7. The results show a big discrepancy between the DIC measurements of the glmer and DSDD models largely due to the penalty term pD.

Secondly, we investigate the DIC measure by focusing on the Bernoulli and Poisson mean canonical links as parameters of interest, their DIC measures become respec-

Sample	DIC_2^{DSDD} (pD)	DIC_2^{glmer} (pD)
Binary Sample 1	883.01 (34.48)	1134.36 (297.82)
Binary Sample 2	1761.56 (692.93)	2000.29 (823.53)
Poisson Sample 1	8412.60 (3543.78)	13163.41 (5926.19)
Poisson Sample 2	4493.25 (980.16)	7028.55 (2867.50)

Table 6.2: DIC_2 results for binary and Poisson simulations of Chapter 3.

tively

$$DIC_p = 2\bar{D} - D\left[E(p)\right] \tag{6.9}$$

and

$$DIC_\lambda = 2\bar{D} - D\left[E(\lambda)\right]. \tag{6.10}$$

$\bar{D}$ is estimated by $\hat{\bar{D}}$ following (6.3). The deviance, $D\left[E(\mu)\right]$, becomes the deviance evaluated at the mean of the parameter,

$$E(\mu) = \sum_i \sum_j \int \mu_{ij}(\boldsymbol{\gamma}) d\hat{G}(\boldsymbol{\gamma})$$

where $\mu_{ij}(\boldsymbol{\gamma}) = \lambda_{ij}(\boldsymbol{\gamma}) = \exp(\mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\boldsymbol{\gamma})$ and $\text{logit}\left[\mu_{ij}(\boldsymbol{\gamma})\right] = \mathbf{z}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{x}_{ij}^T\boldsymbol{\gamma}$ for the Poisson and Bernoulli models respectively. This is easily approximated when the mixing distribution is discrete and nonparametric but can be computationally challenging for other forms of the estimated mixture $\hat{G}(\boldsymbol{\gamma})$. We can overcome the computational difficulty by drawing $\boldsymbol{\gamma}$ from the estimated mixing distribution, $\hat{G}(\boldsymbol{\gamma})$, akin to what is done in (6.7), i.e. $E(\mu_{ij}) = \frac{1}{L}\sum_l \sum_i \sum_j \mu_{ij}(\boldsymbol{\gamma}^l)$ where $\boldsymbol{\gamma}^l$ is the l^{th} draw of $\boldsymbol{\gamma}$ from the estimated mixing distribution, $\hat{G}(\boldsymbol{\gamma})$. Using this method the DIC values for our simulation studies are displayed in Table 6.3.

Note that Spiegelhalter et al. (2002) considers the DIC computation when the canonical parametrization in exponential families is the focus of the model. They

Sample	DIC_{μ}^{DSDD} (pD)	DIC_{μ}^{glmer} (pD)
Binary Sample 1	1010.69 (161.71)	988.69 (151.92)
Binary Sample 2	1475.24 (408.08)	1699.61 (521.77)
Poisson Sample 1	7170.27 (2310.03)	11212.09 (3974.83)
Poisson Sample 2	4882.50 (1368.92)	6066.42 (1900.25)

Table 6.3: DIC_{μ} results for binary and Poisson simulations of Chapter 3 using estimates of the parameters $\mu = p$ and $\mu = \lambda$ for Bernoulli and Poisson samples respectively.

DIC measure	with xb and xw	without xb and xw
DIC_p (pD)	1010.69 (161.71)	876.49 (92.83)
DIC_2 (pD)	883.01 (34.48)	830.95 (47.38)
DIC_1 (pD)	890.55 (101.15)	875.19 (92.17)

Table 6.4: Comparing DIC measures for two suggested models of the first Bernoulli simulation in Section 3.1.1; the first model includes both random covariates while the second excludes the covariates and includes one random intercept only.

explore the simple Poisson and binomial models with conjugate priors, then they examine pD in generalized linear models and generalized linear mixed models. Their computations highlight the lack of invariance of the complexity measure pD.

To compare the performance of the four DIC's discussed in this section we consider the first Bernoulli simulation of Section 3.1.1 where the true model included a between and within random covariates, xb and xw respectively. We compute DIC's of models that include both random covariates and the model that omits them. Table 6.4 displays measurements of the two suggested models. All DIC measures indicate that the model with two random covariates xb and xw is less favourable than the alternative. It also highlights an inconsistancy in measuring pD, specifically in DIC_2 where pD increases from 34.48 to 47.38 despite the presence of less covariates in the model.

The aforementioned DIC derivations can also be generalized to hurdle and zero-

inflated models. Our results show that the effective number of parameter significantly changes with the type of parametrization and the focus of the model, i.e. the type of estimators used to estimate $D(\hat{\boldsymbol{\gamma}})$ will greatly influence the DIC, for example using the overall mean of the random effects as an estimator for $\boldsymbol{\gamma}$ in DIC_1 removes the focus from individual clusters when it comes to random effect estimation. Furthermore, an error of estimation should be considered when using a simulation based method as in 6.3. Our application of the DIC requires more investigation since it is outside the Bayesian framework making the asymptotic justifications provided by Spiegelhalter (2002) invalid.

6.1.2 Conditional Akaike Information

When the model contains random effects, the formula for the AIC measure, -2loglikelihood $+2M$, is vague with respect to the choice of likelihood and whether the random effects should be counted in the penalty term M. In fact, the AIC differs according to the research focus (Vaida and Blanshard, 2005): when the focus is on the population parameter the likelihood in question is the marginal likelihood and M is the number of fixed parameters in the model, however when the research focuses on cluster specific inferences, the likelihood becomes the conditional likelihood and $M = \rho$, where ρ is the effective degrees of freedom of the linear mixed model defined in Hodges and Sargent (2001).

Vaida and Blanshard (2005) proposed the conditional Akaike information (cAI) as a model selection tool for linear mixed models in the analysis of clustered data

when the research focus is cluster specific, which they define as

$$\begin{aligned} cAI &= -2E_{f(\mathbf{y},\boldsymbol{\gamma})}E_{f(\mathbf{y}^0|\boldsymbol{\gamma})}l\left(\mathbf{y}^0|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right) \\ &= -2\int l\left(\mathbf{y}^0|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right) f(\mathbf{y}^0|\boldsymbol{\gamma})f(\mathbf{y},\boldsymbol{\gamma})d\mathbf{y}^0 d\mathbf{y} d\boldsymbol{\gamma} \end{aligned} \tag{6.11}$$

where $\mathbf{y}^0$ is a prediction dataset independent of $\mathbf{y}$ such that $\mathbf{y}^0$ and $\mathbf{y}$ come from the same conditional distribution $f(.|\boldsymbol{\gamma})$. $\hat{\boldsymbol{\beta}}$ is an estimator of the fixed coefficient such as the MLE.

Vaida and Blanshard (2005) proceed to show that an unbiased estimator of cAI is the conditional Akaike information criterion (cAIC) defined as

$$cAIC = -2\log\left[f(\mathbf{y}|\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})\right] + 2\rho \tag{6.12}$$

where $\hat{\boldsymbol{\gamma}}$ is the empirical Bayes estimator of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}} = E(\boldsymbol{\gamma}|\mathbf{y})$. A formula for ρ is given in the general case of unknown variance parameters in finite samples.

The cAIC measure is closely related to the deviance information criterion discussed in Section 6.1.1 which also emphasizes the research focus for hierarchical models and implicitly distinguishes between conditional and marginal inferences. Spiegelhalter et al. (2002) show that under a flat prior the model complexity measure, p_D, is the same as ρ, and as a result Vaida (2005) notes that the cAIC and DIC are the same for mixed effects models focus with known variances.

Donohue et al. (2011) later extended the approach to provide an estimate of the conditional AIC under generalized linear mixed models and proportional hazard mixed models but did not provide exact calculations outside the normal case. They also feature a bootstrap method which has been shown to perform well in estimating the conditional AIC in finite samples with independent and identically distributed

data,

$$cAIC_b = -2l(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2\rho_b \tag{6.13}$$

where ρ_b is the bootstrap estimate of the correction factor

$$\rho_b = E_b \left[l(\mathbf{y}_b|\hat{\boldsymbol{\beta}}_b, \hat{\boldsymbol{\gamma}}_b) - l(\mathbf{y}|\hat{\boldsymbol{\beta}}_b, \hat{\boldsymbol{\gamma}}_b) \right] \tag{6.14}$$

where the subscript b denotes a bootstrap estimate: E_b is obtained by averaging over all the bootstrap datasets, $\mathbf{y}_b$ is the bootstrap response vector obtained by first sampling from the clusters and then sampling within each cluster, $\mathbf{y}$ is the response vector from the sampled clusters, lastly $\hat{\boldsymbol{\beta}}_b$ is the MLE of the fixed coefficient vector for the bootstrapped data, and lastly $\hat{\boldsymbol{\gamma}}$ is an empirical Bayes estimator of the random effect coefficients obtained by fitting the model to $\mathbf{y}_b$. This nonparametric bootstrapping method does not perform well when fitting datasets with small cluster size, e.g. $n_i < 10$ according to Donohue et al. (2011), especially ones containing clusters of one observation.

Donahue et al. (2011) proves the asymptotic unbiasness of cAIC when both the number of clusters and cluster sizes goes to infinity and their work is based on asymptotic normality of the fixed and random coefficient estimators. However there seems to be a lack of a more general approach to obtain an unbiased estimator for generalized linear mixed models. Donahue points to the need of extending the method due to its versatility and wide range of implementations. We encountered convergence problems of the DSDD and glmer algorithm when trying to apply Donahue's bootstrap cAIC method on the datasets discussed in here which can be due to flat gradient surfaces when the same clusters are picked multiple times in each bootstrap iteration.

Lian (2012) proposed an unbiased estimator of cAI for Poisson regression with random effects based on finite sample calculations where he considers the model in (3.5) with the random effects $\boldsymbol{\gamma}$ following a zero mean Gaussian distribution with unknown covariance matrix The unbiased estimator cAIC is

$$cAIC = -2l(\mathbf{y}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) + 2M \tag{6.15}$$

where $l(\mathbf{y}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ is the marginal log-likelihood of $\mathbf{y}$ given any reasonable estimates in the literature, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$, for the random and fixed coefficients respectively, while M is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{y_{ij}\log\left[\hat{\lambda}_{ij}(\mathbf{y})\right] - y_{ij}\log\left[\hat{\lambda}_{ij}(\mathbf{y}^{y_{ij}-1})\right]\right\} \tag{6.16}$$

where $\hat{\lambda}_{ij}(\mathbf{y})$ is the estimated value of λ_{ij} based on the data $\mathbf{y}$, $\mathbf{y}^{y_{ij}-1}$ is the same as $\mathbf{y}$ except that the ij^{th} component is replaced by $\mathbf{y}_{ij} - 1$. Note that when $y_{ij} = 0$, $y_{ij}\log(\hat{\lambda}_{ij}(\mathbf{y}))$ is 0 by convention. Lian (2012) notes that his derivation of the estimator cAIC can be generalized to other contexts such as hierarchical or crossed designs, and it is not sensitive to the assumption of normality of the random effects or the choice of estimator of the fixed or random coefficients.

We applied Lian's method to the second simulated Poisson dataset presented in Section 3.2.1. The fitted value $\hat{\lambda}_{ij}(\mathbf{y})$ is calculated by

$$\hat{\lambda}_{ij}(\mathbf{y}) = \exp\left(\hat{a}_i + \hat{b}_i x_{b_{ij}} + \hat{\beta} z_{ij}\right) \tag{6.17}$$

where $(\hat{a}_i, \hat{b}_i)$ are the posterior modes of the random effects for cluster i and $\hat{\beta}$ is the MLE of the fixed coefficient. Similarly, we calculate $\hat{\lambda}_{ij}(\mathbf{y}^{y_{ij}-1})$ by fitting $\mathbf{y}^{y_{ij}-1}$. The

resulting cAIC was 3654.01. In the glmer case we obtain the fitted values $\hat{\lambda}_{ij}(\mathbf{y})$ by

$$\hat{\lambda}_{ij}(\mathbf{y}) = \hat{B}_a + \hat{a}_i + (\hat{b}_i + \hat{B}_b)x_{b_{ij}} + \hat{\beta}z_{ij} \tag{6.18}$$

where $(\hat{a}_i, \hat{b}_i)$ are posterior modes of the random effects for cluster i and $(\hat{B}_a, \hat{B}_b, \hat{\beta})$ are the MLE's of the fixed coefficients, the resulting cAIC was 1427.45.

6.2 Cross-validation and predictive based goodness-of-fit methods

In this section we present goodness-of-fit methods that focus on predictive assessments and cross-validation. The conditional predictive ordinate (CPO) and numerical posterior predictive check methods are techniques adopted mainly in Bayesian frameworks, but can be used as alternatives to some classical methods that are sensitive to distributional assumptions.

6.2.1 Conditional Predictive Ordinate

Given $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the conditional predictive ordinate for observation i is defined as (Geisser and Eddy, 1979),

$$CPO_i = f(y_i|y_{[i]}) \tag{6.19}$$

where $y_{[i]}$ are the data omitting the i^{th} observation and $f(.)$ is the assumed density of the true observations.

The CPO is a Bayesian approach related to classical ideas for looking at outliers and can also be used as a model comparison tool (Neelon et al. 2009). The marginal

distribution of the data, $f(\mathbf{y})$, is treated as a prior predictive density which is then approximated by a cross-validation predictive density obtained when holding out a set of the data. For our purpose, we modify the CPO technique to suit the clustered data setting by holding out one cluster at a time, therefore the CPO for cluster i becomes

$$CPO_i = f(\mathbf{y}_i|\mathbf{y}_{[i]}, \mathbf{x}_i, \mathbf{z}_i) \tag{6.20}$$

where $\mathbf{y}_i$ is the observed response in the i^{th} cluster, while $\mathbf{x}_i$ and $\mathbf{z}_i$ are the random and fixed covariate vectors of cluster i. $\mathbf{y}_{[i]}$ denote the entire response vector omitting the response of the i^{th} cluster.

For a nonparametric mixed model we apply (6.20) by fitting the model without the i^{th} cluster,

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{y}_{[i]}) &= \sum_{k=1}^{K} f(\mathbf{y}_i|\mathbf{y}_{[i]}; \hat{\boldsymbol{\beta}}_{[i]}, \hat{\boldsymbol{\gamma}}_{[ik]}, \mathbf{x}_i, \mathbf{z}_i)\hat{\pi}_{[k]} \\ &= \sum_{k=1}^{K} \left[\Pi_{j=1}^{n_i} f(y_{ij}|\mathbf{y}_{[i]}; \hat{\boldsymbol{\beta}}_{[i]}, \hat{\boldsymbol{\gamma}}_{[ik]}, \mathbf{x}_i, \mathbf{z}_i)\hat{\pi}_{[k]}\right] \end{aligned} \tag{6.21}$$

where $\hat{\boldsymbol{\beta}}_{[i]}$, $\left[(\hat{\boldsymbol{\gamma}}_{[1]}, \hat{\boldsymbol{\pi}}_{[1]}), \ldots, (\hat{\boldsymbol{\gamma}}_{[K]}, \hat{\boldsymbol{\pi}}_{[K]})\right]$ are estimates of $\boldsymbol{\beta}$ and elements of $\hat{G}$ resulting from fitting a model without the i^{th} cluster. The subscript k indexes the mass point and its corresponding probability in the estimated mixing distribution $\hat{G}$.

The implementation of the CPO method is straight forward, and has the advantage of being easily adaptable to many models. However its implementation in practice can take longer than other methods in relatively large datasets.

We applied the CPO index method to compare the two estimated models resulting from fitting the simulated datasets in Section 3.1.1 and 3.2.1 with the DSDD and glmer routines. In a glmer Poisson mixed model, the random effects have a

Sample	DSDD	glmer
Bernoulli Sample 1	-720.9	-4021.1
Bernoulli Sample 2	-649.9	-4008.1
Poisson Sample 1	-1035.9	-4454.9
Poisson Sample2	-1032.1	-3535.9

Table 6.5: Comparing LPML measures for the DSDD and glmer fitted models of the Bernoulli and Poisson simulation in Chapter 3

multivariate normal distribution, $(a_{glmer}, b_{glmer}) \sim MVN(\mathbf{0}, \Sigma_{(a,b)})$ where $\Sigma_{(a,b)}$ is the estimated variance-covariance matrix. To avoid the computational challenges in (6.20) we discretize the normal distribution by simulating n_{glmer} pairs $(a^*_{glmer}, b^*_{glmer})$. The discretization process is then used to calculate the CPO index for each cluster following (6.21) by considering each pair to be a mass point with a corresponding probability $\pi = 1/n_{glmer}$. Figure 6.1 depicts the indices the CPO's of the Sample 2 simulation in Section 3.2.1 on the log scale where $n_{glmer} = 10000$ is used for the glmer discretization technique. Note that a higher CPO index for the i^{th} cluster means that its observed data is less discordant to the estimated model. For the DSDD algorithm the CPO test shows a group of clusters (60^{th} and above) relatively below the $\log(CPO)$ mean score of -10.32, however in the glmer case the same group of clusters have $\log(CPO)$ relatively above -35.36, the $\log(CPO)$ mean score of the glmer model.

A natural summary statistic of the CPO is the logarithm of the Pseudo-marginal likelihood (LPML), $LPML = \sum_{i=1}^{n} \log(CPO_i)$. Larger values of the LPML mean a better fitting model (Ibrahim et al., 2001b, p: 589). Table 6.5 compares the LPML for the DSDD and glmer algorithms of the simulated samples in Chapter 3. Moreover, results of the DSDD, glmer and GLM models for the NBA dataset analysed in Section 3.3 yield $LPML_{DSDD} = -732$, $LPML_{glmer} = -4317$ and $LPML_{GLM} = -291$, which

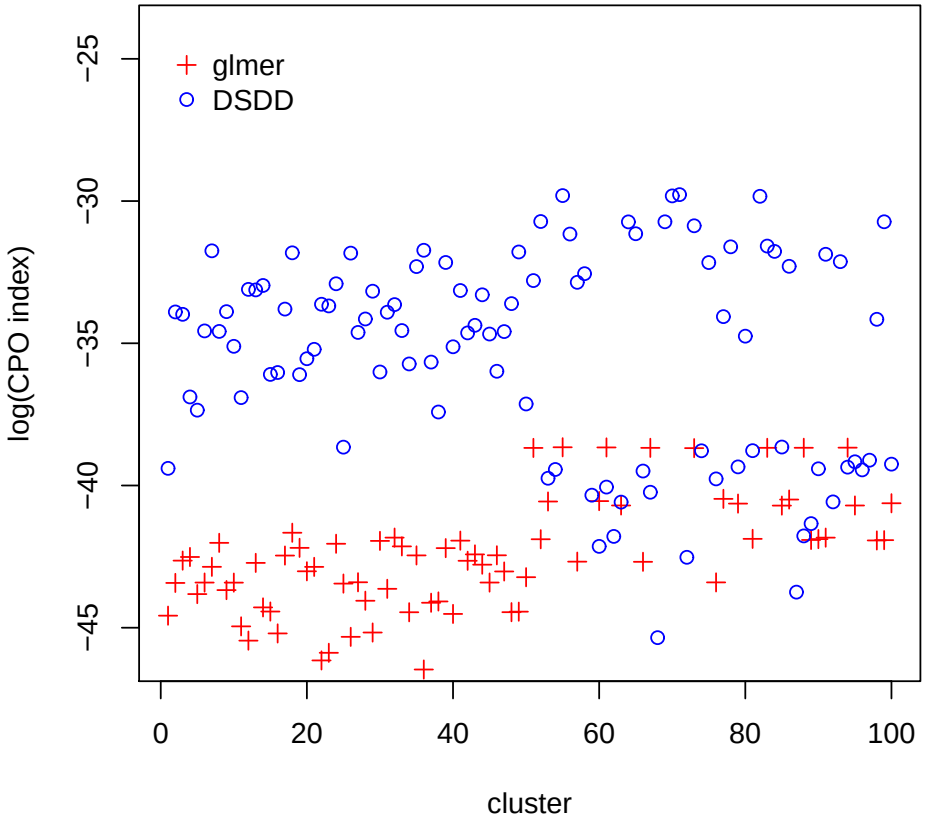

Figure 6.1: CPO index comparison between glmer and DSDD fitted models of the Poisson Sample 2 simulated dataset in Section 3.2.1.

suggests that the simple GLM model might be a better alternative to the mixture models. However, The CPO results suggest that DSDD models perform better than their glmer counterparts.

6.2.2 Numerical posterior predictive checks

Rubin (1981) formulated and gave a Bayesian definition of the posterior predictive model checking (PPMC) technique. Gelman et al. (1996) extended the approach and introduced a discrepancy measure to compare the observed data with the proposed model. Hence, the measure should show an extreme value if the observed data is in conflict with replicated data. The PPMC approach emphasizes predicted data rather than the parameters in the model, Carlin and Louis (2000, p: 223-225) note that by working in a purely predictive space an intuitive penalty emerges without the asymptotics that are needed when calculating the DIC or AIC as discussed in previous sections. The predictive distribution is constructed in a Bayesian fashion

$$f(\mathbf{y}^{rep}|\mathbf{y}^{obs}) = \int f(\mathbf{y}^{rep}|\boldsymbol{\gamma})G(\boldsymbol{\gamma}|\mathbf{y}^{obs})d\boldsymbol{\gamma} \tag{6.22}$$

where $\mathbf{y}^{rep}$ is a replicate vector assumed to have the same distribution as the observed data vector $\mathbf{y}^{obs}$ and $G(\boldsymbol{\gamma}|\mathbf{y}^{obs})$ is the posterior distribution of $\boldsymbol{\gamma}$. The general idea is to choose the model that minimizes

$$E\left[D(\mathbf{y}^{rep}, \mathbf{y}^{obs})|\mathbf{y}^{obs}\right] \tag{6.23}$$

where $D(\mathbf{y}^{rep}, \mathbf{y}^{obs})$ is a discrepancy function. As an example, for Gaussian likelihoods Laud and Ibrahim (1995, p: 250) suggest using the dicrepancy $D(\mathbf{y}^{rep}, \mathbf{y}^{obs}) = (\mathbf{y}^{rep} - \mathbf{y}^{obs})(\mathbf{y}^{rep} - \mathbf{y}^{obs})^T$. While Carlin and Lewis (2000, p: 224) suggest using the

deviance criterion for some non-Gaussian generalized linear mixed models to measure a discrepancy function between $\mathbf{y}$ and $\mathbf{y}^{rep}$. For instance in the Poisson case the discrepancy D is

$$D(\mathbf{y}, \mathbf{y}^{rep}) = 2 \sum_{l} \left[\mathbf{y} \log(\mathbf{y}/\mathbf{y}^{rep,l}) - (\mathbf{y} - \mathbf{y}^{rep,l}) \right] \tag{6.24}$$

where l indexes the replicated dataset. To avoid computational difficulties at extreme values the authors suggest replacing (6.24) with

$$D(\mathbf{y}, \mathbf{y}^{rep}) = 2 \sum_{l} \left[(\mathbf{y} + 0.5) \log \left(\frac{\mathbf{y} + 0.5}{\mathbf{y}^{rep,l} + 0.5} \right) - (\mathbf{y} - \mathbf{y}^{rep,l}) \right]. \tag{6.25}$$

Similarly, For Bernoulli models we express the discrepancy as

$$D(\mathbf{y}, \mathbf{y}^{rep}) = 2 \sum_{l} \left[(\mathbf{y} + \epsilon) \log \left(\frac{\mathbf{y} + \epsilon}{\mathbf{y}^{rep,l} + \epsilon} \right) + (1 - \mathbf{y} + \epsilon) \log \left(\frac{1 - \mathbf{y} + \epsilon}{1 - \mathbf{y}^{rep,l} + \epsilon} \right) \right] \tag{6.26}$$

for some $0 < \epsilon \leq 0.01$ added to avoid numerical complications.

The method presented above can be applied to compare mixture models. To illustrate the details of this application we use the Poisson simulations presented in Section 3.2.1. The l^{th} replication, $y_{ij}^{rep,l}$, is generated from a Poisson distribution with mean $\hat{\lambda}_{ij} = \exp(\hat{\alpha}_i + \hat{b}_i x b_{ij} + \hat{w}_i x w_{ij} + \hat{\beta} z_{ij})$ where $\hat{\beta}$ is the estimated fixed coefficient and $(\hat{\alpha}_i, \hat{b}_i, \hat{w}_i) = \hat{\boldsymbol{\gamma}}_i$ is the posterior mode of $\boldsymbol{\gamma}$ for cluster i, where $P(\hat{\boldsymbol{\gamma}}_i | \mathbf{y}_i) = argmax_{\boldsymbol{\gamma}_k} P(\boldsymbol{\gamma}_{ik}, \mathbf{y}_i)/P(\mathbf{y}_i)$ and k is the index for the estimated mass points of the random effect component. See Section 5.1.2 for a detailed description on obtaining the posterior modes of the normally distributed random effects estimated by the glmer routine.

The discrepancies in (6.25) and (6.26) for the Poisson and Bernoulli simulations

Sample	average$(D_{pois})^{DSDD}$, s.s.e.$(D_{pois})^{DSDD}$	average$(D_{pois})^{glmer}$, s.s.e.$(D_{pois})^{glmer}$
Poisson Sample 1	724.464, 43.73	755.582, 46.34
Poisson Sample 2	676.379, 48.85	5432.338, 124.24
Sample	average $(D_{bin})^{DSDD}$, s.s.e. $(D_{bin})^{DSDD}$	average$(D_{bin})^{glmer}$, s.s.e. $(D_{bin})^{glmer}$
Bernoulli Sample 1	1731.98, 86.68	1900.230, 97.08
Bernoulli Sample 2	1083.94, 74.68	1138.44, 69.52
NBA case study	1807.98, 84.82	1732.20, 93.86

Table 6.6: Discrepancy summary of the Poisson, Bernoulli and the NBA case study datasets presented in Sections 3.1.1, 3.2.1 and 3.3: The average and s.s.e. denote the sample averages and sample standard error of D_{pois} and D_{bin} discrepancies of length L each.

become respectively,

$$D_{pois}(y, y^{rep}) = 2\sum\left[(y_{ij}+0.5)\log\left(\frac{y_{ij}+0.5}{y_{ij}^{rep,l}+0.5}\right)-(y_{ij}-y_{ij}^{rep,l})\right] \tag{6.27}$$

and

$$D_{bin}(y, y^{rep}) = 2\sum\left[(y_{ij}+\epsilon)\log\left(\frac{y_{ij}+\epsilon}{y_{ij}^{rep,l}+\epsilon}\right)-(1-y_{ij}+\epsilon)\log\left(\frac{1-y_{ij}+\epsilon}{1-y_{ij}^{rep,l}+\epsilon}\right)\right]. \tag{6.28}$$

Table 6.6 summarizes the discrepancies of the simulations and NBA case study in Chapter 3.

Posterior predictive checking can be very useful to compare hurdle models as well. The replicated responses, $\mathbf{y}^{rep}$, are generated using the following process:

- Calculate the mode of the random effects for cluster i, $\hat{\boldsymbol{\gamma}}_i$, as described in Section

5.1.2 where $\hat{\boldsymbol{\gamma}}_i = \left\{(\hat{\boldsymbol{\gamma}}_{0_i}, \hat{\boldsymbol{\gamma}}_{1_i}) = \left[(\hat{\alpha}_{0_i}, \hat{\alpha}_{1_i}), (\hat{b}_{0_i}, \hat{b}_{1_i}), (\hat{w}_{0_i}, \hat{w}_{1_i})\right]\right\}$ such that

$$\begin{aligned} P(\hat{\boldsymbol{\gamma}}_i|\mathbf{y}_i) &= argmax_{\boldsymbol{\gamma}_k} P(\boldsymbol{\gamma}_k, \mathbf{y}_i)/P(\mathbf{y}_i) \\ &\propto argmax_{\boldsymbol{\gamma}_k} P(\mathbf{y}_i|\boldsymbol{\gamma}_k)\hat{\pi}_k \end{aligned} \tag{6.29}$$

where $P(\mathbf{y}_i|\hat{\boldsymbol{\gamma}}_k) = L(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\gamma}}_k)$ is the likelihood in (4.4) evaluated at $\hat{\boldsymbol{\gamma}}_k$.

- Generate the random variables $y_{ij}^{ZIP} \sim I_{0_{ij}} * Y_{count_{ij}}$ where Y_{count} is generated according to a Poisson distribution truncated at zero

$$Y_{count_{ij}} \sim TruncatedPoisson\left[\exp\left(\hat{\alpha}_{1_i} + \hat{b}_{1_i} xb_{ij} + \hat{w}_{1_i} xw_{ij} + \hat{\beta}_1 z_{ij}\right)\right]$$

and

$$I_{0_{ij}} \sim \text{Bernoulli}\left[\frac{\exp(\hat{\alpha}_{0_i} + \hat{b}_{0_i} xb_{ij} + \hat{w}_{0_i} xw_{ij} + \hat{\beta}_0 z_{ij})}{1 + \exp(\hat{\alpha}_{0_i} + \hat{b}_{0_i} xb_{ij} + \hat{w}_{0_i} xw_{ij} + \hat{\beta}_0 z_{ij})}\right].$$

Due to the over-abundance of zeros within clusters, Neelon et al. (2010) suggests using the sample skewness and an overdispersion index as discrepancy measures for zero-inflated data to compare replicated and observed response variables. The comparison of the discrepancies can be done through various measures such as p-values defined in a Bayesian framework as $P[D(\mathbf{y}^{rep}) \geq D(\mathbf{y})|A, \boldsymbol{\gamma}]$ where A is the assumed model and $\boldsymbol{\gamma}$ is the set of unknown parameter with distribution G. The Bayesian p-value is the probability that the discrepancy measure $D(\mathbf{y}^{rep})$ is more extreme than the observed discrepancy. Generally p-values between 0.2 and 0.8 represent an adequate fit of the model (Neelon et al., 2010).

For the pharmaceutical case study in Section 4.3.2, the posterior predictive check method is used to compare the DSDD and EM algorithms ' results. Based on 5000

replicated samples we used two measures for the discrepancy: the overdispersion index defined as $D_1 = var(\mathbf{y}^{rep})/E(\mathbf{y}^{rep})$ and the adjusted Fisher-Pearson standardized moment coefficient $D_2 = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left(\frac{y_i - \bar{y}}{s}\right)^3$ where s is the sample standard deviation and N is the sample size. The resulting p-values for the DSDD algorithm are 0.76 and 0.99 for D_1 and D_2, and the EM algorithm's p-values are 0.51 and 0.97 corresponding to D_1 and D_2 respectively.

For zero-inflated data, we find that the sensitivity to the choice of a discrepancy measure is an obvious limitation of this approach. Gelman et al. (p: 176, 2004) notes that finding an extreme p-value suggests improvements of the model or places to check the data. They also note that the posterior predictive checking technique assesses whether the data could have arisen by chance under the model assumption and not whether the data come from the assumed model, therefore the authors suggest that these methods should only serve as preliminary goodness-of-fit test.

6.3 Coefficient of determination for generalized linear mixed models

The coefficient of determination is well known in the ordinary least square regression and is defined in the univariate case as (Rao, 1973)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{6.30}$$

where $\hat{y}_i$ is the i^{th} fitted value and $\bar{y}$ is the grand mean of the observed values.

The R^2 measure is intuitive, independent of any units of measurements, is well bounded by 0 and 1 with 0 representing a complete lack of fit and 1 representing

the perfect fit, and is applicable to any type of model regardless of the distributional properties. It is interpreted as the proportion of total variation in y (about its sample mean) that is explained or accounted for by the fitted model. Despite some of its limitations as described by Kvalseth (1985), it remains one of the most popular goodness-of-fit measure in modelling. The R^2 satistifies all of the conditions outlined by Kvalseth (1985) to be considered a reasonable goodness-of-fit measure.

Due to its versatility a number of adjusted R^2 have emerged depending on model settings. McFadden (1974) suggested a log-likelihood ratio for logistic regressions with several advantages such as its direct relation to the valid test statistics for the significance of all slope coefficients and the compliance of its derivation with respect to the test statistic with corresponding derivation in linear regression. Cox and Snell (1989) generalized the coefficient of determination to a more general linear model, which was later modified by Nagelkerke (1991) to assure that its range of possible values extended to 1 and its properties and interpretation complied with the classical R^2 measure along with other important merits that we will not discuss here. Cameron and Windmeijer (1997) proposed an R^2 measure based on the Kullback-Leibler divergence for exponential families.

Generally in mixed models, there is no single definition of R^2 because the concept of total variation depends on the criterion used and the method of estimation (Nagelkerke, 1991). Vonesh et al. (1996) derived a concordance coefficient closely related to the R^2 measure

$$R_C = 1 - \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)' (\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n} (\mathbf{y}_i - \bar{y}\mathbf{1}_i)' (\mathbf{y}_i - \bar{y}\mathbf{1}_i) + \sum_{i=1}^{n} \left(\hat{\mathbf{y}}_i - \bar{\hat{y}}\mathbf{1}_i\right)' (\mathbf{y}_i - \bar{y}\mathbf{1}_i) + N(\bar{y} - \bar{\hat{y}})^2} \tag{6.31}$$

where n is the number of clusters, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the vectors of observed and fitted

values for the i^{th} cluster, $\bar{y}$ and $\bar{\hat{y}}$ are the grand averages of the observed and fitted values, N is the total number of observations, $\mathbf{1}_i$ is an n_i X 1 vector of 1's. R_c describes the level of agreement between the observed and fitted values, it ranges between -1 and 1 where a value of 1 means a perfect positive agreement between the fitted and observed values, a value of -1 denotes a perfect negative agreement (i.e. the vectors $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are orthogonal to each other), and a value of 0 corresponds to no agreement between the fitted and observed values. Note that any $R_C \leq 0$ corresponds to a lack of fit (Vonesh et al. 1996).

Liao (2003) improved the concordance coefficient, R_c, to assess the goodness-of-fit of nonlinear mixed effect models. The new concordance coefficient is defined as

$$cc = \rho \frac{4S_1S_2 - \rho(S_1^2 + S_2^2)}{(2-\rho)(S_1^2 + S_2^2) + (\bar{y} - \bar{\hat{y}})^2} \tag{6.32}$$

where S_1^2 and S_2^2 are the variances of the observed and fitted response vectors respectively and ρ is the correlation between the two vectors.

In addition to the two concordance measures R_C and cc one can calculate a modelling efficiency or fit index also known as the conditional R^2, denoted by R^2_{cond} here, which corresponds to the traditional R^2 measure in (6.30) except that the fitted response vector, $\hat{\mathbf{y}}$, takes into account the random effect predictions (Huang et al., 2009). For clustered data, the R^2_{cond} is defined as

$$R^2_{cond} = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2} \tag{6.33}$$

The three different coefficients described here are displayed in Table 6.7 for the simulated datasets and NBA case study in Chapter 3. The fitted values vector, $\hat{\mathbf{y}}$ is obtained by using the modes to predict the random effect vector of the i^{th} cluster as

		Coefficient		
		R_C	cc	R^2_{cond}
Bernoulli 1	DSDD	0.44	0.40	0.28
	glmer	0.42	0.30	0.29
Bernoulli 2	DSDD	0.70	0.67	0.54
	glmer	0.74	0.69	0.61
Poisson 1	DSDD	0.91	0.91	0.84
	glmer	0.92	0.92	0.85
Poisson 2	DSDD	0.91	0.90	0.83
	glmer	0.92	0.91	0.85
NBA case study	DSDD	0.47	0.42	0.31
	glmer	0.45	0.38	0.30

Table 6.7: Coefficients of determination for the simulated datasets and NBA case study of Chapter 3

described in Section 5.1.2. All three coefficients indicate that the DSDD model fitted values agree better with the observed vector $\mathbf{y}$ for the NBA case study than the glmer fitted values. We can also notice that all three coefficients indicate that the Poisson models are adequate fits for both methods.

The use of the coefficient of determination have numerous advantages over other goodness-of-fit methods, mainly because its free of any distributional assumptions and one does not need to specify a likelihood in order to evaluate how well the current model fits the data. However Vonesh (1996) points to the fact that the coefficients of determination are not well suited for data highly discrete in nature such as the binary data in this dissertation.

Chapter 7

Discussion and topics for further research

We introduce a nonparametric estimation method for generalized linear mixed models with random intercept and slopes in a clustered setting. The R package lme4 addresses this type of model where the random effects are normally distributed. The use of nonparametric mixtures, however, can have many advantages when the model's random effects are the focus of the study.

Fixed coefficient estimations are often based on maximum likelihood theory, which assumes that the underlying probability model is correctly specified. When the emphasis of the estimation is placed on the fixed effects and the random effects in the model are treated as nuissance parameters, the latter are traditionally assumed to be normally distributed. Although recent research shows that the choice of random effects distribution seem to have little influence on the maximum likelihood estimator (see McCulloch, 2010 and Agresti, 2004), the predictive power of the model can be affected. The Poisson and Bernoulli simulations shown in Chapter 3 confirm this conclusion as the estimation of the fixed coefficients are consistent between the parametric

and semi-parametric models despite a slight loss of efficiency in the nonparametric case. Nevertheless, the nonparametric results are more reliable in terms of predicting the random effects for each cluster (see the results in Section 5.1.2). Consequently, such misspecification will affect predictions of the response variable as evidenced in Figures 3.2 and 3.3.

7.1 Algorithm performance

The DSDD expands the CNM algorithm in order to estimate the nonparametric mixing distribution, G, in a Poisson or Bernoulli GLMM with random intercepts and slopes. Previous algorithms, such as the CNM and ISDM, use gradient based methods to find local maxima of the directional derivative at each step. The direct search method can be used as an alternative. Their proofs of convergence, shown in Polak (1971) and Torczon (1997), guarantee that the algorithm can not jam at any point hence avoiding any numerical complications. Moreover, the simplicity of direct search methods make them easily applicable to a variety of models, such as zero-inflated models, without the need for first and second order derivatives. The speed of convergence of the direct search method was not compared to other gradient based methods, but our experience is that the DSDD performs reasonably well and is quicker when there is a clear separation between clusters.

The DSDD does well in estimating fixed effects, as shown in our simulation studies. It is worth noting that the estimated mixtures in the simulation studies include extra support points compared to the true mixtures, although this is compensated by the fact that their corresponding weights are relatively low.

For future work, it is beneficial to assess whether there is a significant change in the performance of the DSDD if the direct search method is replaced with a traditional

gradient based optimization method such as the DFP and BFGS formulas as suggested in Wang (2010).

7.2 Goodness-of-fit Summary

Graphical tools such as the residual plots presented in Chapter 5 are one way to visualize how well the model fits the data. Nevertheless, model comparison and goodness-of-fit tests that are not sensitive to distributional assumptions are needed for comparing mixture models with different mixing distributions. The literature on the topic is large and diverse, and we explore a few of the available methods to assess and compare the models discussed throughout the dissertation.

We start with penalized-likelihood based methods; DIC and cAIC. Our results showed a large discrepancy between the nonparametric mixture models estimated by the DSDD and the glmer methods which highlights their instability. The cAIC bootstrap calculations provided by Donohue (2011) were computationally challenging especially for Bernoulli mixed models, which is mainly due to small cluster sizes. Furthermore, DIC results showed that the measure is not invariant to parametrization and is sensitive to the type of estimator used for the random effects as evidenced by the inconsistencies in the penalty term, pD, between DIC^1, DIC^2 and DIC^p. In terms of model comparison we found that the DIC tends to favour models with less random covariates in the case of nonparametric mixtures.

Simulation based Bayesian cross-validation methods such as the conditional predictive ordinates and numerical posterior predictive checks are extensions of classical model checking with the advantage of being relatively simple to apply and interpret in the various models studied here regardless of the mixing distribution used. Pettit (1990) notes that the CPO method should only be used as an initial diagnostics step

to give an indication of any surprising observation (or in our case, observations within each cluster), if the CPO gives a similar value for all observations then follow-up examinations will unlikely suggest any contaminants in the model.

The discrepancy used in numerical posterior predictive checks vary depending on the aspect of the data we wish to investigate. Computationally, both methods have relatively long execution times especially when the complexity and sample size of the model increase.

The choice of measure between the penalized likelihoods criteria and predictive checks depends on the focus of the study as noted in Spiegelhalter et al. (2002). Penalized criteria are preferred when the aim is to compare models conditional on the parameters, while posterior predictive checks are preferable when the aim is to compare the predictive ability of two models marginally.

Finally, the coefficient of determination, R^2, is a classical goodness-of-fit test in least square regression. However, R^2 for mixed models can be defined in different ways due to the presence of several variance components. We investigated three of its variations, and the results indicate a relative agreement between the DSDD and glmer estimated models for all the R^2 coefficients.

Appendix A

NBA Dataset summary

Table A.1: NBA data, summary statistics and estimates

Team	Proportion of seasons in playoffs	Cluster	$\hat{a}$	$\hat{b}$	Mean standardized payroll	Mean age
Atlanta	0.50	3	0.49	0.31	-0.55	27.1
Boston	0.50	4	0.54	3.00	-0.08	26.6
Charlotte	0.50	3	0.49	0.31	-0.92	27.5
Chicago	0.56	3	0.49	0.31	-0.21	27.0
Cleveland	0.50	4	0.54	3.00	0.04	27.0
Dallas	0.63	4	0.54	3.00	0.49	27.3
Denver	0.50	3	0.49	0.31	-0.26	26.9
Detroit	0.75	6	2.07	2.19	-0.40	28.1
Golden State	0.06	1	-1.35	3.00	-0.35	28.5
Houston	0.63	3	0.49	0.31	-0.06	27.6
Indiana	0.69	3	0.49	0.31	0.38	26.2
Los Angeles Clippers	0.13	4	0.54	3.00	-0.95	27.2
Los Angeles Lakers	0.94	6	2.07	2.19	0.73	28.2
Miami	0.75	3	0.49	0.31	0.17	27.5
Milwaukee	0.44	4	0.54	3.00	-0.32	27.4
Minnesota	0.50	3	0.49	0.31	-0.24	27.7
New Jersey	0.44	2	-0.12	-0.17	-0.07	28.5
New York	0.50	2	-0.12	-0.17	2.14	27.9
Orlando	0.69	3	0.49	0.31	0.33	26.6
Philadelphia	0.50	3	0.49	0.31	-0.15	28.1
Phoenix	0.81	3	0.49	0.31	0.36	27.2
Portland	0.69	3	0.49	0.31	0.85	27.4
Sacramento	0.56	3	0.49	0.31	-0.10	29.9
San Antonio	0.94	3	0.49	0.31	0.21	29.9
Seattle/Oklahoma	0.50	3	0.49	0.31	-0.42	27.4
Utah	0.81	3	0.49	0.31	-0.35	27.7
Washington	0.31	2	-0.12	-0.17	-0.01	26.9
New Orleans	0.50	2	-0.12	-0.17	-0.45	27.7
Toronto	0.33	2	-0.12	-0.17	-0.54	27.5
Memphis	0.20	4	0.54	3.00	-0.46	26.6

Bibliography

[1] Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. IN *Proceedings of the* 2^{nd} *International Symposium on Information Theory* (eds B.N. Petrovand F. Csaki), pp. 267-281. Budapest: Akademiai Kiado.

[2] Anderson, D.A., and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society B*, **47**, 203-210.

[3] Arulampalam, W., and Booth, A. (1997). Who gets over the training hurdle? a study of the training experiences of young men and women in Britain. *Journal of Population Economics*, **10**, 197-217.

[4] Atwood, C.L. (1976). Convergent design sequences, for sufficiently regular optimality criteria. *The Annals of Statistics*, **4**, 1124-1138.

[5] Bender, Patricia. (2011). Patricia's Various Basketball Stuff. Web log post. Internet Service Provider Broadband DSL Dial Access Hosting. Web. 02 Aug. 2011. http://www.eskimo.com/∼pbender/index.html

[6] Bates, D.M., Maechler, M., and Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39. http://CRAN.R-project.org/package=lme4.

[7] Bates, D.M. (2011). Computational methods for mixed models. *University of Wisconsin-Madison*.

[8] Bates, D.M., Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

[9] Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

[10] Brillinger D.R., Preisler, H.K., Benoit, J.W. (2006). Probabilistic risk assessment for wildfire. *Environmetrics*, **17**, 623-633.

[11] Broström, G., Holmberg, H. (2011) glmmML: Generalized linear models with clustering. R package version 0.81-8. http://CRAN.R-project.org/package=glmmML.

[12] Böhning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning Inference*, **11**, 57-69.

[13] Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, **13**, 257-265.

[14] Cameron, A.C., and Windmeijer, F.A.G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, **77**, 329-342.

[15] Carlin, B.P., and Louis, T.A. (2000). *Bayes And Empirical Bayes Methods For Data Analysis: Second Edition*. Chapman and Hall/CRC.

[16] Céa, J. (1971). Optimisation: Théorie et algorithmes. Dunod, Paris.

[17] Celeux, G., Forbes, F., Robert, C., Titterington, M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651-674.

[18] Chambers, R., Chandra, H. (2011). A semiparametric block bootstrap for clustered data. *Centre for Statistical and Survey Methodology*, University of Wollongong. Working paper **30**, 01-11.

[19] Claeskens, G., Hart, J.D. (2009). *Goodness-of-fit tests in mixed models*. https://lirias.kuleuven.be/bitstream/123456789/220817/3/KBI (May 12,2009).

[20] Crouch, E. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int f(t)e^{-t^2}dt$: Application to Logistic-normal models. *Journal of the American Statistical Association*, **85**, 464-469.

[21] Deely, J.J., Lindley, D.V. (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association*, **76** (376), 833-841.

[22] Cox, D.R., and Snell, E.J. (1989). *The Analysis of Binary Data: Second Edition*. Chapman and Hall, London.

[23] Donohue, M.C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika Trust*, **98**, 685-700.

[24] Fedorov V.V. (1972). *Theory of Optimal Experiments (English translation)*. New York: Academic Press.

[25] Follmann, D.A., and Lambert, D.(1991). Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, **27**, 375-381.

[26] Geisser, S., Eddy, W. (1979). A predictive approach to model selection. *Journal of American Statistical Association*, **74**, 153-160.

[27] Gelman, A., Carlin, J.B., Stern, H., and Rubin, D.B. (2004). *Bayesian Data Analysis: Second Edition* Chapman and Hall/CRC.

[28] Gelman, A., Meng, X.L. and Stern, H. (1996). Posterior predicitve assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733-807.

[29] Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, **57**, 369-375.

[30] Hall, D.B. (2000). Zero-inflated Poisson and Binomial regression with random effects: A case study. *Biometrics*, **56**, 1030-1039.

[31] Haskell K.H., and Hanson R.J. (1981). Algorithm for linear least squares problems with equality and nonnegativity constraints. *Math. Programming*, **21**, 98-118.

[32] Harrell, F.E. (2001). *Regression Modeling Strategies: First Edition*. Springer Series in Statistics.

[33] Henderson, C.R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, **21**, 309-310.

[34] Henderson, C.R. (1984).Applications of linear of linear models in animal breeding. Guelph, Canada: University of Guelph.

[35] Hinde, J. (1982). Compound Poisson regression models. *Lecture Notes in Statistics*, **14**, 109-121.

[36] Hodges, J.S, and Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly parametrized models. *Biometrika*, **88**, 367-379.

[37] Hooke, R., and Jeeves, T.A. (1961). Direct search solution of numerical and statistical problems. *Journal of the Association of Computing Machinery*, **8**, 212-229.

[38] Huang, S., Meng, S. X., and Yang, Y. (2009). Assessing the goodness of fit of forest models estimated by nonlinear mixed-model methods. *Canadian Journal of Research*, **39**, 2418-2436.

[39] Hwang YT, Wei PF (2006). A novel method for testing normality in a mixed model of a nested classification. *Computational Statistics and Data Analysis*, **51 No. 2**, 1163-1183.

[40] Ibrahim, JG., Chen, M-H., Sinha, D. (2001b). *Bayesian Survival Analysis*. New York: Springer-Verlag.

[41] Jewell, N. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, **10**, 479-484.

[42] Kiefer, J., and Wolfowitz J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Statistics*, **27**, 887-906.

[43] Kvalseth, T.O. (1985). Cautionary note about the R^2. *The American Statistician*, **39** No. 4, 279-282.

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

[45] Lange, N., Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, **17 No. 2**, 624-642.

[46] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

[47] Laud, L.P., and Ibrahim, J.G. (1995). Predictive model selection. *Journals of the Royal Statistical Society B*, **57**, 247-262.

[48] Lesperance, M.L., and Kalbfleisch, J.D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, **87**, 120-126.

[49] Liao, J.J.Z. (2003). An improved concordance correlation coefficient. *Pharmaceutical Statistics*, **2 No. 4**, 253-261.

[50] Lian, H. (2012). A note on conditional Akaike information for Poisson regression with random effects. *Electronic Journal of Statistics*, **6**, 1-9.

[51] Lindsay, B.G. (1983). The Geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, **11**, 86-94.

[52] Lindsay, B.G. (1995). *Mixture models: theory, geometry and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5. U.S.A.

[53] Lindstrom, M. J., and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673-687.

[54] Maritz, J.S (1970). *Empirical Bayes Methods*. London: Chapman and Hall.

[55] McCulloch, C.E. and Neuhaus, J.M. (2010). Prediction of Random Effects in Linear and Generalized linear Models under Model Misspecification. *Biometrics*, no. doi: 10.1111/j.1541-0420.2010.01435.x.

[56] McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, linear, and mixed models: Second Edition*. Wiley Series in Probability and Statistics.

[57] McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior, in Frontiers of Econometrics*, P. Zarembka, ed., Academic Press, New York, 105-142.

[58] Min, Y., Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1-19.

[59] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341-365.

[60] Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, **78 No.3**, 691-692.

[61] Neelon, B.H., O'Malley, A.J., Normand S.L. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, **4**, 421-439.

[62] Pettit, L.I. (1990). The Conditional Predictive Ordinate for the Normal Distribution. *Journals of the Royal Statistical Society*, B, **52 No. 1**, 175-184.

[63] Pearson, K. (1984). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A,* **185**, 71-110.

[64] Polak, E. (1971). *Computational Methods in Optimization: A United Approach*. Academic Press, New York.

[65] Rao, C. R. (1973). *Linear Statistical Inference and its Applications: Second Edition*. Wiley, New York.

[66] Roberts, A.W., Varberg, D.E. (1973). *Convex Functions*. Academic Press, New York and London.

[67] Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6** No. 1, 15-51.

[68] Rubin, 1981. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, **6**, 377-401.

[69] Schwartz, G., (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

[70] Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, **4**, 1200-1209.

[71] Spiegelhalter, D.J., Best, N.G., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

[72] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society B*, **64**, 583-639.

[73] Spiegelhalter, D. J., Carlin, B.P, Best, N. G. Lunn, D. (2003). Winbugs User Manual. Version 1.4. MRC Biostatistics Unit. Cambridge, England.

[74] Stiratelli R., Laird, N., and Ware, J.H. (1984). Random-effects model for serial observations with binary response. *Biometrics*, **40**, 961-971.

[75] Tao, H., Palta, M., Yandell, B.S., and Newton, M.A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics*, **55**, 102-110.

[76] Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, **32**, 244-248.

[77] Torczon V. (1991). On the convergence of the multidirectional search algorithm. *The SIAM Journal on Optimization*, **1**, 123-145.

[78] Torczon V. (1997). On the convergence of pattern search algorithm. *The SIAM Journal on Optimization*, **7** No. 1, 1-25.

[79] Vaida F., and Blanchard S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92** Part 2, 351-370.

[80] Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in random-effects population. *Journal of the American Statistical Association*, **91**, 217-221.

[81] Vonesh, E.F., Chinchilli, V.M., and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, **2**, 572-587.

[82] Wang, Y. (2007). On fast estimation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journals of the Royal Statistical Society B*, **69** Part 2, 185-198.

[83] Wang, Y. (2010). Maximum likelihood computation for fitting semiparametric mixture models. *Statistics and Computing*, **20**, 75-86.

[84] Welsh, A. H., Cunningham, R. B., Donnelly, C. F., and Lindenmayer, D. B. (1996). Modelling abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297-308.

[85] Wu, C.F. (1978a). Some algorithmic aspects of the theory of optimal designs. *The Annals of Statistics*, **6**, 1286-1301.

[86] Wu, C.F. (1978b). Some iterative procedures for generating nonsingular optimal designs. *Communications in Statistics - Theory and Methods*, **7**, 1399-1412.

[87] Wynn, H.P.(1970). The sequential generation of D-optimal experimental design. *The Annals of Mathematical Statistics*, **41**, 1655-1664.