UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Science

Conjugate Generalized Linear Mixed Models with Applications

by

Jarod Yan Liang Lee

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

Doctor of Philosophy

Sydney, Australia

2017

Certificate of Original Authorship

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that this thesis has been written by me. Any help that I have received in my research and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

© Copyright 2017 Jarod Lee

Signature of Student:

Date:

Acknowledgements

The successful completion of this research work is not only the result of my own effort, but also a series of contribution from many others. My deepest gratitude goes to my supervisors Professor Louise Ryan, Professor Peter Green and Professor James Brown for always being encouraging, supportive and inspirational. Their great insight into statistical methodology and applied statistics have profoundly influenced the way I approach statistical research. Without them, I could not have finished my PhD.

I am grateful to the Australian Bureau of Statistics for providing funding that allowed me to present my work at various conferences. Also, thanks to Professor Michael Martin and Professor Alan Welsh who provided advice and encouragement during my undergraduate years at the Australian National University, where I first became interested in statistics.

Last but certainly not least I would like to thank my family and my partner for all of their support, patience and understanding. I consider myself very fortunate to have such a great family.

Jarod Lee
Sydney, Australia, 2017

ABSTRACT

Conjugate Generalized Linear Mixed Models with Applications

by

Jarod Yan Liang Lee

This thesis focuses on the development of conjugate generalized linear mixed models (CGLMMs), which is a computationally efficient modelling framework for longitudinal and multilevel data where the likelihood can be expressed in closed-form. We focus on the scenario where the random effects are mapped uniquely onto the grouping structure and are independent between groups. Compared with conventional inference methods for generalized linear mixed models (GLMMs), CGLMMs allow the parameters to be estimated directly without the need for computational intensive numerical approximation methods. The proposed framework has important implications in terms of distributed computing, privacy preservation in large-scale administrative databases and discrete choice models, which we illustrate using several real data. Altogether, CGLMMs prove to be a credible inference framework and a good alternative to GLMMs, especially when dealing with a large amount of data and/or privacy is of concern.

Contents

Chapter 1

Introduction

1.1 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) extend generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) by including random effects in the linear predictor in addition to the usual fixed effects, hence the term mixed models. Schall (1991) and Breslow and Clayton (1993) popularized GLMMs by proposing approximate inference methods and providing example applications, although these models had a history before that (Harville, 1977; Laird and Ware, 1982; Jennrich and Schluchter, 1986). Since then, there has been an explosion of developments in GLMMs and their applications. In fact, a search of "generalized linear mixed models" on Google Scholar returns 1.69 million results as of 17 August 2017. The field is vast and expanding rapidly as to preclude any attempt at exhaustive coverage, but a selection of major books include Pinheiro and Bates (1978); Rao (1997); Verbeke and Molenberghs (2000); Diggle et al. (2002); Fitzmaurice et al. (2004); Hedeker and Gibbons (2006); Gelman and Hill (2007); Jiang (2007); McCulloch et al. (2008); Wu (2010); Goldstein (2011); Demidenko (2013); and Stroup (2013).

GLMMs have been used in a plethora of applications, ranging from species abundance studies (Fornaroli et al., 2015), forecasting elections (Wang et al., 2015), school performance studies (Marks and Printy, 2003), phylogenetic analysis in bioinformatics (Ronquist and Huelsenbeck, 2003; Stamatakis, 2006) to cancer screening (Morrell et al., 2012), just to name a few. Certain richly parameterized models such as penalized splines, additive, spatial and time-series models can also be re-expressed as mixed models, making them a rich and generic class of models (Hodges, 2013). In general, GLMMs are useful for, but not limited to

- Modelling the dependence among outcome variables inherent in multilevel and longitudinal data, including complex grouping structures such as crossed random effects (Diggle et al., 2002; Fitzmaurice et al., 2004; Gelman and Hill, 2007; Goldstein, 2011),
- Accommodating overdispersion, often encountered in binomial and Poisson distributions (Breslow and Clayton, 1993; Molenberghs et al., 2010),
- Producing shrinkage estimates, for example in small area disease mapping (Chambers and Clark, 2012; Rao and Molina, 2015),
- Accounting for variation among experimental blocks, where the blocks are viewed as a small subset of the larger set of blocks (Montgomery, 2017).

As an example, outcomes of patients within the same hospital are likely to be dependent due to similar risk profiles and a common clinical management practice. GLMMs provide a natural framework for modelling dependencies by allowing for random group specific effects.

1.1.1 Model Formulation

Observation on unit j ($j = 1, 2, \ldots, n_i$) in cluster i ($i = 1, 2, \ldots, I$) consists of a univariate response variable y_{ij}, together with covariate vectors x_{ij} (p by 1) and z_{ij} (q by 1), with z_{ij} often being a subset of x_{ij}. Denote the total number of observations by $N = \sum_{k=1}^{I} n_k$. Given a q-dimensional random effects vector b, $y_{ij}'s$ are assumed to be conditionally independent with distribution chosen from the one parameter exponential families. This formulation includes situations where the random effects are nested within groups and when they are not. The conditional mean $\mu_{ij} = \mathrm{E}(Y_{ij}|b)$ is related to the linear predictor η_{ij} via a monotonic link function $g(\mu_{ij}) = \eta_{ij}$, i.e.

$$\eta_{ij} = g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T b,$$

where β is a p by 1 vector of fixed effects.

In matrix form, denote

$$y = (y_{11}, y_{12}, \ldots, y_{In_I})^T,$$
$$\eta = (\eta_{11}, \eta_{12}, \ldots, \eta_{In_I})^T,$$

and the design matrices X (N by p) and Z (N by q) with rows x_{ij}^T and z_{ij}^T. The linear predictor can then be written as

$$\eta = X\beta + Zb.$$

It is traditional to assume an independent multivariate normal distribution for b with mean zero and variance-covariance matrix Σ, i.e.

$$b \sim N(0, \Sigma),$$

although there is little information available to guide this choice.

1.1.2 Likelihood Inference

Inference for GLMMs are typically based on the marginal likelihood, although there exist other methods based on generalized estimating equations (Ziegler, 2011) and Bayesian inference (Fong et al., 2010). The marginal likelihood of GLMMs is obtained by integrating over all possible values of the random effects, i.e.

$$L = \int f(y|b)f(b)\mathrm{d}b.$$

Due to the assumption of multivariate normal distribution, the marginal likelihood does not generally has a closed-form solution. Various approximation methods have been proposed to circumvent this problem, which we review in Sections 2.1 and 4.1.

1.1.3 Are Normal Random Effects Necessary?

Imposing a normal distribution on the random effects is convenient when we want to build in correlation structure. Normal distribution is also easy to interpret, with computational methods widely implemented in common statistical software packages. However, the normality assumption for random effects does result in analytical difficulty of the marginal likelihood, except for a few simple models. Although the vast majority of applications assume normal random effects, Stroup (2013) speculates that non-normal random effects will be a commonplace in the future, just like GLMs and linear mixed models are commomplace as of today but would have been considered beyond the state of the art a few decades ago. In fact, there are already developments in the mixed models literature that move away from the Gaussian random effects. A selection of examples include Lee and Nelder (1996); Verbeke and Lesaffre (1997); Zhang and Davidian (2001); Zhang et al. (2008); and Tzavidis et al. (2015).

1.2 Contribution

This thesis aims to contribute advances in the computational aspects of GLMMs. Focussing primarily on two-level data where the random effects are mapped uniquely onto the grouping structure and are independent between groups, the novel contributions of this thesis can be summarized as follows:

- *Chapter 2, Conjugate Generalized Linear Mixed Models for Clustered Data.*
  - Motivated by tractable marginal likelihood representation of GLMMs, a connection is drawn between the posterior in the Bayesian setting and the marginal likelihood in the frequentist setting. By relaxing the independent and identical assumptions of the standard Bayesian conjugacy theory, a class of computationally efficient GLMMs for two-level data is developed that is able to incorporate unit-level covariates while maintaining a closed-form representation of the marginal likelihood. The resulting marginal likelihood can be maximized directly without having to resort to approximate inference.
  - Distributions considered include Gaussian, Poisson, binomial and gamma, where we derived the most generic formulation such that the marginal likelihood is tractable.
  - It is shown that closed-form representation of the marginal likelihood for binomial (and hence multinomial) mixed models does not exist regardless of the random effect distribution, except for the special case when there are no covariates. This is taken as a basis for subsequent methodological development in Chapter 4.
- *Chapter 3, Exploring the Robustness of Normal vs. Log-Gamma for Random Effects Distributions: The Case of Count Data.*
  - Empirical performance of Poisson GLMMs vs. CGLMMs is investigated in terms of estimation of fixed effects and prediction of random effects.
  - The performance of both models are shown to be quite comparable in a simulation study.

- *Chapter 4, On the "Poisson Trick" and its Extensions for Fitting Multinomial Regression Models.*
    - A comprehensive review of the relationship between multinomial and Poisson regresion models is provided for independent responses, where a parsimonious strategy is presented when all the covariates are categorical.
    - A novel approach for modelling correlated responses based on the Poisson CGLMMs is proposed, where the marginal likelihood can be expressed in closed-form. An estimation procedure based on the Expectation/Conditional Maximization algorithm is derived, which can be implemented using functions for fitting generalized linear models readily available in standard statistical software packages.
    - Our formulation is shown to perform favourably compared to the existing approaches using Poisson GLMMs, when fitted to a yogurt brand choice dataset.
- *Chapter 5, Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era.*
    - The closed-form marginal likelihood property of CGLMMs is exploited for model fitting using sufficient and summary statistics in the context of large-scale administrative databases. The summaries are based on simple constructions such as sums and simple mathematical functions (exponential and log), which is easy to calculate. Under Poisson CGLMMs, using summary information does not result in any loss of information compared to unit record data.
    - Aside from offering benefits in terms of privacy protection, the proposed methodology allows for potential analysis of confidentialized tabular data at a finer geographical level.

- The potential benefits of the proposed framework in terms of modelling distributed data is illustrated via a hypothetical scenario of hospital variation studies. Data from different hospitals can be analyzed directly without the need to combine them, avoiding the sharing of sensitive information.

Recommendations for future work are given in *Chapter 6*. Literature review that is tailored for each project is provided within individual chapters throughout the thesis, rather than in a separate section.

1.3 List of Publications

Part of the methodological chapters (Chapters 2, 4 and 5) in this thesis have been published or are in submission. The most current version of the articles are included “as it is”, and it is inevitable that the chapters differ in styles and notational conventions due to the requirements of the journals being different. The full list of journal papers are provided below:

J-1. **Jarod Y. L. Lee**, Peter J. Green and Louise M. Ryan. Conjugate Generalized Linear Mixed Models for Clustered Data. *In Submission.*

J-2. **Jarod Y. L. Lee**, Peter J. Green and Louise M. Ryan. On the “Poisson Trick” and its Extensions for Fitting Multinomial Regression Models. *In Submission.*

J-3. **Jarod Y. L. Lee**, James J. Brown and Louise M. Ryan (2017). Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era. *The American Statistician*, **71(3)**, xxx-xxx (page numbers pending).

1.4 List of Conference Presentations

Some of the materials in this thesis have been presented in conference held locally and internationally, both invited and contributed. The full list of conference presentations are provided below:

C-1. “Working with Distributed Data.” *A Degustation of Australian Bureau of Statistics Big Data Problems.* Australian Bureau of Statistics, Canberra, Australia, March 13, 2015 (Invited).

C-2. “Multilevel Modelling of Tabular Counts via Deconstructed Maximum Likelihood.” *World Statistics Congress.* Riocentro, Rio De Janeiro, Brazil, July 26-31, 2015.

C-3. “Multilevel Modelling of Counts with Gamma-Poisson Model.” *International Biometric Conference.* Victoria Convention Centre, Victoria, Canada, July 10-15, 2016.

C-4. “On the “Poisson Trick” and its Extensions for Fitting Multinomial Regression Models.” *Australian Statistical Conference.* Hotel Realm, Canberra, Australia, December 5-9, 2016.

C-5. “Extracting More Value from Confidentialised Tabular Data.” *International Statistical Institute Regional Statistics Conference.* Bali International Convention Center, Bali, Indonesia, March 20-24, 2017 (Invited).

C-6. “Exploring the Robustness of Normal vs Log Gamma for Random Effects Distribution: The Case of Count Data.” *International Conference on Robust Statistics.* University of Wollongong, Wollongong, Australia, July 3-7, 2017 (Invited).

Chapter 2

Conjugate Generalized Linear Mixed Models for Clustered Data

Summary

This chapter concerns a class of generalized linear mixed models for clustered data, where the random effects are mapped uniquely onto the grouping structure and are independent between groups. We derive necessary and sufficient conditions that enable the marginal likelihood of such class of models to be expressed in closed-form. Illustrations are provided using the Gaussian, Poisson, binomial and gamma distributions. These models are unified under a single umbrella of conjugate generalized linear mixed models, where "conjugate" refers to the fact that the marginal likelihood can be expressed in closed-form, rather than implying inference via the Bayesian paradigm. Having an explicit marginal likelihood means that these models are more computationally convenient, which can be important in big data contexts. Except for the binomial distribution, these models are able to achieve simultaneous conjugacy, and thus able to accommodate both unit and group level covariates.

Keywords: Generalized linear mixed model; longitudinal data; multilevel model; unit level model; random effect.

2.1 Introduction

Generalized linear mixed models (Jiang, 2007; Stroup, 2013; Wu, 2010) are a broad class of models that can account for the dependency structure inherent within multilevel and longitudinal data, where the responses of units within a group are correlated. The grouping structure can be hospital, postal area, school, individual etc., and the goal is to model the response as a function of unit and group level covariates while accounting for group to group variability. For example, outcomes of patients within the same hospital are likely to be dependent due to similar risk profiles and a common clinical management practice. Generalized linear mixed models provide a natural framework for modelling dependencies by allowing for random group-specific effects.

Despite being popular for application in areas such as marketing, biological and social sciences, generalized linear mixed models are computationally intensive to fit, especially for large scale applications such as recommender systems (Perry, 2017) and discrete choice modelling (Hensher et al., 2015; Train, 2009). Inference for generalized linear mixed models is typically likelihood based, involving a multidimensional integral which usually does not have an analytic expression. Common estimation procedures include "exact" methods such as numerical quadrature (Rabe-Hesketh et al., 2002) and Monte Carlo methods; approximate methods such as Laplace approximation (Tierney and Kadane, 1986) and penalized quasi-likelihood (Breslow and Clayton, 1993); hierarchical likelihood (Lee and Nelder, 1996); simulated maximum likelihood (Train, 2009, p.238–239). Some of these approaches apply an expectation-maximization algorithm that treats the random effects as missing data (McCulloch, 1997). "Exact" methods can approximate the likelihood with arbitrary accuracy but are computational expensive. Approximate methods avoid the intractable integrals but may result in non-negligible bias (Lin and Breslow, 1996).

For large scale applications, it is important that models can be fit in a reasonable time. Several methods have been proposed in various settings. Zhang and Koren (2007) exploit the sparsity of predictors to achieve speedup for Bayesian hierarchical models. Luts et al. (2014) use variational approximations to fit real-time Bayesian hierarchical models to streaming data. Perry (2017) proposes a moment-based procedure that is non-iterative. Scott et al. (2016) propose a fitting strategy based on the divide and recombine principle, where data are partitioned into manageable subsets and the intended statistical analysis are performed independently on each subsets before combining the results. The methods proposed by Perry (2017) and Scott et al. (2016) are well suited for implementation in the context of distributed computing.

In this chapter, we are concerned with a class of generalized linear mixed models for two-level data, where the random effects are mapped uniquely onto the grouping structure and are independent between groups. We derive necessary and sufficient conditions that enable the marginal likelihood of such class of models to be expressed in closed-form. Having an explicit marginal likelihood means that one can proceed directly to maximization without having to resort to approximate inference. We consider the most common distribution families, that is, Gaussian, Poisson, binomial and gamma. These models are unified under a single umbrella of *conjugate generalized linear mixed models*, where "conjugate" in this context refers to the tractable form of the marginal likelihood, rather than implying inference via the Bayesian paradigm.

2.2 Exponential Family and Conjugate Prior

The likelihood of a one parameter exponential family with dispersion can be written in the general form:

$$f_{Y|\theta}(y|\theta,\phi)=\exp\left\{(y\theta-b(\theta))/\phi+c(y,\phi)\right\}, \tag{2.1}$$

for some specified functions $b(\theta)$ and $c(y,\phi)$, where θ is the *canonical parameter* and can be expressed as a function of the mean $\theta(\mu)$, and ϕ is the *dispersion parameter*, assumed known.

For such an exponential family, there exists a family of prior distributions on θ such that the posterior is in the same family as the prior. Such a conjugate prior for θ is defined as:

$$f_\Theta(\theta|\chi,\nu)=g(\chi,\nu)\exp\{\chi\theta-\nu b(\theta)\}, \tag{2.2}$$

where χ and ν are parameters and $g(\chi,\nu)$ denotes the normalizing factor.

The posterior can be obtained by multiplying the likelihood and the prior (up to a constant of proportionality):

$$f_{\Theta|y}(\theta|y,\chi,\nu,\phi)\propto\exp\{c(y,\phi)\}g(\chi,\nu)\exp\{\theta(\chi+y/\phi)-b(\theta)(\nu+1/\phi)\}, \tag{2.3}$$

which has the same kernel as the prior, but with different parameters. The updated parameters, based on a single observation y, are

$$\tilde{\chi}=\chi+y/\phi,\quad \tilde{\nu}=\nu+1/\phi.$$

For n independent and identically distributed observations y_j, $j=1,\ldots,n$, it is straightforward to show that conjugacy still holds and the updated parameters are

$$\tilde{\chi}=\chi+\sum_j y_j/\phi,\quad \tilde{\nu}=\nu+n/\phi.$$

These are the standard results for independent and identically distributed data in the Bayesian context. In this chapter, we aim to achieve explicit marginal likelihood for generalized linear mixed models in the frequentist setting. This is attained by establishing a connection between the posterior in the Bayesian paradigm and the marginal likelihood in the frequentist paradigm, and relaxing the assumption of identical distribution. The result is a class of models where unit level covariates can be conveniently incorporated while maintaining a closed-form representation of the marginal likelihood, which we refer to as *conjugate generalized linear mixed models*.

2.3 Conjugate Generalized Linear Mixed Models

2.3.1 From Bayesian formalism to frequentist inference - group level models

We now make a transition from the Bayesian paradigm, where θ is a parameter and its distribution is the prior, to the frequentist paradigm, where θ is a group specific random effect and its distribution describes the variation between groups.

Specifically, consider the two-level setting where the responses $y_{ij}, j = 1, \dots, n_i$ are grouped within a higher level structure indexed by $i = 1, \dots, I$, with $n = n_1 + \cdots + n_I$ being the total number of observations across all groups. The responses are assumed to come from the same exponential family. *Random effects* with a specified distribution are introduced at the group level to account for the correlation between units in a given group. Within each group, the responses are conditionally independent given the group specific random effects. Such data structure is common in many scenarios, for instance students within schools, patients within hospitals, residents within postal areas and repeated measurements from individuals.

For this model setup, the distribution from which the responses are drawn is governed by a group specific parameter θ_i which itself is drawn from a distribution chosen so that the resulting marginal likelihood is explicit. The marginal likelihood, obtained by integrating out the random effects, is given by

$$L = \prod_i \int \prod_j f_{Y|\theta_i}(y_{ij}|\theta_i, \phi) f_{\Theta_i}(\theta_i|\chi, \nu) d\theta_i,$$

where the integrand is proportional to the posterior in (2.3).

Imposing a conjugate prior distribution on the random effects would ensure that the integrand comes from a recognizable density function, which would enable the marginal likelihood to be expressed in closed-form. Solving for the integral, the likelihood contribution for the entire data is

$$L = \prod_i \left\{ \frac{\exp\left(\sum_j c(y_{ij}, \phi)\right) g(\chi, \nu)}{g\left(\chi + \sum_j y_{ij}/\phi, \nu + n/\phi\right)} \right\}. \tag{2.4}$$

This is the formulation for group level models in the absence of of unit level covariates. Although the random effects $\theta_i = \theta(\mu_i)$ are typically expressed in terms of a monotonic transformation of μ_i, interest usually lies in the distribution of μ_i. Consonni and Veronese (1992) and Gutiérrez-Pẽna and Smith (1995) showed that the conjugate distribution on μ_i coincides with the prior on μ_i induced by the conjugate distribution on θ_i if and only if the exponential family has a quadratic variance function. This holds for some of the most widely used distribution, including the Gaussian, Poisson, binomial and gamma (Morris, 1983), providing a convenient way to incorporate group level variables, for example, via the mean μ_i using a monotonic link function.

2.3.2 Relaxing the assumption of identical distribution - unit level models

Relaxing the assumption of identical distribution, we consider the regression setting where each observation y_{ij} is allowed to have a separate parameter $\theta_{ij} = \theta(x_{ij})$ that is a function of the covariates, while ϕ, if present, is constant across all observations. We want to explore the most generic formulation that leads to marginal likelihood simplification using the idea of Bayesian conjugacy, and thus we leave open the functional dependence of θ_{ij} on x_{ij} at this stage.

Denote $\theta_0 = \theta(x_0)$ as the baseline parameter where x_0 is an arbitrary baseline covariate value. In this chapter, we assume $x_0 = 0$, but user can take any baseline appropriate for the problem at hand. Technically, θ_0 is also indexed by i to reflect the group correlated data structure, but this can be suppressed without ambiguity. Likewise, for ease of notation, the i and j indexing are suppressed for most of the remaining chapter.

Remark 1. *With this formulation, within a group, we can think of units with covariate configuration that deviate from the baseline characteristics as modifying θ_0. This is as opposed to the standard formulation of generalized linear mixed models, where for a given unit with a particular covariate configuration, it is the group membership that modifies the linear predictor.*

Imposing a conjugate prior distribution on θ_0, the integrand of the marginal likelihood for a single observation has the form

$$\exp\left[\{y\theta(x) - b(\theta(x))\}/\phi + \chi\theta_0 - \nu b(\theta_0)\right]. \tag{2.5}$$

Remark 2. *The conjugate prior distribution is placed on $\theta_0 = \theta(x_0)$, rather than explicitly on each $\theta_{ij} = \theta(x_{ij})$.*

Equation (2.5) lies in the same family as (2.2) in its dependence on θ_0 if and only if both $\theta(x)$ and $b(\theta(x))$ are affine functions of θ_0 and $b(\theta_0)$, i.e. if there exist functions p, q, r, s, t and u of x such that

$$\theta(x) = p(x)\theta_0 + q(x)b(\theta_0) + r(x) \tag{2.6}$$

$$b(\theta(x)) = s(x)\theta_0 + t(x)b(\theta_0) + u(x). \tag{2.7}$$

We are interested in families where $\theta(x)$ has non-trivial dependence on x, that is, at least one of p, q or r must depend on x. When this occurs, the induced prior for $\theta(x)$ exhibits simultaneous conjugacy across all values of x, and the resulting model is capable of incorporating unit level covariates while maintaining a closed-form likelihood. Otherwise, this formulation reduces back to a group level model.

Remark 3. *Since $\theta_0 = \theta(x_0)$ and $b(\theta_0) = b(\theta(x_0))$, it is clear that $p(x_0) = 1$, $q(x_0) = 0$, $r(x_0) = 0$, $s(x_0) = 0$, $t(x_0) = 1$ and $u(x_0) = 0$. These constraints need to be satisfied when choosing the functional solutions for p, q, r, s, t and u.*

Conditions (2.6) and (2.7) can be combined to obtain

$$b\left\{p(x)\theta_0 + q(x)b(\theta_0) + r(x)\right\} = s(x)\theta_0 + t(x)b(\theta_0) + u(x). \tag{2.8}$$

This is the key equation in deriving the functional solutions for p, q, r, s, t and u. Under (2.8), the integrand of the marginal likelihood for a single observation is

$$\exp\left[\theta_0\{\chi + (yp(x) - s(x))/\phi\} - b(\theta_0)\{\nu + (t(x) - yq(x))/\phi\}\right].$$

Solving for this integral, the likelihood contribution for the observations within a single group is

$$L = \frac{\exp\left(\sum_j c(y_{ij}, \phi)\right) g(\chi, \nu) \exp\left(\sum_j (r(x_{ij})y_{ij} - u(x_{ij}))/\phi\right)}{g\left(\chi + \sum_j \left(y_{ij}p(x_{ij}) - s(x_{ij})\right)/\phi, \nu + \sum_j \left(t(x_{ij}) - y_{ij}q(x_{ij})\right)/\phi\right)}. \tag{2.9}$$

For multiple groups, the likelihood contribution can be obtained by multiplying (2.9) across the group index i.

2.4 Examples

2.4.1 Gaussian Distribution (with Known Variance)

The Gaussian density function (with known variance $\sigma^2 \geq 0$) can be written in the form

$$\exp\left\{\frac{y\mu_0 - \mu_0^2/2}{\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right) - \frac{y^2}{2\sigma^2}\right\},$$

where $\mu_0 \in \mathbb{R}$ is the mean of y. This can be written in the form of (2.1) if we write $\theta_0 = \mu_0$, $b(\theta_0) = \theta_0^2/2$, $\phi = \sigma^2$ and $c(y, \phi) = -\{\log(2\pi\phi) + y^2/\phi\}/2$.

To determine the conjugate distribution for θ_0, we compute the normalization factor

$$g(\chi, \nu) = \left\{\int \exp\left(\chi\theta_0 - \nu\frac{1}{2}\theta_0^2\right) d\theta_0\right\}^{-1} = \sqrt{\frac{\nu}{2\pi}} \exp\left\{-\left(\frac{\chi^2}{2\nu}\right)\right\},$$

where the integrand is the kernel of a Gaussian density function with mean $\mathrm{E}(\theta_0) = \lambda = \chi/\nu$ and variance $\mathrm{Var}(\theta_0) = \kappa^2 = 1/\nu$. This implies

$$\mu_0 = \theta_0 \sim \text{Gaussian}\left(\lambda, \kappa^2\right).$$

Group level covariates can be incorporated via the mean of μ_0, by replacing λ with $\lambda_i = x_i^T\beta$ for example. To incorporate unit level covariates, (2.8) requires

$$b(\theta(x)) = \frac{1}{2}\left\{p(x)\theta_0 + q(x)\frac{1}{2}\theta_0^2 + r(x)\right\}^2 \equiv s(x)\theta_0 + t(x)\frac{1}{2}\theta_0^2 + u(x),$$

which gives the following solution set:

$$p(x) = \zeta_1(x), \quad q(x) = 0, \quad r(x) = \zeta_2(x),$$
$$s(x) = \zeta_1(x)\zeta_2(x), \quad t(x) = \zeta_1^2(x), \quad u(x) = \zeta_2^2(x)/2,$$

where $\zeta_1(x)$ and $\zeta_2(x)$ are user-specified functions of x, subject to $\zeta_1(x_0) = 1$ and $\zeta_2(x_0) = 0$. This implies $\theta(x) = \mu(x) = \zeta_1(x)\mu_0 + \zeta_2(x)$.

As an example, choosing $\zeta_1(x) = 1$ and $\zeta_2(x) = x^T\beta$ gives rise to the random intercept model $\mu(x) = \mu_0 + x^T\beta$, where x does not include the constant 1 so that $\zeta_2(x_0) = 0$. Random slopes can be incorporated by writting $\mu(x) = z^T\mu_0 + x^T\beta$, where μ_0 is now a vector and z is a known design matrix for the random effects (usually a subset of x).

2.4.2 Poisson Distribution

The Poisson density function can be written in the form

$$\exp\left(y\log\mu_0 - \mu_0 - \log y!\right),$$

where $\mu_0 > 0$ is the rate parameter. This can be written in the form of (2.1) if we write $\theta_0 = \log\mu_0$, $b(\theta_0) = e^{\theta_0}$, $\phi = 1$ and $c(y,\phi) = -\log y!$.

To determine the conjugate distribution for θ_0, we compute the normalization factor

$$g(\chi,\nu) = \left\{ \int \exp\left(\chi\theta_0 - \nu\exp(\theta_0)\right) d\theta_0 \right\}^{-1} = \frac{\nu^{\chi}}{\Gamma(\chi)},$$

where the integrand is the kernel of a log-gamma density function with shape $A = \chi > 0$ and scale $B = \nu^{-1} > 0$, $\Gamma(\cdot)$ is the gamma function. This implies

$$\mu_0 = \exp(\theta_0) \sim \text{Gamma}\left(A, B\right).$$

Christiansen and Morris (1997) considered a similar model without covariates in the Bayesian setting. Group level covariates can be incorporated via the mean of μ_0, by letting $\mathrm{E}(\mu_0) = AB \equiv \exp(x_i^T\beta)$ for example. As a result, we replace B in the likelihood equation by $B_i = \exp(x_i^T\beta)/A$. To incorporate unit level covariates, (2.8) requires

$$b(\theta(x)) = \exp(p(x)\theta_0 + q(x)\exp(\theta_0) + r(x)) \equiv s(x)\theta_0 + t(x)\exp(\theta_0) + u(x),$$

which gives the following solution set:

$$\begin{aligned} p(x) &= 1, \quad q(x) = 0, \quad r(x) = \zeta(x), \\ s(x) &= 0, \quad t(x) = e^{\zeta(x)}, \quad u(x) = 0, \end{aligned}$$

where $\zeta(x)$ is a user-specified function of x, subject to $\zeta(x_0) = 0$. This implies $\theta(x) = \log(\mu(x)) = \theta_0 + \zeta(x)$, or equivalently, $\mu(x) = \mu_0\exp(\zeta(x))$.

As an example, choosing $\zeta(x) = x^T\beta$ leads to $\mu(x) = \mu_0\exp(x^T\beta)$, where x does not include the constant 1 so that $\zeta(x_0) = 0$. This is a sensible choice as $\mu(x)$ is guaranteed to be always positive. Similar multiplicative models with unit level covariates have been considered by Lee and Nelder (1996) , Lee et al. (2017) and Lee et al. (2017b) and in various settings.

2.4.3 Binomial Distribution (with Known Number of Trials)

The binomial density function (with fixed number of trials $n \in \mathbb{N}$) can be written in the form

$$\exp\left\{y \log\left(\frac{\mu_0}{1-\mu_0}\right) + n\log(1-\mu_0) + \log\binom{n}{y}\right\},$$

where $0 \leq \mu_0 \leq 1$ is the probability of success. This can be written in the form of (2.1) if we write $\theta_0 = \log(\mu_0(1-\mu_0)^{-1})$, $b(\theta_0) = n\log(1+\exp(\theta_0))$, $\phi = 1$ and $c(y,\phi) = \log\binom{n}{y}$.

To determine the conjugate distribution for θ_0, we compute the normalization factor

$$g(\chi,\nu) = \left[\int \exp\{\chi\theta_0 - \nu\log(1+\exp(\theta_0))\}d\theta_0\right]^{-1} = \frac{1}{\mathrm{B}(\chi,\nu-\chi)},$$

where the integrand is the kernel of the log of a beta prime density function with shape parameters $A = \chi > 0$ and scale $B = \nu - \chi > 0$, $\mathrm{B}(\cdot)$ is the beta function. This implies

$$\mu_0 = \exp(\theta_0)/(1+\exp(\theta_0)) \sim \mathrm{Beta}\,(A,B)\,.$$

Kleinman (1973), Crowder (1978) and He and Sun (1998) considered similar models without covariates in various settings. Group level covariates can be incorporated via the mean of μ_0. Reparameterizing the beta density function by setting the mean $\lambda = A/(A+B)$ and precision $\phi = A+B$, we can allow λ_i to be some function of $x_i^T\beta$, say, $\lambda_i = \{1+\exp(-x_i^T\beta)\}^{-1}$ (Ferrari and Cribari-Neto, 2004). As a result, we replace A and B in the likelihood equation by $\lambda_i\phi$ and $\phi - \lambda_i\phi$, respectively. To incorporate unit level covariates, (2.8) requires

$$b(\theta(x)) = \log\left[1+\exp\left\{p(x)\theta_0 + q(x)\log(1+\exp(\theta_0)) + r(x)\right\}\right] \equiv$$
$$s(x)\theta_0 + t(x)\log\left(1+\exp(\theta_0)\right) + u(x),$$

which gives the following solution set:

$$p(x) = 1, \quad q(x) = 0, \quad r(x) = 0,$$
$$s(x) = 0, \quad t(x) = 1, \quad u(x) = 0.$$

Since neither p, q nor r depend on x, it is impossible to simultaneously incorporate unit level covariates while maintaining closed-form likelihood.

2.4.4 Gamma Distribution (with Known Shape)

For modelling purposes, it is convenient to reparameterize the gamma distribution with shape $A > 0$ and scale $B_0 > 0$ in terms of A and mean $\mu_0 = AB_0 > 0$. The reparameterized gamma density function (with fixed shape A) can be written in the form

$$\exp\left[\frac{-y\mu_0^{-1} - \log\mu_0}{A^{-1}} + A\log(Ay) - \log y - \log\Gamma(A)\right].$$

This can be written in the form of (2.1) if we write $\theta_0 = -\mu_0^{-1}$, $b(\theta_0) = -\log(-\theta_0)$, $\phi = A^{-1}$ and $c(y, \phi) = A\log(Ay) - \log y - \log\Gamma(A)$.

To determine the conjugate distribution for η_0, we compute the normalization factor

$$g(\chi, \nu) = \left[\int \exp\left\{\chi\theta_0 - \nu(-\log(-\theta_0))\right\} d\theta_0\right]^{-1} = \frac{\chi^{\nu+1}}{\Gamma(\nu+1)},$$

where the integrand is the kernel of the negative of a gamma density function with shape $C = \nu + 1$ and scale $D = \chi^{-1}$, and $\Gamma(\cdot)$ is the gamma function. This implies

$$\mu_0 = -\theta_0^{-1} \sim \text{Inverse-Gamma}\,(C, D)\,.$$

Group level covariates can be incorporated via the mean of μ_0, by letting $\mathrm{E}(\mu_0) = D_i(C-1)^{-1} \equiv \exp(x_i^T\beta)$ for example, provided $C > 1$. As a result, we replace

D in the likelihood equation by $D_i = (C-1)\exp(x_i^T\beta)$. To incorporate unit level covariates, (2.8) requires

$$b(\theta(x)) = -\log\{p(x)\theta_0 + q(x)\log(-\theta_0) - r(x)\} \equiv s(x)\theta_0 - t(x)\log(-\theta_0) + u(x),$$

which gives the following solution set:

$$p(x) = \zeta(x), \quad q(x) = 0, \quad r(x) = 0,$$
$$s(x) = 0, \quad t(x) = 1, \quad u(x) = -\log\zeta(x),$$

where $\zeta(x)$ is a user-specified function of x, subject to $\zeta(x_0) = 1$. This implies $\theta(x) = -\mu^{-1}(x) = \zeta(x)\theta_0$, or equivalently, $\mu(x) = \mu_0/\zeta(x)$.

As an example, choosing $\zeta(x) = \exp(x^T\beta)$ leads to $\mu(x) = \mu_0/\exp(x^T\beta)$, where x does not include the constant 1 so that $\zeta(x_0) = 1$. This is a sensible choice as $\mu(x)$ is guaranteed to be always positive.

2.4.5 Summary

Table 2.1 and 2.2 summarize the results discussed in this section, for group level and unit level models, respectively. We have covered the four distribution families that are most important in practice. Results for other distributions could be derived as needed.

2.5 An Illustrative Example: Poisson responses

Consider the well-known epileptic seizure count data previously analyzed by Thall and Vail (1990), Breslow and Clayton (1993), Lee and Nelder (1996) and Ma and Jorgensen (2007), where 59 epileptics were randomized to a new drug (Trt =

Table 2.1 : Summary of group level models. Log-likelihood functions are contributed by a single observation. The i and j indexes are omitted for ease of notation.

	Gaussian (with known variance σ^2)
Model	$y \mid \mu_0 \sim \text{Gaussian}(\mu_0, \sigma^2) \quad \mu_0 \sim \text{Gaussian}(\lambda, \kappa^2)$
Log-likelihood	$-\frac{1}{2}\left\{\log(\sigma^2+\kappa^2)+\frac{y^2}{\sigma^2}+\frac{\lambda^2}{\kappa^2}-\frac{\lambda^2\sigma^4+2\lambda\kappa^2\sigma^2 y+\kappa^4 y^2}{\kappa^2\sigma^2(\sigma^2+\kappa^2)}\right\}$
Covariates	Replace λ by $\lambda_i = x_i^T\beta$
	Poisson
Model	$y \mid \mu_0 \sim \text{Poisson}(\mu_0) \quad \mu_0 \sim \text{Gamma}(A, B)$
Log-likelihood	$\log\Gamma\left(A+y\right)-(A+y)\log(B^{-1}+1)-\log\Gamma(A)-A\log B$
Covariates	Replace B by $B_i = e^{x_i^T\beta}/A$
	Binomial (with known number of trials n)
Model	$y \mid \mu_0 \sim \text{Bernoulli}(\mu_0) \quad \mu_0 \sim \text{Beta}(A, B)$
Log-likelihood	$\log B\left(A+y, B+1-y\right)-\log B(A, B)$
Covariates	Replace A and B by $\lambda_i\phi$ and $\phi-\lambda_i\phi$ respectively, where $\lambda_i = \left\{1+e^{-x_i^T\beta}\right\}^{-1}$
	Gamma (with known shape A)
Model	$y \mid \mu_0 \sim \text{Gamma}(A, \mu_0/A) \quad \mu_0 \sim \text{Inverse-Gamma}(C, D)$
Log-likelihood	$-\log B(A, C)+A\log(ADy)-\log y+(A+C)\log(1+ADy)$
Covariates	Replace D by $D_i = (C-1)e^{x_i^T\beta}$, provided $C > 1$

Table 2.2 : Summary of unit level models. Log-likelihood functions are contributed by a single observation. The i and j indexes are omitted for ease of notation.

	Gaussian (with known variance σ^2)
Model	$y \mid \mu_0 \sim \text{Gaussian}\left(\zeta_1(x)\mu_0 + \zeta_2(x), \sigma^2\right) \quad \mu_0 \sim \text{Gaussian}(\lambda, \kappa^2)$
Constraint	$\zeta_1(x_0) = 1 \quad \zeta_2(x_0) = 0$
Log-likelihood	$-\frac{1}{2}\left\{\log\left(\sigma^2 + \kappa^2 \sum_j \zeta_1^2(x)\right) + \frac{\sum_j y^2}{\sigma^2} + \frac{\lambda^2}{\kappa^2} - \frac{2\sum_j \zeta_2(x)y}{\sigma^2} + \frac{\sum_j \zeta_2^2(x)}{\sigma^2} + \frac{P}{Q}\right\}$ where $P = -\lambda^2\sigma^4 + 2\lambda\kappa^2\sigma^2\left(\sum_j \zeta_1(x)y\right) - \kappa^4\left(\sum_j \zeta_1(x)y\right)^2 +$ $2\kappa^4\left(\sum_j \zeta_1(x)y\right)\left(\sum_j \zeta_1(x)\zeta_2(x)\right) - \kappa^4\left(\sum_j \zeta_1(x)\zeta_2(x)\right)^2 -$ $2\lambda\kappa^2\sigma^2\left(\sum_j \zeta_1(x)\zeta_2(x)\right)$ $Q = \kappa^2\sigma^2\left(\sigma^2 + \kappa^2\sum_j \zeta_1^2(x)\right)$
Remark	Can incorporate random slopes if $\mu(x)$ is linear in terms of μ_0

	Poisson
Model	$y \mid \mu_0 \sim \text{Poisson}\left(\mu_0 e^{\zeta(x)}\right) \quad \mu_0 \sim \text{Gamma}(A, B)$
Constraint	$\zeta(x_0) = 0$
Log-likelihood	$\log\Gamma\left(A + \sum_j y\right) - \left(A + \sum_j y\right)\log\left(B^{-1} + \sum_j e^{\zeta(x)}\right) -$ $\log\Gamma(A) - A\log B + \sum_j \zeta(x)y$

	Gamma (with known shape A)
Model	$y \mid \mu_0 \sim \text{Gamma}(A, \mu(x)/A) \quad \mu_0 \sim \text{Inverse-Gamma}(C, D)$ where $\mu(x) = \mu_0/\zeta(x)$
Constraint	$\zeta(x_0) = 1$
Log-likelihood	$\log\Gamma(An_i + C) - n_i\log\Gamma(A) - \log\Gamma(C) + An_i\log A + (A-1)\sum_j \log y -$ $(An_i + C)\log\left\{1 + AD\left(\sum_j \zeta(x)y\right)\right\} + An_i\log D + A\sum_j \log\zeta(x)$ where n_i is the number of units within group i

progabide) or a placebo (Trt = placebo) at a clinical trial. Baseline data included the log seizure counts during the 8-week period before the trial (lbase) and the log age in years (lage), both centered to have zero mean. A multivariate response variable consisted of the counts seizures during the 2-weeks before each of four clinic visits. An indicator variable for the fourth visit (V4) was constructed to reflect the fact that counts are substantially lower during the fourth visit. The dataset are stored in the epil object within the MASS package in R (R Development Core Team, 2017).

Our reanalysis is primarily oriented toward comparing two different methods of incorporating random effects, namely, generalized linear mixed models (GLMM) using additive Gaussian random effects: $y_{ij}|u_i \sim \text{Poisson}(\exp(x_{ij}^T\beta + u_i))$, $\mu_i \sim \text{Gaussian}(0, \sigma^2)$; and conjugate generalized linear mixed models (CGLMM) using multiplicative Gamma random effects: $y_{ij}|u_i \sim \text{Poisson}(u_i \exp(x_{ij}^T\beta))$, $\mu_i \sim \text{Gamma}(A, 1/A)$. To allow for a direct comparison between the models, we included an intercept in the Poisson conjugate generalized linear mixed model, but fixed the mean of u_i to be one to ensure identifiability. Due to the intractable nature of the marginal likelihood of Poisson generalized linear mixed models, various approximation methods have been employed to estimate the marginal likelihood. The results are presented in Table 4.5.

In comparing the estimates and standard errors between the models, we note that the fixed effects model is likely to produce biased estimates as it did not take into account of the correlation induced by multiple measurements from the same individual. The parameter estimates and the standard errors of the random effect models are quite similar, regardless of the distribution of the random effects. This is probably due to the fact that the variance of the random effects not being too large, implying moderate subject-to-subject variability in seizure counts after taking into

Table 2.3 : Regression estimates for the epileptics data, and the associated standard errors.

		GLMM			
Variables	**GLM**	**Laplace**[1]	**AGQ**[2]	**PQL**[3]	**CGLMM**
Intercept	1.898 (0.043)	1.833 (0.105)	1.833 (0.106)	1.870 (0.106)	1.932 (0.105)
lbase	0.949 (0.044)	0.883 (0.131)	0.883 (0.131)	0.882 (0.129)	0.880 (0.126)
trtprogabide	-0.346 (0.061)	-0.334 (0.147)	-0.334 (0.148)	-0.310 (0.149)	-0.282 (0.146)
lage	0.888 (0.116)	0.481 (0.346)	0.481 (0.347)	0.534 (0.346)	0.505 (0.357)
V4	-0.160 (0.055)	-0.160 (0.054)	-0.160 (0.055)	-0.160 (0.077)	-0.160 (0.055)
lbase:trtprogabide	0.562 (0.064)	0.339 (0.202)	0.339 (0.203)	0.342 (0.203)	0.344 (0.193)
σ	N/A	0.501 (N/A)	0.502 (N/A)	0.444 (N/A)	N/A
A	N/A	N/A	N/A	N/A	3.935 (0.863)

[1] Laplace approximation: fitted using the glmer() function within the lme4 package in R.

[2] Adaptive Gauss-Hermite quadrature: fitted using the glmer() function within the lme4 package in R, using nAGQ=100.

[3] Penalized Quasi-Likelihood: fitted using the glmmPQL() function within the MASS package in R.

account of the covariate effects.

2.6 Remarks

Group level conjugate models have long been used in the context of Bayesian small area estimation and disease mapping, the most common ones being the gamma-Poisson (Rao and Molina, 2015, p. 383) and the beta-binomial models (Rao and Molina, 2015, p. 389). This chapter considers the frequentist setting where the most general conditions that allow for explicit marginal likelihood in unit level generalized linear mixed models are derived. The primary advantage of the proposed modelling framework is mathematical convenience, but the conjugate random effect distribution this assumes may not accurately reflect the real variation between groups. Mathematical convenience should not deter the exploration of alternative formulations for the distribution of random effects in this situation. Other applications of the proposed modelling framework include privacy preservation in large-scale administrative databases (Lee et al., 2017) and the fitting of discrete choice models (Lee et al., 2017b).

Some of the models derived from our conjugate generalized linear mixed models framework are similar to those of the *conjugate hierarchical generalized linear models* framework proposed by Lee and Nelder (1996). While the word "conjugate" in our framework reflects the fact that the marginal likelihood can be made explicit, it has quite a different meaning in the hierarchical likelihood framework (Lee and Nelder, 1996, p. 621), where it refers to the fact that a Bayesian conjugate prior is imposed on the random effects distribution but does not necessarily result in a closed-form likelihood.

Molenberghs et al. (2010) considered models that can simultaneously accommodate both overdispersion and correlation induced by grouping structures via two separate sets of random effects. They consider a combined model where the conjugate and Gaussian random effects induce overdispersion and association, respectively. Although they use the conjugate distribution for a set of random effects, the resulting marginal likelihood is generally not explicit.

Chapter 3

Exploring the Robustness of Normal vs. Log-Gamma for Random Effects Distributions: The Case of Count Data

Summary

When analyzing multilevel and longitudinal data, it is common to utilize generalized linear mixed models (GLMMs) with Gaussian random effects. In this work, we contrast Poisson GLMMs with Poisson models combined with log-gamma random effects. The latter belongs to the conjugate generalized linear mixed models (CGLMMs) framework presented in Chapter 2 and has advantages in terms of closed-form likelihood and privacy preservation in large-scale administrative databases. In this chapter, we explore the robustness of Poisson GLMMs vs. CGLMMs in terms of estimation of fixed effects and prediction of random effects, when the random effects distribution is misspecified. We show that the performance of Poisson GLMMs and CGLMMs is generally quite comparable under model misspecification, except for a few extreme cases.

Keywords: Clustered count data; Generalized linear mixed models; Mean squared error of prediction; Misspecified random effects distributions; Non-normality.

3.1 Introduction

It is traditionally assumed that the random effects in generalized linear mixed models (GLMMs) follow a multivariate normal distribution, although this assumption is very difficult to verify in practice. Several authors (McCulloch and Neuhaus, 2011a,b; Neuhaus et al., 2012) have investigated the effects of misspecifying the random effects distribution and most of them conclude that it has little effect on analyses. In this chapter, we aim to compare the performance of Poisson GLMMs vs. CGLMMs that assume a log-gamma random effects distribution. The latter has advantages in terms of closed-form likelihood (Lee et al., 2017a) and privacy preservation in large-scale administrative databases Lee et al. (2017). It is also arguable that the closed-form likelihood property of Poisson CGLMMs provide a convenient mean to incorporate survey weights using the pseudolikelihod approach, compared to that of Poisson GLMMs (Rabe-Hesketh and Skrondal, 2006). The main drawback of Poisson CGLMMs is perhaps the lack of readily available statistical softwares for fitting such models, although they can be fitted quite easily using standard optimization procedures available in almost every statistical software packages. If we can show comparable performance of Poisson GLMMs vs. CGLMMs under random effects misspecification, we can then make convincing arguments for using Poisson CGLMMs.

3.2 Models

Let P be a finite population of size N which can be partitioned into I domains, with P_i denoting population of known size N_i on domain i, $i = 1, \ldots, I$. Let y_{ij} denotes the outcome, i.e. number of events for observation j in domain i; x_{ij} denotes a $(p+1) \times 1$ vector of individual and domain level covariates. Given a domain-specific random effects u_i, we assume y_{ij} is independently Poisson distributed with mean $\lambda_{ij} = N_{ij} \exp(x_{ij}^T \alpha + u_i)$, where α is a $(p + 1) \times 1$ vector of unknown regression coefficients to be estimated, including the intercept. Different distributional assumptions for the random effects u_i lead to different models. Here, we consider two such models: Generalised Linear Mixed Models (GLMMs) and Conjugate Generalised Linear Mixed Models (CGLMMs).

3.2.1 Poisson Generalized Linear Mixed Models (GLMMs)

This is for the case where $(u_i)_{i=1}^{I}$ are normally distributed with mean 0 and variance-covariance matrix Σ, but we only restrict our attention to the case where $(u_i)_{i=1}^{I}$ are assumed to be independent with variance σ^2. The resulting marginal likelihood is intractable, but procedures for estimating the marginal likelihood of Poisson GLMMs are widely implemented in many common statistical software packages. In this article, we use the glmer() function within the lme4 package in R, which implements Laplace approximation by default.

3.2.2 Poisson Conjugate Generalized Linear Mixed Models (CGLMMs)

In Poisson CGLMMs, u_i are assumed to be independently and identically distributed according to a log-gamma (LG) distribution, i.e.

$$u_i \overset{ind.}{\sim} \mathrm{LG}\left(A, B\right)$$
$$f(u_i) = \frac{1}{\Gamma(A)B^A} \exp(Au_i) \exp\left(-\frac{e^{u_i}}{B}\right),$$

LG distribution is inherently negatively skewed, but approaches to symmetry as the shape parameter A tends to infinity. Refer to Figure 3.1 for a pictorial representation. The expected value and variance of u are given by $\mathrm{E}(u) = \psi(A) + \ln(B)$ and $\mathrm{V}(u) = \psi_1(A)$, respectively. Here, $\psi(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function, and $\ln(\cdot)$ is the natural logarithmic function. The marginal likelihood can then be expressed in closed-form, i.e.

$$L(\alpha, A, B; y) \propto \frac{\exp(\sum_j x_{ij}^T \alpha y_{ij})}{\Gamma(A) B^A} \Gamma\left(\sum_j y_{ij} + A\right) \left(\sum_j e^{x_{ij}^T \alpha} + \frac{1}{B}\right)^{-\left(\sum_j y_{ij} + A\right)} . \tag{3.1}$$

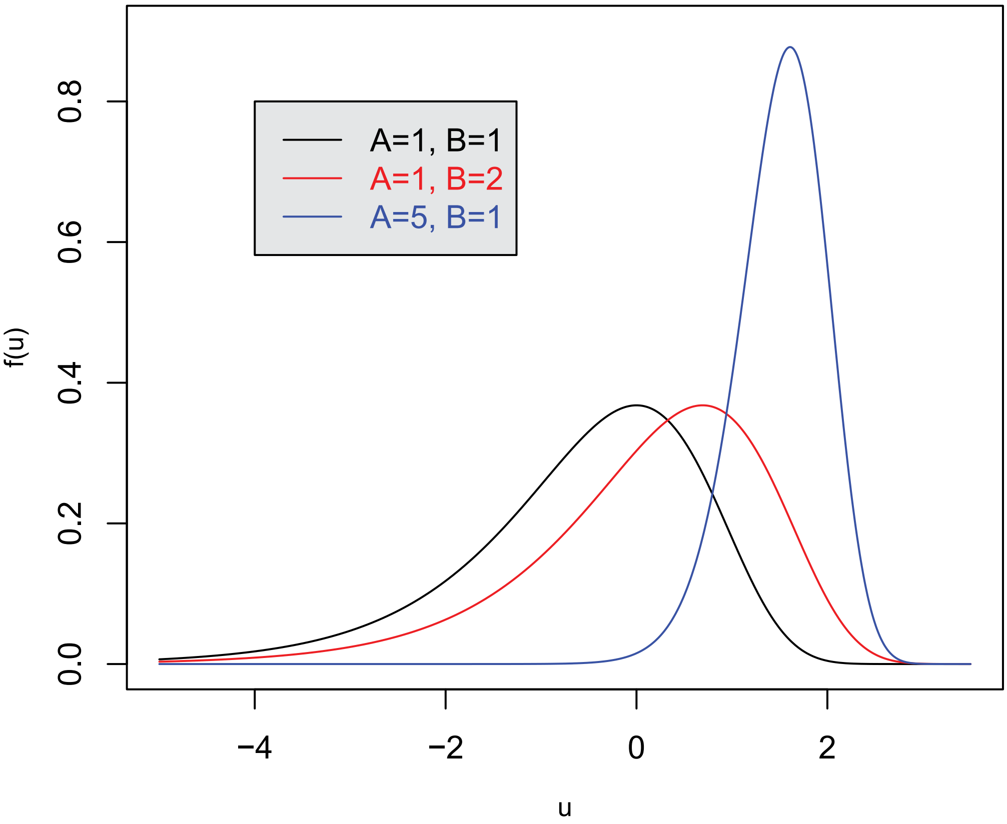


Figure 3.1 : Probability density function of log-gamma distribution with different parameter combinations. As the shape parameter increases, the distribution becomes more symmetric.

Taking log of Equation 3.1, we obtain the log-likelihood as

$$\begin{aligned}\ell(\alpha, A, B; y) = \text{constant} + \sum_i \sum_j x_{ij}^T \alpha y_{ij} - \sum_i \ln \Gamma(A) - \sum_i A \ln B \\ + \sum_i \ln \Gamma \left(\sum_j y_{ij} + A \right) - \sum_i \left[\left(\sum_j y_{ij} + A \right) \ln \left(\sum_j e^{x_{ij}^T \alpha} + \frac{1}{B} \right) \right]\end{aligned} \tag{3.2}$$

A constraint on either the random effects u_i or the intercept α_0 is needed to ensure identifiability, since

$$\lambda_{ij} = N_{ij} \exp \left((x_{ij,0} + c) + \sum_{k=1}^{p} x_{ij,k} \alpha + (u_i - c) \right) \qquad \text{for all } c \neq 0.$$

We choose to impose a constraint on u_i such that $\mathrm{E}(u_i) = \psi(A) + \log(B) = 1$. As a consequence, $B = e^{-\psi(A)}$.

The additive log-gamma formulation is also equivalent to the multiplicative gamma formulation $\lambda_{ij} = b_i \exp(x_{ij}^T \alpha)$, where $b_i = e^{u_i}$ and is independently gamma distributed with shape A and scale B (Lee et al., 2017a). Consequently the expected value is $\mathrm{E}(b) = AB$ and variance is $\mathrm{V}(b) = AB^2$. Both formulations are equivalent in the sense that they give rise to the same marginal likelihood. However, a constraint on one parameterisation does not translate naturally into the other parameterisation. For instance, a constraint of $\mathrm{E}(b_i) = \mathrm{E}(e^{u_i}) = 1$ is not equivalent to $\mathrm{E}(u_i) = 0$. With the latter random effects formulation, we have

$$\begin{aligned} b_i &\overset{ind.}{\sim} \text{Gamma}\,(A, B) \\ f(b_i) &= \frac{1}{\Gamma(A) B^A} b_i^{A-1} \exp \left(-\frac{b_i}{B} \right). \end{aligned}$$

The random effects b_i captures the deviation of individual rate from the average individuals with the same characteristics, due to remaining unexplained domain variations. With the multiplicative formulation, $b_i > 1$ corresponds to a positive deviation of area j from the mean rate, whereas $0 < b_i < 1$ corresponds to a negative deviation. In this chapter, we shall stick with the log-gamma formulation so that it is directly comparable with Poisson GLMMs, since both are formulated on the same addtive scale.

3.2.3 Prediction of Random Effects

Here, we present two different methods for predicting the random effects of the various Poisson mixed models:

Posterior Mode (minimizing the 0-1 error loss function)

(A) GLMMs: default implementation of the glmer() function within the lme4 package in R.

(B) CGLMMs (additive LG):

$$\hat{u}_i = \ln\left(\frac{\sum_j y_{ij} + A}{\sum_j e^{x_{ij}^T\alpha} + \frac{1}{B}}\right). \tag{3.3}$$

(C) CGLMMs (multiplicative Gamma):

$$\hat{b}_i = \frac{\sum_j y_{ij} + A - 1}{\sum_j e^{x_{ij}^T\alpha} + \frac{1}{B}} \qquad \text{iff} \quad \sum_j y_{ij} + A \geq 1. \tag{3.4}$$

Posterior Mean (minimizing the quadratic error loss function)

This is also known as the *best predictor* (BP), i.e. the predictor that minimizes the overall mean squared error of prediction.

(A) GLMMs: no closed-form expression.

(B) CGLMMs (additive LG):

$$\tilde{u}_i = \ln\left(\frac{1}{\sum_j e^{x_{ij}^T\alpha} + \frac{1}{B}}\right) + \psi\left(\sum_j y_{ij} + A\right). \tag{3.5}$$

(C) CGLMMs (multiplicative Gamma):

$$\tilde{b}_i = \frac{\sum_j y_{ij} + A}{\sum_j e^{x_{ij}^T\alpha} + \frac{1}{B}}. \tag{3.6}$$

These predictors depend on the estimated fixed effects and the parameters of the random effects distribution, in which we replace by their estimates, leading to the *empirical predictor*. In this chapter, we shall focus on the posterior mode predictor as opposed to the more commonly used posterior mean, as it is readily implemented in the glmer() function within the lme4 package in R. This allows us to concentrate on analyzing the results and minimize the risk of computing issues becoming a distraction.

3.3 Simulation Studies

3.3.1 Data Generation

We perform simulation studies to evaluate the performance of Poisson GLMMs vs. CGLMMs under various true distributions. We consider the following true distributions for u_i,:

1. Symmetric: Gaussian distribution with mean 0 and varying standard deviation σ.

2. Negatively Skewed: log-gamma distribution with varying shape A and varying scale $B = e^{-\psi(A)}$, so that the mean is 0.

3. Positively skewed: exponential distribution with varying rate λ, shifted to have mean 0.

The first two distributions are chosen to represent the case where one of the Poisson GLMMs (symmetric) and CGLMMs (negatively skewed but symmetric in the limit) is the "correct" model. The last distribution represents an extreme deviation from both Gaussian and log-gamma. The exponential distribution is heavily skewed, has high kurtosis, and a limited support.

Data consist of two mutually orthogonal covariates: an individual level and a domain specific covariate, both generated from the standard normal distribution. These quantities are kept *constant* throughout simulations. We simulate 250 areas, where each area is associated with a random effects u_i generated from one the true distributions specified above. The mean for individual observations is then calculated via $\lambda_{ij} = e^{x_{ij}^T \alpha + u_i}$. We fix $\alpha_0 = 0$, $\alpha_1 = -1$ and $\alpha_2 = 1$. In each run of the simulation, the response counts y_{ij} for each i and j are generated from a Poisson distribution with rate λ_{ij}. This simulation setup is repeated for different values of the random effects standard deviation $(0.1, 0.2, 0.5, 1)$ and cluster size $(5, 10, 20, 40)$. With 250 areas, this implies the corresponding total population is 1250, 2500, 5000 and 10000, respectively.

3.3.2 Bias Results for Fixed Effects

Figure 3.2 to 3.4 present biases in fixed effects estimates from the Poisson GLMMs and CGLMMs under various true distributions, where bias is defined as median of the simulation estimate minus true value. Consistent with the results presented in existing literature, the intercept and domain specific covariates are more sensitive to shape misspecification compared to individual level covariates. The biases from both models are nearly indistinguishable when random effects standard deviation is small.

For large standard deviation, the “true” model performs better in general, but the differences diminish as cluster size increases.

3.3.3 Coverage Results for Fixed Effects

Figure 3.5 to 3.7 present estimated coverage rates for fixed effects estimates from the Poisson GLMMs and CGLMMs under various true distributions. For each regression coefficient within each simulation, we construct a 95% confidence interval using the Wald method, i.e.

$$\begin{gathered}\hat{\alpha} \ \pm\ 1.96 \times \text{model-based standard error estimate obtained} \\ \text{from the information matrix of the fitted likelihood.}\end{gathered}$$

The coverage rate is defined as the proportion of the 95% intervals that captured the true parameter value.

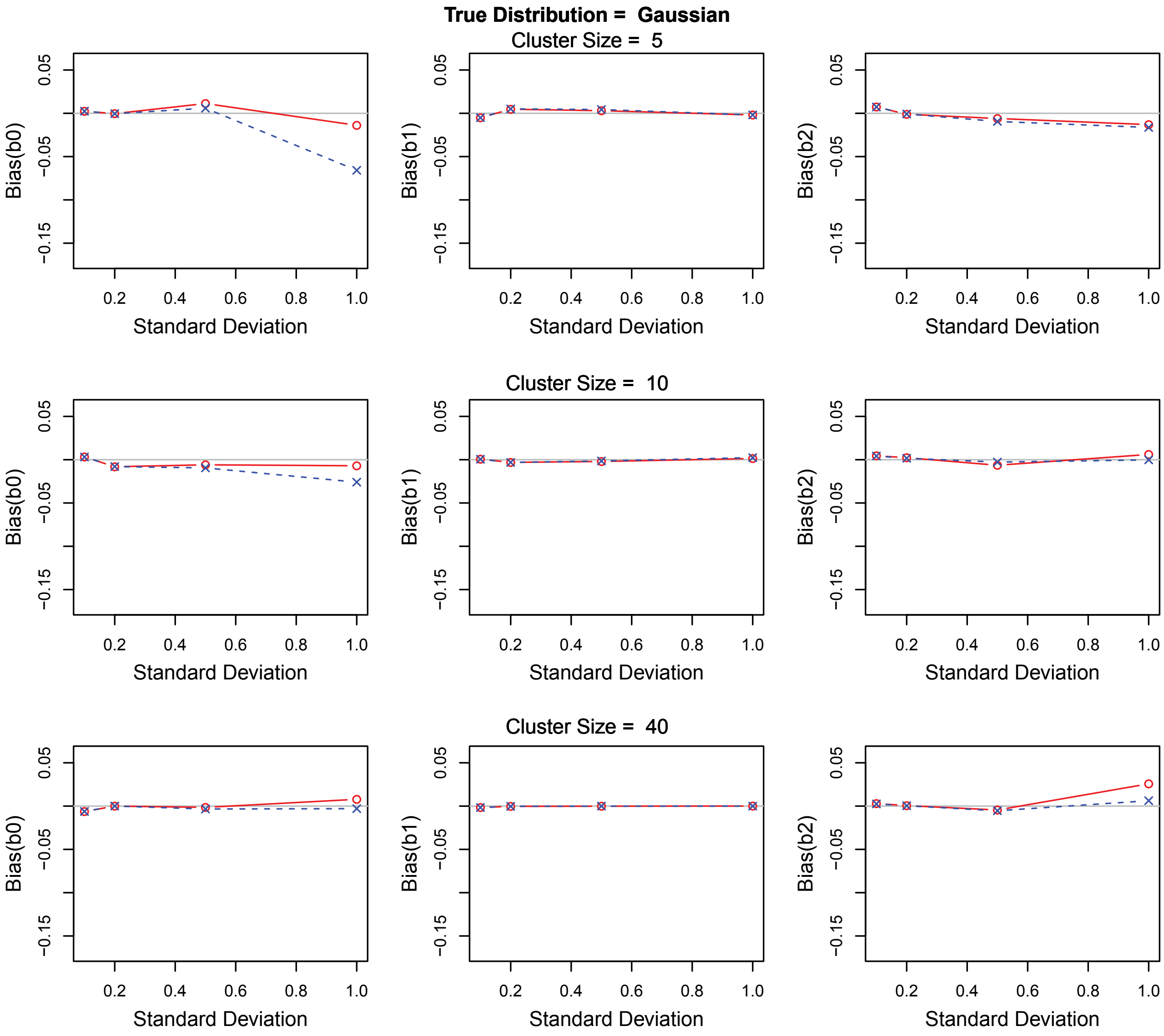
The coverage rates from both models are comparable under true Gaussian and log-gamma, except for the case of true Gaussian when cluster size = 40 and standard deviation = 1, where the coverage rate from Poisson CGLMMs is quite low compared to Poisson CGLMMs. For true exponential, the coverage rates of Poisson CGLMMs are substantially lower than that of GLMMs when the standard deviation is large, and the difference does not disappear as cluster size increases.

3.3.4 Mean Square Error of Prediction Results for Random Effects

Figure 3.8 to 3.10 present the empirical mean squared error of prediction (MSEP) from the Poisson GLMMs and CGLMMs under various true distributions, where MSEP is defined as $\mathrm{E}(\hat{u} - u)^2$ and $\hat{u}$ is the posterior mode estimate for the random effects, obtained by minimizing the zero-one error function. When the standard deviation is small, both Poisson GLMMs and CGLMMs give virtually identical results. For large standard deviation, there is a modest gain in MSEP for using the "true" distribution as the fitted distribution, and the benefit does not seem to diminish as cluster size increases. This is in contrast with the message from McCulloch and Neuhaus (2011b) that "the primary determinant of the MSEP is the cluster size". The discrepancy in MSEP is the largest when the true distribution is exponential and the standard deviation is large, in which Poisson GLMMs outperforms CGLMMs by quite a large amount. This is not surprising as Gaussian distribution (symmetric) is "closer" to exponential distribution, as opposed to log-gamma distribution that is inherently left-skewed.

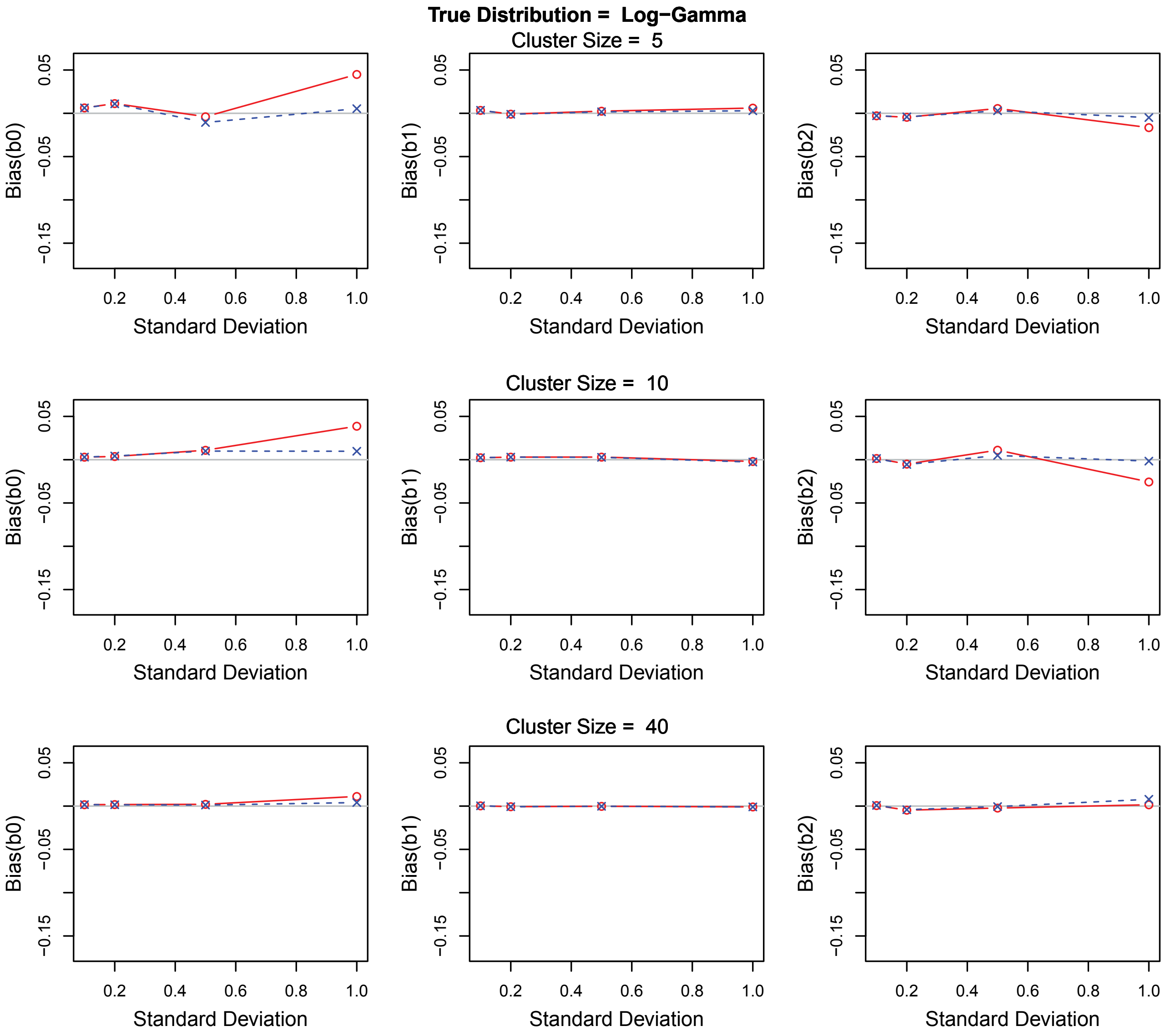
3.4 Discussion

The simulation results showed that the performance of Poisson GLMMs vs. CGLMMs is generally quite comparable, except for a few extreme cases when the true distribution is exponential and the standard deviation is large. In fact, when the standard deviation of then random effects is small and/or the cluster size is reasonably large, both Poisson GLMMs and CGLMMs are nearly indistinguishable. The extreme cases can be avoided in practice if we focus our attention on finding a "good" set of covariates.



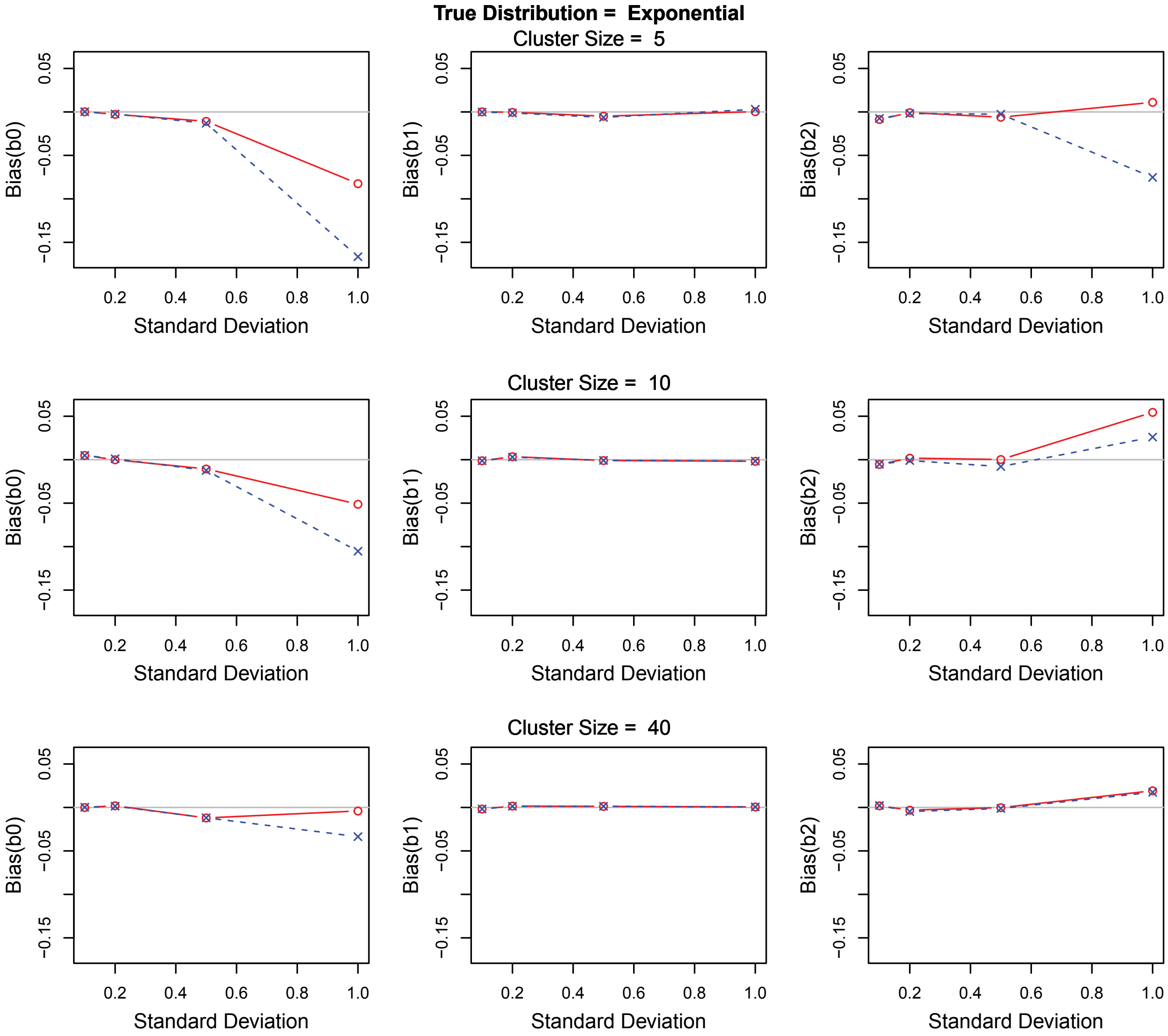
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.2 : Bias in fixed effects estimates from the Poisson GLMMs and CGLMMs when the true distribution is Gaussian. Each row is a different value of the cluster size and each plot shows the bias as median of the simulation estimate minus true value as a function of random effects standard deviation with separate curves for different assumed distributions.



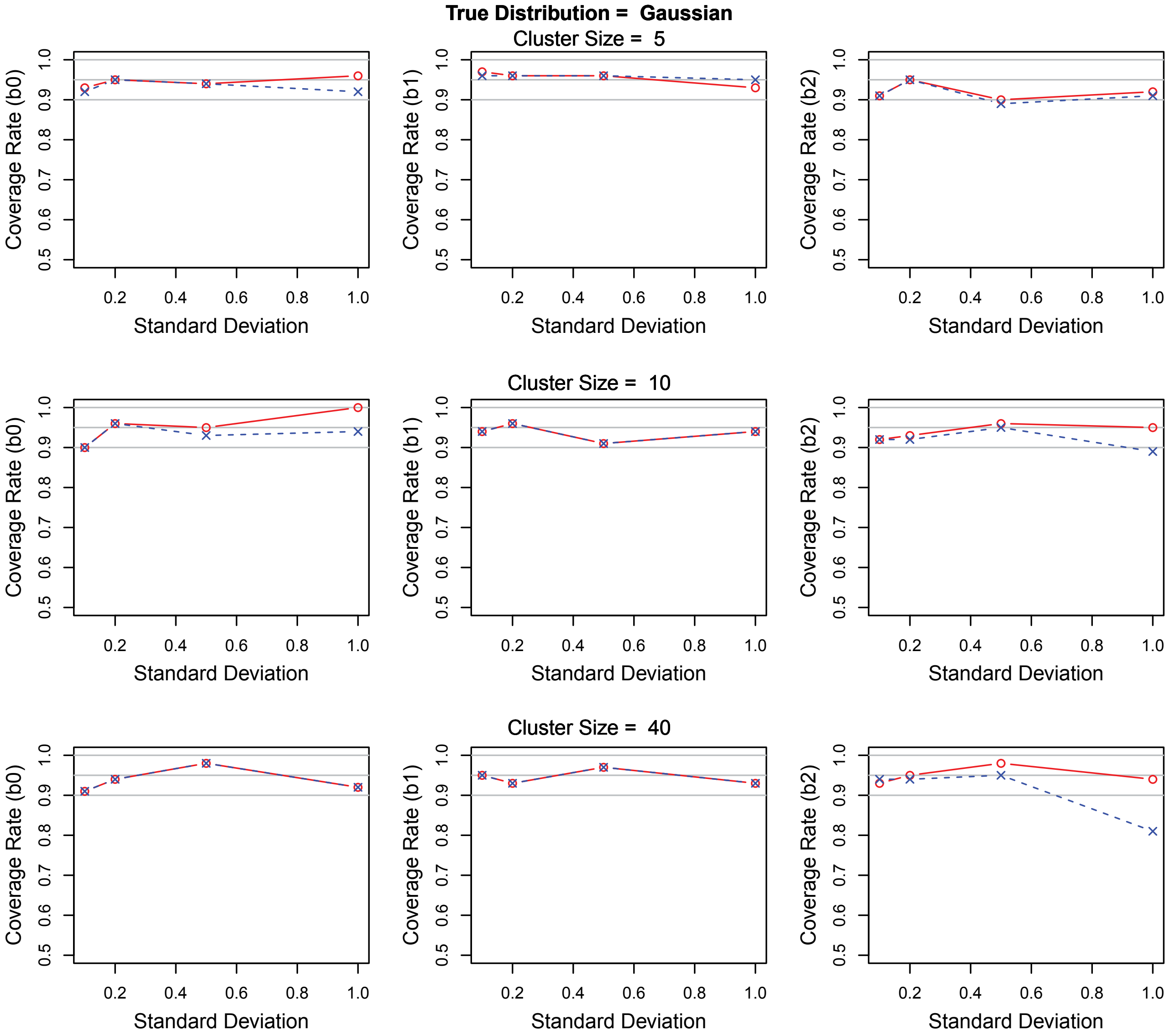
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.3 : Bias in fixed effects estimates from the Poisson GLMMs and CGLMMs when the true distribution is log-gamma. Each row is a different value of the cluster size and each plot shows the bias as median of the simulation estimate minus true value as a function of random effects standard deviation with separate curves for different assumed distributions.



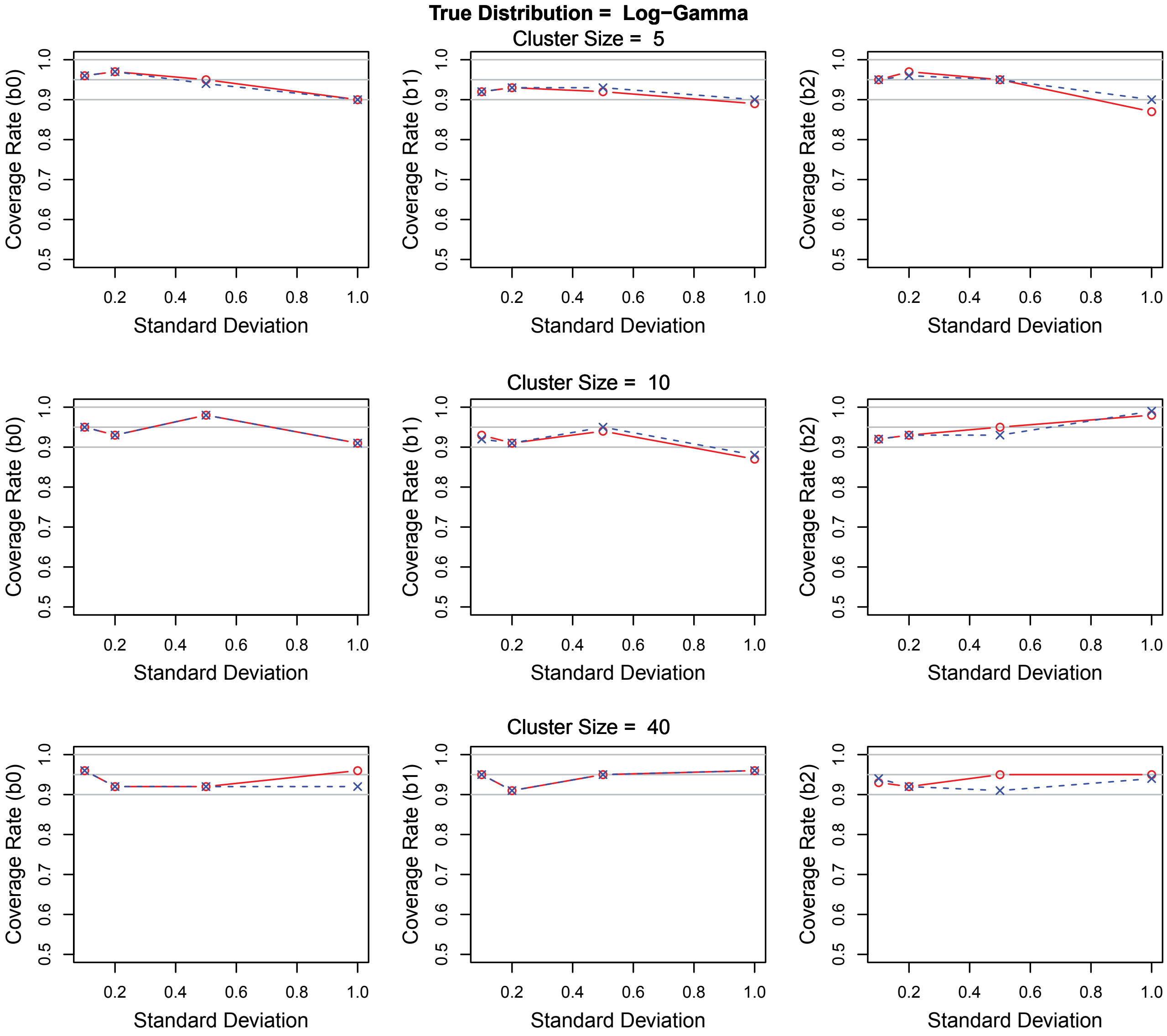
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.4 : Bias in fixed effects estimates from the Poisson GLMMs and CGLMMs when the true distribution is exponential. Each row is a different value of the cluster size and each plot shows the bias as median of the simulation estimate minus true value as a function of random effects standard deviation with separate curves for different assumed distributions.



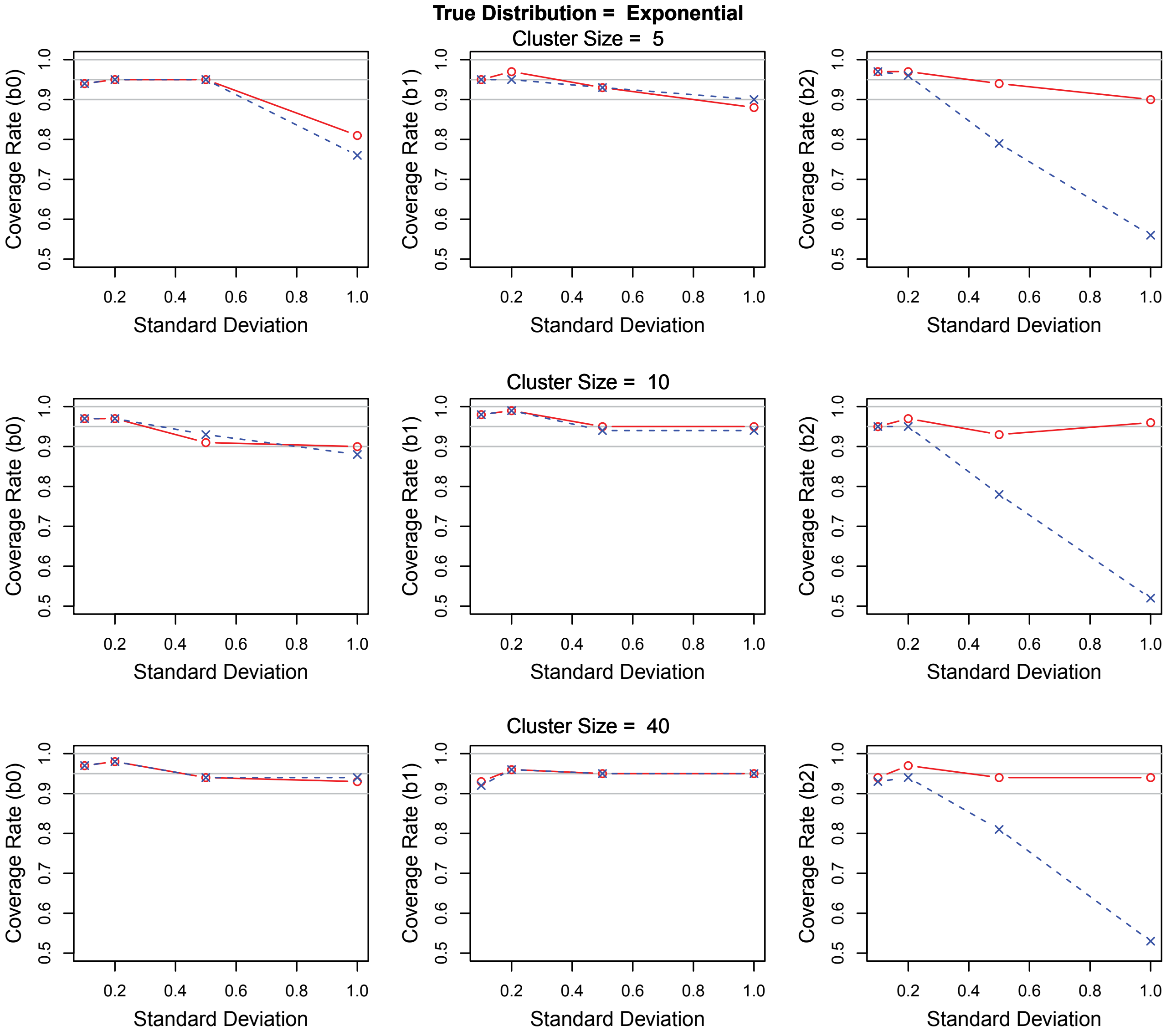
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.5 : Coverage rates for model-based 95% confidence intervals from the Poisson GLMMs and CGLMMs when the true distribution is Gaussian. Each row is a different value of the cluster size and each plot shows the coverage rate as a function of random effects standard deviation with separate curves for different assumed distributions. Horizontal lines at coverage rates = $0.90, 0.95$ and 1.00.



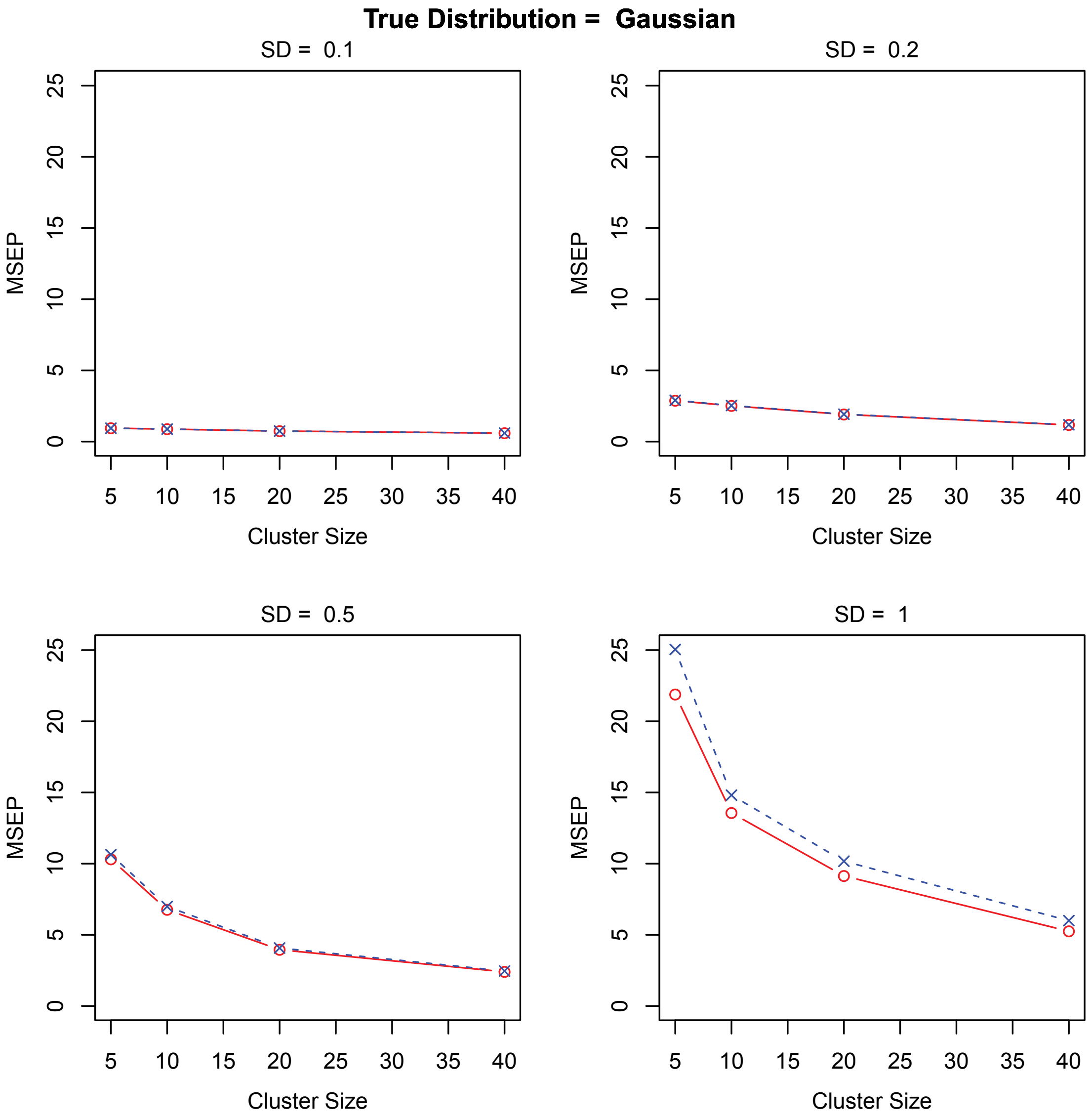
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.6 : Coverage rates for model-based 95% confidence intervals from the Poisson GLMMs and CGLMMs when the true distribution is log-gamma. Each row is a different value of the cluster size and each plot shows the coverage rate as a function of random effects standard deviation with separate curves for different assumed distributions. Horizontal lines at coverage rates $= 0.90, 0.95$ and 1.00.



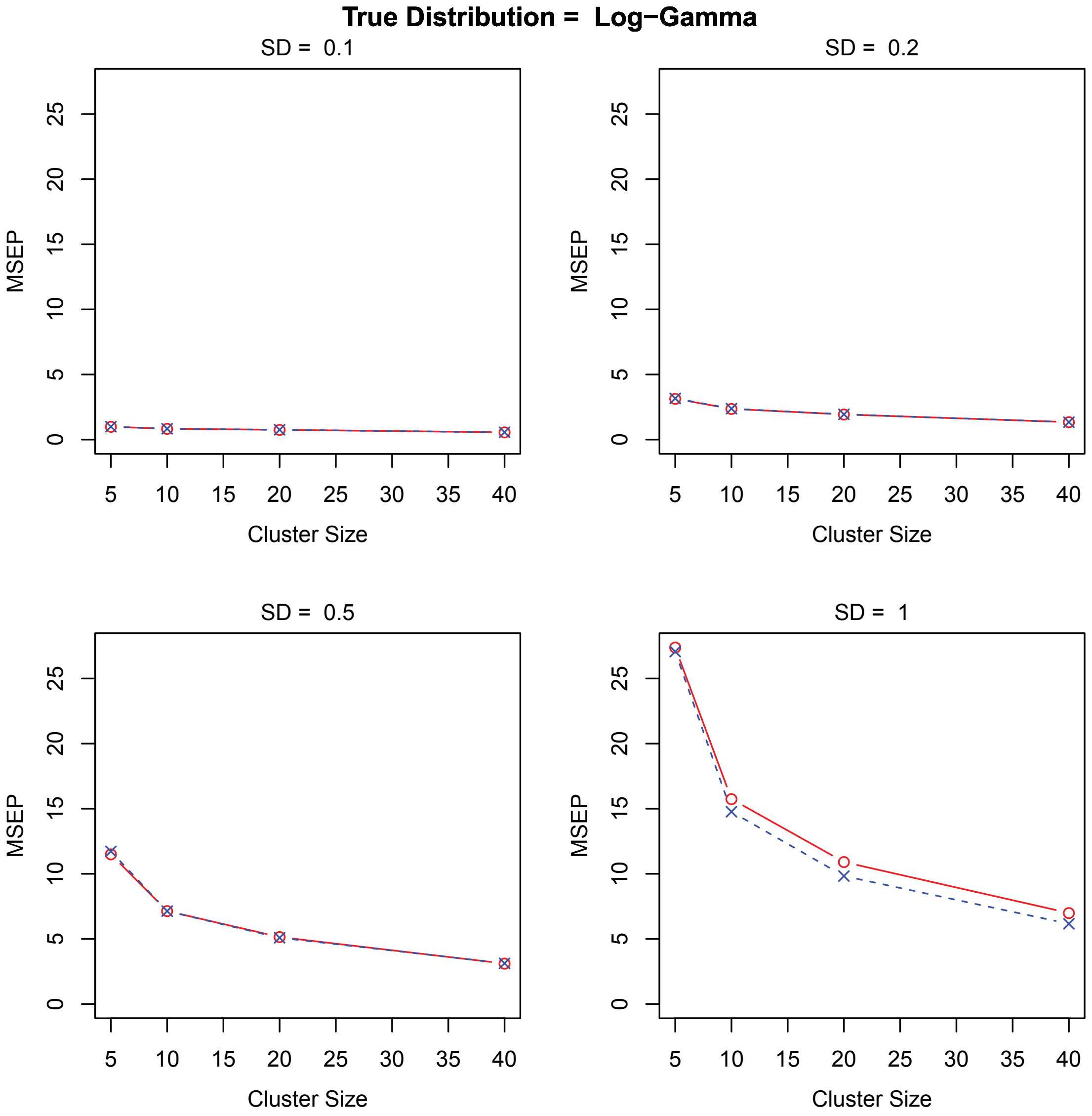
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.7 : Coverage rates for model-based 95% confidence intervals from the Poisson GLMMs and CGLMMs when the true distribution is exponential. Each row is a different value of the cluster size and each plot shows the coverage rate as a function of random effects standard deviation with separate curves for different assumed distributions. Horizontal lines at coverage rates $= 0.90, 0.95$ and 1.00.



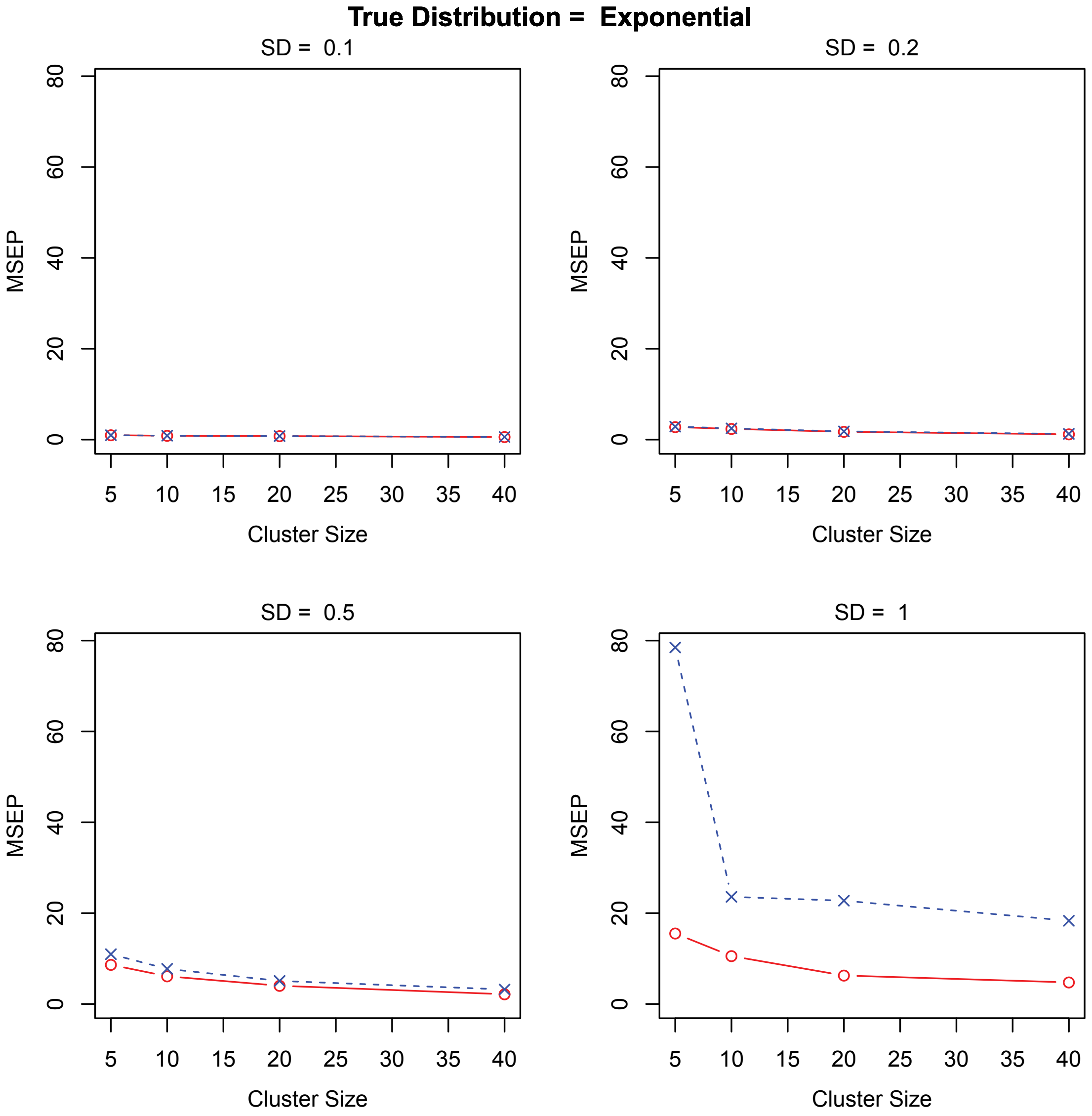
Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.8 : Empirical mean squared error of prediction (MSEP) from the Poisson GLMMs and CGLMMs when the true distribution is Gaussian. Each plot shows the MSEP as a function of cluster size with separate curves for different assumed distributions. Each panel is a different value of the random effects standard deviation.



Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.9 : Empirical mean squared error of prediction (MSEP) from the Poisson GLMMs and CGLMMs when the true distribution is log-gamma. Each plot shows the MSEP as a function of cluster size with separate curves for different assumed distributions. Each panel is a different value of the random effects standard deviation.



Assumed distributions: Dash/X = Log-Gamma, Solid/Circle = Gaussian

Figure 3.10 : Empirical mean squared error of prediction (MSEP) from the Poisson GLMMs and CGLMMs when the true distribution is exponential. Each plot shows the MSEP as a function of cluster size with separate curves for different assumed distributions. Each panel is a different value of the random effects standard deviation.

Chapter 4

On the "Poisson Trick" and its Extensions for Fitting Multinomial Regression Models

Summary

This chapter is concerned with the fitting of multinomial regression models using the so-called "Poisson Trick". The work is motivated by Chen and Kuo (2001) and Malchow-Møller and Svarer (2003) which have been criticized for being computationally inefficient and sometimes producing nonsense results. We first discuss the case of independent data and offer a parsimonious fitting strategy when all covariates are categorical. We then propose a new approach for modelling correlated responses based on an extension of the Gamma-Poisson model, where the likelihood can be expressed in closed-form. The parameters are estimated via an Expectation/Conditional Maximization (ECM) algorithm, which can be implemented using functions for fitting generalized linear models readily available in standard statistical software packages. Compared to existing methods, our approach avoids the need to approximate the intractable integrals and thus the inference is exact with respect to the approximating Gamma-Poisson model. The proposed method is illustrated via a reanalysis of the yogurt data discussed by Chen and Kuo (2001).

Keywords: Discrete choice model; Longitudinal data; Mixed logit model; Multinomial mixed model; Nominal polytomous data; Unobserved heterogeneity.

4.1 Introduction

Data with correlated categorical responses arise frequently in applications. This may arise from units grouped into clusters (clustered data) or multiple measurements taken on the same unit (longitudinal data). For instance, we might expect the unemployment outcomes (employed, unemployed, not in labour force) of residents living in the same region to be correlated, due to similar job opportunities and socioeconomic levels. Ignoring the correlation structure and assuming that all observations are independent by fitting an ordinary multinomial regression model may result in biased estimates and inaccurate predictions. Multinomial mixed models can account for correlation by using group level random effects (Daniels and Gatsonis, 1997; Hartzel et al., 2001; Hedeker, 2003).

For multinomial mixed models, it is a common practice to assume a multivariate normal distribution for the random effects. The multivariate normal distribution is easy to interpret and is convenient when we want to build more complicated correlation structures into our model. However, the resulting likelihood involves multidimensional integrals that cannot be solved analytically. The computational effort to evaluate the likelihood increases with the number of groups and categories, making it not suitable for large scale applications. In fact, Lee et al. (2017a) showed that closed-form likelihoods for multinomial mixed models do not exist regardless of the random effect distribution, except for the special case when there are no covariates.

Various methods have been proposed to circumvent the computational obstacle for fitting multinomial mixed models. Among them are quadrature (Hartzel et al., 2001; Hedeker, 2003), Monte Carlo EM algorithm, pseudo-likelihood approach (Hartzel et al., 2001) and Markov Chain Monte Carlo methods (Daniels and Gatsonis, 1997).

Jain et al. (1994) proposed a random effects estimation approach using a discrete probability distribution approximation. Simulation based methods such as method of simulated moments (McFadden, 1989) and method of simulated maximum likelihood (Gong et al., 2004; Hann and Uhlendorff, 2006) are widely used in the econometrics literature. Recently, Perry (2017) proposed a fast moment-based estimation method that scales well for large samples and which arguably can be extended for fitting multinomial mixed models. Kuss and McLerran (2007) used the fact that the multinomial model is a multivariate binary model and exploited a procedure proposed by Wright (1998) for model fitting. Their approach has been criticized by de Rooij and Worku (2012) as they failed to realize that a multivariate link function is needed in the context of multinomial models. An alternative strategy using clustered bootstrap was subsequently proposed by de Rooij and Worku (2012). Although some authors have considered the Dirichlet-multinomial model that results in a closed-form likelihood, it does not allow the incorporation of individual level covariates.

Chen and Kuo (2001) advocate using Poisson log-linear or non-linear mixed models, both with random effects, as surrogates to multinomial mixed models. Their method capitalizes on existing mixed models software packages for fitting generalized linear models with random effects. This allows multinomial mixed models to be fitted using an approximate likelihood from the Poisson surrogate models. Their results are based on extensions of the well known “Poisson Trick” (Baker, 1994; McCullagh and Nelder, 1989; Venables and Ripley, 2002) that relates multinomial models with Poisson models via a respecification of the model formulae. Although clever, their methods have been criticized for being computationally inefficient (Malchow-Møller and Svarer, 2003; Kuss and McLerran, 2007) and sometimes producing nonsense results (Kuss and McLerran, 2007). This might be due to the intractable likelihoods of their models and the various approximation methods being used in different software packages. The considerable execution time is especially problematic, where it can take up to months to fit the model to a moderate-sized dataset (Malchow-Møller and Svarer, 2003)! In this chapter, we propose a new approach based on an extension of the Gamma-Poisson model (Lee et al., 2017), where the likelihood can be expressed in closed-form. Using the proposed estimation procedure, the parameters can be estimated via readily available packages for fitting generalized linear models.

The remaining chapter is organized as follows. Section 4.2 reviews the "Poisson Trick" for multinomial regression models with independent responses, and suggests a parsimonious fitting strategy when all covariates are categorical. In Section 4.3 we propose a new approach for approximating the likelihood of multinomial regression models with random effects. The empirical performance of the proposed model is demonstrated via a simulation study and a reanalysis of the yogurt brand choice dataset as discussed by Chen and Kuo (2001) in Section 5.4. Finally, we conclude with a discussion and a summary of our findings in Section 5.5.

4.2 "Poisson Trick" for Independent Multinomial Responses

This section describes the relationship between multinomial and Poisson regression models for independent responses, which we refer to as the "Poisson Trick". The results are based on the well known fact that given the sum, Poisson counts are jointly multinomially distributed (McCullagh and Nelder, 1989).

4.2.1 Derivation

Let $Y_j = (Y_{jq})_{q=1}^{Q}$ be the $Q \times 1$ response vector for observation j with the corresponding probability vector $p_j = (p_{jq})_{q=1}^{Q}$, where q indexes the multinomial category. A common approach to satisfy the two characteristics of probability (i) $0 \leq p_{jq} \leq 1$ for all j and q; and (ii) $\sum_{q=1}^{Q} p_{jq} = 1$ is via

$$p_{jq} = \zeta_{jq}/\zeta_{j+}, \tag{4.1}$$

where ζ_{jq} is a positive user-specified function of covariates x and fixed effects γ, and $\zeta_{j+} = \sum_{q=1}^{Q} \zeta_{jq}$. Depending on the type of variable under consideration, x and γ can be indexed by various combinations of j and q (Croissant, 2013, pp. 7-8). Conditional on the multinomial sums $Y_{j+} = \sum_{q=1}^{Q} Y_{jq}$, $Y_j s$ are independently multinomially distributed for each j, i.e.

$$Y_j | Y_{j+} \sim \mathcal{M}\left(Y_{j+}, p_j\right). \tag{4.2}$$

In multinomial models, $Y_{j+} = y_{j+}$ is treated as fixed. Suppose we instead treat Y_{j+} as random and assume

$$Y_{j+} \sim \mathcal{P}(\delta_j \zeta_{j+}), \tag{4.3}$$

independently for each j. This results in a multinomial-Poisson mixture with the following joint probability function for each j:

$$\begin{aligned} \mathrm{P}(Y_j = y_j \cap Y_{j+} = y_{j+}) &= \mathrm{P}(Y_{j+} = y_{j+})\mathrm{P}(Y_j = y_j | Y_{j+} = y_{j+}) \\ &= e^{-\delta_j \zeta_{j+}} \frac{(\delta_j \zeta_{j+})^{y_{j+}}}{y_{j+}!} \times \frac{y_{j+}!}{\prod_q y_{jq}!} \prod_q \left(\frac{\zeta_{jq}}{\zeta_{j+}}\right)^{y_{jq}} \\ &= \prod_q \left\{ \frac{e^{-\delta_j \zeta_{jq}} \left(\delta_j \zeta_{jq}\right)^{y_{jq}}}{y_{jq}!} \right\} \quad \text{iff} \quad Y_{j+} = y_{j+}. \end{aligned} \tag{4.4}$$

The marginal probability of Y_j can then be obtained by summing the joint probability over all possible values of Y_{j+}:

$$\begin{aligned} \mathrm{P}(Y_j = y_j) &= \sum_{Y_{j+}=0}^{\infty} \prod_q \left\{ \frac{e^{-\delta_j \zeta_{jq}} (\delta_j \zeta_{jq})^{y_{jq}}}{y_{jq}!} \right\} \\ &= \prod_q \left\{ \frac{e^{-\delta_j \zeta_{jq}} (\delta_j \zeta_{jq})^{y_{jq}}}{y_{jq}!} \right\}. \end{aligned} \tag{4.5}$$

Thus, allowing the multinomial sums to be random according to a Poisson distribution results in

$$Y_{jq} \sim \mathcal{P}(\delta_j \zeta_{jq}), \tag{4.6}$$

independently for each j and q. Summing over all observations, the log-likelihood is

$$\sum_j \ell^{\mathcal{P}}(\delta_j \zeta_{j+}; Y_{j+}) + \sum_j \ell^{\mathcal{M}}(\zeta_j; Y_j|Y_{j+}) = \sum_j \sum_q \ell^{\mathcal{P}}(\delta_j \zeta_{jq}; Y_{jq}), \tag{4.7}$$

where $\ell^{\mathcal{P}}$ and $\ell^{\mathcal{M}}$ denote the Poisson and multinomial log-likelihood functions respectively, and $\zeta_j = (\zeta_{jq})_{q=1}^{Q}$. The second term on the left hand side is the model we would like to fit, and the term on the right hand side is the model we actually fit.

To show that the Poisson surrogate model is an exact fit to the multinomial model, first note the log-likelihood corresponding to the multinomial model is

$$\sum_j \log(y_{j+!}) - \sum_j \sum_q \log(y_{jq}!) + \sum_j \sum_q y_{jq} \log \zeta_{jq} - \sum_j y_{j+} \log \zeta_{j+}, \tag{4.8}$$

and the log-likelihood of the Poisson surrogate model is

$$-\sum_j \delta_j \zeta_{j+} + \sum_j y_{j+} \log \delta_j + \sum_j \sum_q y_{jq} \log \zeta_{jq} - \sum_j \log(y_{j+}!). \tag{4.9}$$

Differentiating Equation 4.9 with respect to δ_j and setting it to 0, we obtain $\hat{\delta}_j = y_{j+}/\zeta_{j+}$. Plugging in the maximizing value of δ_j into Equation 4.9, we have

$$-\sum_j y_{j+} + \sum_j y_{j+} \log y_{j+} - \sum_j y_{j+} \log \zeta_{j+} + \sum_j \sum_q y_{jq} \log \zeta_{jq}. \tag{4.10}$$

Equation 4.10 is identical to Equation 4.8, up to an additive constant. It follows that the maximum likelihood estimates, their asymptotic variances and tests for the fixed effects can be *exactly* recovered under the Poisson surrogate model (Richards, 1961). That is, likelihood inference for ζ_{jq} is the same whether we regard Y_{j+} as fixed (multinomial) or randomly sampled from independent Poissons. This result applies to the fixed effects model, with any parameterization of ζ_{jq}, including:

- Exponential transformations of linear combinations of *categorical variables* and regression coefficients (McCullagh and Nelder, 1989; Agresti, 2013),
- Exponential transformations of linear combinations of *continuous variables* and regression coefficients,
- *Any* monotonic transformations of linear combinations of covariates and regression coefficients,
- Nonlinear functions of covariates and regression coefficients,
- Nonparametric formulations.

The Poisson surrogate model eliminates ζ_{j+} from the denominator of the multinomial probabilities. This makes sense intuitively, as we do not expect the multinomial sums to provide any useful information in estimating the fixed effects. Given that $\hat{\delta}_j$ can also be obtained by setting the fitted values of the multinomial sums $\hat{Y}_{j+} = \mathrm{E}(Y_{j+})$ equal to the observed counts y_{j+} in the Poisson surrogate model, δ_j has the effect of recovering the multinomial sums. The key idea is to include a separate constant δ_j for each unique combination of covariates in the Poisson surrogate models.

4.2.2 Specifying Model Formulae for Poisson Surrogate Models

For purposes of exposition, the model formulae in this section are written in terms of the R language (R Development Core Team, 2017), although this chapter is not concerned with software packages per se. Multinomial models are fitted using the multinom() function within the nnet package (Ripley and Venables, 2016); Poisson models are fitted using the glm() function within the stats package.

For concreteness, consider the *non-parallel baseline category logit models*. The "baseline category logit" assumption refers to the following: treating category 1 as the baseline category with $\zeta_{j1} = 1 \; \forall j$ without loss of generality, we model $\log(p_{jq}/p_{j1})$ $= \log(\zeta_{jq})$ as a linear function of covariates x and regression coefficients γ. This assumption is not necessary, but chosen so that the model formulae can be illustrated using functions within the stats package. The "non-parallel" assumption refers to covariate effects that vary across categories (Fullerton and Xu, 2016), i.e. all elements of the γ vector are indexed by q. That is, if the set of logits are plotted against the covariate on the same graph, a set of straight lines with slopes that are in general different will be obtained. Later we shall discuss cases where we relax this assumption.

Consider a hypothetical dataset with two predictors X_1 and X_2 (these can be categorical or continuous) and a multinomial outcome vector Y with $Q = 3$ categories. In *short format*, each row of data represents an observation with a 3-dimensional outcome vector (Y_1, Y_2, Y_3). Poisson models treat the outcomes of each observation as independent and glm() requires data to be presented in *long format*. This requires an additional factor C that denotes the category memberships. Each row now comprises a scalar outcome, resulting in 3 rows of data per observation. The first few rows of data in both short and long format are shown in Table 4.1.

Table 4.1 : Multinomial data.

(a) Short format.

Obs	X_1	X_2	Y_1	Y_2	Y_3
1	0	0	3	5	2
2	0	1	5	5	0
3	1	0	7	2	1
4	1	1	1	3	6
–	–	–	–	–	–

(b) Long format.

I	X_1	X_2	C	Y
1	0	0	1	3
1	0	0	2	5
1	0	0	3	2
2	0	1	1	5
2	0	1	2	5
2	0	1	3	0
–	–	–	–	–

Table 4.2 shows the equivalant relationship between non-parallel multinomial models and the corresponding Poisson models, where the parameters satisfy the usual constraints for identifiability. The Poisson surrogate models possess several important features:

1. The model includes an indicator variable I (that corresponds to $\log \delta_j$ in Section 4.2.1) for each observation, although this can be simplified when all covariates are categorical. This is to ensure the exact recovery of the multinomial sums, as the fixed sums are treated as random in the Poisson models. As a result, we do not interpret the coefficients of I since they are just nuisance parameters.

2. The category membership indicator C enters as a covariate in the Poisson models, where the coefficients correspond to the intercepts in the multinomial modelsv so that the counts are allowed to vary by category.

3. The model includes interaction terms between X and C (denoted by $*$ in

the model formula), where the coefficients correspond to the slopes in the multinomial models. This is due to the non-parallel assumption where each category has a separate slope, and also the fact that multinomial models treat the response counts jointly for each observation, whereas Poisson models treat each response count as a separate observation. It is important that these interaction terms are included even if they are not significant. For multinomial models where some (*partial models*) or all (*parallel models*) of the covariate effects do not vary across categories, the equivalent Poisson models can be obtained by modifying the interaction structure between X and C accordingly. For instance, in parallel models where all categories share the same covariate effects, there is no need to include the interation terms between X and C, since the slopes do not vary across categories.

Table 4.2 : Equivalent relationship between non-parallel multinomial models and the corresponding Poisson models.

Multinomial[1]	**Poisson**[2]
$Y \sim 1$	$Y \sim I + C$
$Y \sim X_1$	$Y \sim I + C + C * X_1$
$Y \sim X_1 + X_2$	$Y \sim I + C + C * X_1 + C * X_2$
$Y \sim X_1 + X_2 + X_1 * X_2$	$Y \sim I + C + C * X_1 + C * X_2 + C * X_1 * X_2$

[1] Syntax for using multinom() in R, where data are presented in short format and Y is a vector of response counts.

[2] Syntax for using glm() in R, where data are presented in long format and Y is a scalar response count.

When writing the model formula, it is important to specify I and C as *factors* due to their categorical nature. This can be achieved via the factor() function in R.

Special Case: Categorical Covariates

When all covariates are categorical, the model formulae in Table 4.3 offer a more parsimonious option for fitting the Poisson models without having to estimate a separate parameter for each observation.

Table 4.3 : Equivalent relationship between non-parallel multinomial models and the corresponding Poisson models, when all the covariates are categorical.

Multinomial[1]	**Poisson**[2]
$Y \sim 1$	$Y \sim C$
$Y \sim X_1$	$Y \sim X_1 + X_1 * C$
$Y \sim X_1 + X_2$	$Y \sim X_1 * X_2 + X_1 * C + X_2 * C$
$Y \sim X_1 + X_2 + X_1 * X_2$	$Y \sim X_1 * X_2 * C$

[1] Syntax for using multinom() in R, where data are presented in short format and Y is a vector of response counts.

[2] Syntax for using glm() in R, where data are presented in long format and Y is a scalar response count.

As stated above, the key to achieving the 1-1 correspondence between multinomial and Poisson models (with the same link function) is to include a separate constant for each unique combination of covariates. For categorical covariates, this can be achieved by including the full interaction among the predictors in the Poisson model. When all the covariates are categorical, the interaction term has the precise effect of pooling groups of observations with identical covariates. Fitting such models is equivalent to fitting the observation index as a factor (Table 4.2), but the pooling results in a smaller effective data frame, and therefore smaller storage requirements and faster fitting speed, with no loss of information. Of course, if there are many factors, there may not be much saving, because it will be comparatively rare for different observations to have all the same factor level combinations.

4.3 Extending the "Poisson Trick" for Correlated Multinomial Responses

4.3.1 Derivation

Consider a set of observations which fall into a collection of I groups and let $\lambda_i = (\lambda_{iq})_{q=1}^{Q}$ be a vector-valued random effect for group i. Each observation belongs to only a single group. Extending the notation in Section 4.2, the $Q \times 1$ response vector for observation j in group i is $Y_{ij} = (Y_{ijq})_{q=1}^{Q}$, with the corresponding probability vector $p_{ij} = (p_{ijq})_{q=1}^{Q}$, where $p_{ijq} = \lambda_{iq}\zeta_{ijq} / \sum_{q=1}^{Q} \lambda_{iq}\zeta_{ijq}$. Conditional on the multinomial sums $Y_{ij+} = \sum_{q=1}^{Q} Y_{ijq}$ and the random effects λ_i, the counts are multinomially distributed:

$$Y_{ij}|Y_{ij+}, \lambda_i \sim \mathcal{M}\left(Y_{ij+}, p_{ij}\right). \tag{4.11}$$

In analogy to the results in Section 4.2, given the random effects, we treat Y_{ij+} as random and assume

$$Y_{ij+}|\lambda_i \sim \mathcal{P}\left(\delta_{ij} \sum_{q=1}^{Q} \lambda_{iq}\zeta_{ijq}\right), \tag{4.12}$$

independently for each i and j. This gives

$$Y_{ijq}|\lambda_{iq} \sim \mathcal{P}(\delta_{ij}\lambda_{iq}\zeta_{ijq}), \tag{4.13}$$

independently for each j and q. The probability argument in Equation 4.7 still holds, now conditional on the random effects:

$$\sum_i \sum_j \ell^{\mathcal{P}} \left(\delta_{ij} \sum_{q=1}^{Q} \lambda_{iq} \zeta_{ijq}; Y_{ij+} | \lambda_i \right) + \sum_i \sum_j \ell^{\mathcal{M}} \left(\zeta_{ij}; Y_{ij} | Y_{ij+}, \lambda_i \right)$$
$$= \sum_i \sum_j \sum_q \ell^{\mathcal{P}} \left(\delta_{ij} \lambda_{iq} \zeta_{ijq}; Y_{ijq} | \lambda_{iq} \right), \quad (4.14)$$

where $\zeta_{ij} = (\zeta_{ijq})_{q=1}^{Q}$. If the random effects are observed, the conditional probability statement above imply a 1-1 exact correspondence between the multinomial and the Poisson surrogate models. However, due to the unobserved nature of the random effects, interest lies in the marginal distribution, obtained by integrating out the random effects. This results in an approximate relationship between the two models. It turns out that the marginal likelihood of the approximating Poisson surrogate model (right hand side of Equation 4.14) can be expressed in closed-form if we assume an independent gamma model for the random effects, with $\mathrm{E}(\lambda_{iq}) = \alpha_q / \beta_q$ and $\mathrm{Var}(\lambda_{iq}) = \alpha_q / \beta_q^2$, i.e. $\lambda_{iq} \sim \mathcal{G}(\alpha_q, \beta_q)$ (Lee et al., 2017a).

With this assumption for the distribution of the random effects, the marginal likelihood of the multinomial model that we would like to fit (second term on the left hand side of Equation 4.14) is given by

$$L^{\mathcal{M}} = \prod_i \int \cdots \int \prod_j \left\{ \frac{y_{ij+}!}{\prod_q y_{ijq}!} \prod_q \left(\frac{\lambda_{iq} \zeta_{ijq}}{\sum_{q=1}^{Q} \lambda_{iq} \zeta_{ijq}} \right)^{y_{ijq}} \right\} \times \prod_q \frac{\beta_q^{\alpha_q} \lambda_{iq}^{\alpha_q - 1} e^{-\beta_q \lambda_{iq}}}{\Gamma(\alpha_q)} \, \mathrm{d}\lambda_{i1} \ldots \mathrm{d}\lambda_{iQ}. \quad (4.15)$$

This does not generally exhibit a closed-form solution regardless of the random effect distribution, unless in the special cases of no covariate or with only group specific covariates (Lee et al., 2017a). Numerical or simulation methods can be used to approximate the likelihood, with computational efforts increasing with increasing number of groups and categories. On the other hand, the marginal likelihood of the

Poisson surrogate model can be expressed in closed-form:

$$\begin{aligned} L^P &= \prod_i \left\{ \int \cdots \int \prod_j \prod_q \frac{e^{-\delta_{ij}\lambda_{iq}\zeta_{ijq}} (\delta_{ij}\lambda_{iq}\zeta_{ijq})^{y_{ijq}}}{y_{ijq}!} \prod_q \frac{\beta_q^{\alpha_q} \lambda_{iq}^{\alpha_q - 1} e^{-\beta_q \lambda_{iq}}}{\Gamma(\alpha_q)} \, \mathrm{d}\lambda_{i1} \ldots \mathrm{d}\lambda_{iQ} \right\} \\ &= \prod_i \left\{ \prod_q \frac{\Gamma(\alpha_q + y_{i+q})\beta_q^{\alpha_q}}{\Gamma(\alpha_q)(\beta_q + \sum_j \delta_{ij}\zeta_{ijq})^{\alpha_q + y_{i+q}}} \times \prod_j \prod_q \frac{(\delta_{ij}\zeta_{ijq})^{y_{ijq}}}{y_{ijq}!} \right\}. \end{aligned} \tag{4.16}$$

The Poisson surrogate model is an extension of the gamma-Poisson model as proposed by Lee et al. (2017) and Lee et al. (2017a) to allow the modelling of counts for multiple categories.

As a consequence of Equation 4.16, we have

$$\mathrm{E}(Y_{ijq}) = \frac{\alpha_q}{\beta_q} \delta_{ij}\zeta_{ijq}. \tag{4.17}$$

Refer to the appendix for details. This is the population-averaged expected value and is not suitable for prediction in general, as it does not take into account the cluster effect. However, it can be useful for out of sample prediction, when there are no samples present in a particular group.

Special Case: Var($\boldsymbol{\lambda_{iq}}$) approaches 0

When $\mathrm{Var}(\lambda_{iq})$ approaches 0 for all q, the model reduces to the special case of no random effects as outlined in Section 4.2, and the exact correspondence between the multinomial and the Poisson models can be regained.

4.3.2 Identifiability

There is some lack of identifiability with the model formulation given by Equation 4.16, characterized by non-uniqueness of the maximum likelihood estimates. There is an identifiability issue between λ_{iq} and δ_{ij}, and also between λ_{iq} and the category intercepts. To fix this, we impose the constraint of $\alpha_q = 1/\beta_q$ so that $\mathrm{E}(\lambda_{iq}) = 1$. As a consequence, $\lambda_{iq} \sim \mathcal{G}(1/\beta_q, \beta_q)$ and $\mathrm{Var}(\lambda_{iq}) = \beta_q$. Also, we only require a random effect for each logit, and thus a constraint for the random effects associated with the baseline category $q = 1$ is needed. Denote $u_{iq} = \log \lambda_{iq}$. Several authors such as Agresti (2013) (pp.514) and Hartzel et al. (2001) considered a multivariate normal distribution for the random effects, i.e. $(u_{iq})_{q=2}^{Q} \sim \mathcal{N}(0, \Sigma)$, where Σ is a $Q - 1$ by $Q - 1$ variance-covariance matrix. This is equivalent to saying that $u_{i1} = 0$ for all i, or $\sigma_{11} = 0$. The equivalent statement in our proposed model is to fix $\lambda_{i1} = 1$ for all i. This is tantamount to saying $\mathrm{Var}(\lambda_{i1}) = \beta_1$ approaches 0, and thus α_1 approaches ∞.

4.3.3 Prediction of Random Effects and Fitted Values

We focus on the *best predictor* (BP) for random effects prediction, i.e. the predictor that minimises the overall mean squared error of prediction. McCulloch et al. (2008) shows that the BP is given by the posterior expectation of the random effect. Under the proposed Poisson surrogate model, the BP is given by

$$\mathrm{BP}(\lambda_{iq}) = \hat{\lambda}_{iq} \equiv \underset{\lambda^\star}{\mathrm{argmin}}\ \mathrm{E}(\lambda_{iq} - \lambda^\star)^2 := \mathrm{E}(\lambda_{iq}|y), \tag{4.18}$$

which can be calculated via

$$\hat{\lambda}_{iq} = \frac{\displaystyle\int_{-\infty}^{\infty} \lambda_{iq} f(\lambda_{iq}) f(y|\lambda_{iq})\ d\lambda_{iq}}{\displaystyle\int_{-\infty}^{\infty} f(\lambda_{iq}) f(y|\lambda_{iq})\ d\lambda_{iq}}\ . \tag{4.19}$$

Solving for the integral, the BP is

$$\hat{\lambda}_{iq} = \frac{Y_{i+q} + 1/\beta_q}{\sum_j \delta_{ij}\zeta_{ijq} + \beta_q}, \tag{4.20}$$

where $Y_{i+q} = \sum_j Y_{ijq}$. $\hat{\lambda}_{iq}$ depends on the parameters δ_{ij}, γ and β_q, in which we replace by their estimators, leading to the *empirical best predictor* (EBP). The fitted values can then be defined as

$$\hat{Y}_{ijq} = \delta_{ij} \hat{\lambda}_{iq} \zeta_{ijq}, \tag{4.21}$$

where we replace δ_{ij} and ζ_{ijq} by their respective estimators $\hat{\delta}_{ij}$ and $\hat{\zeta}_{ijq}$.

4.3.4 Parameter Estimation

Consider the parameterization $\zeta_{ijq} = \exp(\eta_{ijq})$ which is widely adopted in practice, where $\eta_{ijq} = x_{ijq}^T \gamma$, where x_{ijq} and γ are both vectors. The chosen index structure for x and γ encompasses a variety of possible scenarios: (i) category-specific predictors with generic coefficients $x_{ijq}^T \gamma$, (ii) category-specific predictors with category-specific coefficients $x_{ijq}^T \gamma_q$, and (iii) observation-specific predictors with category-specific coefficients $x_{ij} \gamma_q$. This can be achieved by creating the appropriate interaction terms between the predictor and the category indicator variable, thus modifying the model matrix. Note that observation-specific predictors must be paired with choice-specific coefficients. Otherwise they will disappear in the differentiation when we consider the log-odds.

Denote $\theta = (\gamma, (\beta_q)_{q=2}^{Q})$, where γ includes the incidental parameters $\log(\delta_{ij})$ for all i and j. Algorithm 1 presents an Expectation/Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993) for parameter estimation of the Poisson surrogate model. Refer to the appendix for a detailed derivation.

4.4 Yogurt Brand Choice Dataset

We consider the yogurt brand choice dataset previously analyzed by Jain et al. (1994) and Chen and Kuo (2001). Jain et al. (1994) approximated the likelihood of a

```
Initialize $\theta$.
Cycle:
while relative differences in the parameter estimates are not negligible do
    E-Step: Calculate for each $i$ and $q$:
        $\hat{\lambda}_{iq}^{(t+1)} = \mathrm{E}\left(\lambda_{iq}\middle|(y_{ijq})_j, \theta^{(t)}\right) = \frac{y_{i+q} + 1/\beta_q^{(t)}}{\sum_j e^{x_{ijq}\gamma^{(t)}} + 1/\beta_q^{(t)}}$
        $\hat{\chi}_{iq}^{(t+1)} = \mathrm{E}\left(\log(\lambda_{iq})\middle|(y_{ijq})_j, \theta^{(t)}\right) = \psi\left(y_{i+q} + 1/\beta_q^{(t)}\right) - \log\left(\sum_j e^{x_{ijq}\gamma^{(t)}} + 1/\beta_q^{(t)}\right),$
        where $\psi(\cdot)$ is the digamma function.
    CM-Step:
        • Obtain $\gamma^{(t+1)}$ by fitting a Poisson log-linear model with $y_{ijq}$ as the response
          and $X_{ijq}$ as the design matrix, with $\hat{\lambda}_{iq}^{(t+1)}$ as offset. $X_{ijq}$ includes indicator
          variables for each unique combination of covariates.
        • Obtain $\beta_q^{(t+1)}$ for each $\beta_q$ for $q = 2$ to $Q$ by maximizing
              $\sum_i \left\{(1/\beta_q - 1)\hat{\chi}_{iq}^{(t+1)} - \hat{\lambda}_{iq}^{(t+1)}/\beta_q - \log(\beta_q)/\beta_q - \log\Gamma(1/\beta_q)\right\},$
          where $\Gamma(\cdot)$ is the gamma function.
end
```

Algorithm 1: *Expectation/Conditional Maximization (ECM) algorithm for fitting the Poisson surrogate model.*

multinomial logit model with Gaussian random effects using a discrete distribution. Chen and Kuo (2001) approximated the multinomial logit model using the Poisson log-linear model and Poisson nonlinear model, both with Gaussian random effects.

The dataset consists of purchases of yogurt by a panel of 100 households in

Springfield, Missouri, and were originally provided by A. C. Nielsen. The data were collected by optical scanners for about two years and correspond to $2,412$ purchases. Variables collected include brand, price and presence of newspaper feature advertisements for each purchase made by households in the panel. Price and feature advertisements are choice-specific variables. We assume a *parallel baseline logit* model by assigning generic coefficients γ to these variables, as we do not expect the effect of price and feature advertisements on the probability of purchase to vary according to brands. The four brands of yogurt: Yoplait, Dannon, Weight Watchers, and Hiland account for market shares of 34%, 40%, 23%, and 3% respectively. Following Chen and Kuo (2001), we put Hiland as the reference brand. Table 4.4 presents the yogurt data in both long and short format. The letters 'f' and 'p' represent the feature and price variables respectively, with the letter that follows denoting the brand. For instance, 'fy' stands for 'feature of Yoplait' and 'pd' stands for 'price of Dannon'.

We fit the models proposed in Sections 4.2 and 4.3, and compare our results to that of Chen and Kuo (2001), fitted using the SAS macro GLIMMIX and the SAS procedure NLMIXED. The results are presented in Table 4.5. The preference ordering of the brands are the same for all models, i.e. Yoplait is the most preferred brand, followed by Dannon, Weight Watchers and Hiland. The slope parameters estimates have the expected signs for all models. An increase in price is associated with a decrease in the probability of purchase. Feature advertisement tends to increase the chance of purchase. The household-to-household variation in the probability of purchase for Weight Watchers is much larger than the other brands, although none are significant ($p > 0.05$).

In comparing the estimates between models, we note that the fixed effects model is likely to produce biased estimates as it did not take into account of the correlation

Table 4.4 : Yogurt data.

(a) Short format.

id	obs	yoplait	dannon	weight	hiland	fy	fd	fw	fh	py	pd	pw	ph
1	1	0	1	0	0	0	0	0	0	0.108	0.081	0.079	0.061
1	2	0	1	0	0	0	0	0	0	0.108	0.098	0.075	0.064
1	3	0	1	0	0	0	0	0	0	0.108	0.098	0.086	0.061
–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	9	1	0	0	0	0	0	0	0	0.108	0.098	0.079	0.050
–	–	–	–	–	–	–	–	–	–	–	–	–	–
100	2412	0	0	1	0	0	0	0	0	0.108	0.086	0.079	0.043

(b) Long format.

id	obs	feature	price	count	brand
1	1	0	0.108	0	yoplait
1	1	0	0.081	0	dannon
1	1	0	0.079	1	weight
1	1	0	0.061	0	hiland
1	2	0	0.108	0	yoplait
1	2	0	0.098	1	dannon
–	–	–	–	–	–
100	2412	0	0.043	1	hiland

induced by multiple purchases from the same household. The parameter estimates of NLMIXED and Gamma-Poisson are uniformly larger than that of GLIMMIX, except for the intercept associated with Weight Watchers (for NLMIXED) and for the slope associated with price (for Gamma-Poisson). The estimates of the standard errors from NLMIXED and Gamma-Poisson are also uniformly larger than that of GLIMMIX. These differences can be attributed to the different distributional assumptions of the random effects, and also the different approximations used in GLIMMIX and NLMIXED to estimate the intractable likelihood. In this regard, our model exhibit a closed-form likelihood that allows exact inference to be performed with respect to the approximating model.

We tried to fit a simplified version of GLIMMIX using the glmer() function within the lme4 package in R, with just a random effect per household (ignoring the choice effect). However, the model failed to converge within a few months, even though Chen and Kuo (2001) claimed that the GLIMMIX model coverged in SAS.

4.5 Concluding Remarks

In this chapter, we presented methods for fitting various multinomial regression models via the so-called "Poisson Trick" and its extensions. The "Poisson Trick" for fitting fixed effects multinomial regression models is handy when the direct fitting of multinomial models is not supported, for instance the INLA package (Rue et al., 2009) in R. Murray (2017) used a related trick to derive efficient Markov Chain Monte Carlo sampler in the context of Bayesian additive regression trees for multinomial responses. For multinomial regression models with random effects, there exist a variety of experience for using the existing extensions proposed by Chen and Kuo (2001), from taking months to fit a moderate sized dataset (Malchow-

Table 4.5 : Regression estimates for the yogurt data, and the associated standard errors.

		Random Effects		
Variables	**Fixed Effects**[1]	**GLIMMIX**[2]	**NLMIXED**[3]	**Gamma-Poisson**[4]
Dannon	3.716 (0.145)	3.838 (0.231)	4.130 (0.648)	4.616 (0.309)
Weight	3.074 (0.145)	2.242 (0.241)	1.046 (0.671)	3.677 (0.392)
Yoplait	4.450 (0.187)	4.626 (0.261)	4.805 (0.699)	5.275 (0.342)
Feature	0.491 (0.120)	0.730 (0.121)	0.956 (0.185)	0.785 (0.178)
Price	-36.658 (2.437)	-40.012 (2.562)	-36.686 (3.725)	-40.881 (3.778)
β_{Dannon}	N/A	N/A	N/A	2.203 (0.134)
β_{Weight}	N/A	N/A	N/A	6.067 (0.374)
$\beta_{Yoplait}$	N/A	N/A	N/A	1.918 (0.135)

[1] Fitted using the glm() function in R, using the “Poisson Trick” as outlined in Section 4.2.

[2] Poisson log-linear model with Gaussian random effects, fitted by Chen and Kuo (2001) using the SAS macro GLIMMIX.

[3] Poisson nonlinear model with Gaussian random effects, fitted by Chen and Kuo (2001) using the SAS procedure NLMIXED.

[4] Poisson log-linear model with Gamma (multiplicative) random effects fitted using the ECM algorithm, as outlined in Section 4.3.

Møller and Svarer, 2003), producing nonsense results (Kuss and McLerran, 2007) to non-convergence in our experience of fitting the yogurt brand choice dataset. We proposed an extension of the “Poisson Trick” using Gamma (multiplicative) random effects. In contrast to the models by Chen and Kuo (2001), our model exhibits a closed-form likelihood and can be maximized using existing functions for fitting generalized linear models that are stable and heavily optimized, without having to approximate the integrals.

Appendix

Derivation of the Population-Averaged Expected Values in Equation 4.17

Equation 4.16 is also equivalent to

$$\prod_i \left\{ \prod_q \left[\frac{\Gamma(\alpha_q + y_{i+q})}{\Gamma(\alpha_q) y_{i+q}!} \left(\frac{\sum_j \delta_{ij}\zeta_{ijq}}{\beta_q + \sum_j \delta_{ij}\zeta_{ijq}} \right)^{y_{i+q}} \left(\frac{\beta_q}{\beta_q + \sum_j \delta_{ij}\zeta_{ijq}} \right)^{\alpha_q} \right] \times \right.$$
$$\left. \prod_q \left[\frac{y_{i+q}!}{\prod_j y_{ijq}!} \frac{\prod_j (\delta_{ij}\zeta_{ijq})^{y_{ijq}}}{(\sum_j \delta_{ij}\zeta_{ijq})^{y_{i+q}}} \right] \right\}. \quad (4.22)$$

This results in two different interpretations for the extended Gamma-Poisson surrogate model:

1. For each i and q, Y_{ijq} is independent negative multinomial $\mathcal{NM}\left(\alpha_q, \frac{\delta_{ij}\zeta_{ijq}}{\beta_q + \sum_j \delta_{ij}\zeta_{ijq}}\right)$ (Equation 4.16).
2. For each i and q, the category sums Y_{i+q} are independent Negative Binomial $\mathcal{NB}\left(\alpha_q, \frac{\sum_j \delta_{ij}\zeta_{ijq}}{\beta_q + \sum_j \delta_{ij}\zeta_{ijq}}\right)$, and conditional on the Y_{i+q}, Y_{ijq} is independent multinomial $\mathcal{M}\left(Y_{i+q}, \frac{\delta_{ij}\zeta_{ijq}}{\sum_j \delta_{ij}\zeta_{ijq}}\right)$ (Equation 4.22).

Taking expectation of both Equations 4.16 and 4.22 with respect to Y_{ijq} gives rise to the population-averaged expected value given in Equation 4.17. The definitions of negative multinomial and negative binomial distributions are given in the following subsections.

Negative Multinomial Distribution

This is the distribution on the $n+1 > 2$ non-negative integers outcomes $\{X_0, \ldots, X_n\}$, with corresponding probability of occurence $p = \{p_0, \ldots, p_n\}$ and probability mass function

$$\Gamma\left(\sum_{i=0}^{n} x_i\right)\frac{p_0^{x_0}}{\Gamma(x_0)}\prod_{i=1}^{n}\frac{p_i^{x_i}}{x_i!},$$

for parameters $x_0 > 0$ and $p = (p_i)_{i=1}^n$, where $p_i \in (0,1)$ for all i, $\sum_{i=0}^n p_i = 1$ and $\Gamma(\cdot)$ is the Gamma function. We write $Y \sim \mathcal{NM}(x_0, p)$. For positive integer x_0, the negative multinomial distribution can be recognized as the joint distribution of the n-tuple $\{X_1, \ldots, X_n\}$ when performing sampling until X_0 reaches the predetermined value x_0. The mean vector of negative multinomial distribution is given by $\frac{x_0}{p_0}p$.

Negative Binomial Distribution

This is the distribution on the non-negative integers outcome X, with corresponding probability of occurence p and probability mass function

$$\frac{\Gamma(r+x)}{x!\Gamma(r)}(1-p)^r p^x,$$

for parameters $r > 0$ and $p \in (0,1)$. We write $X \sim \mathrm{NB}(r,p)$. For positive integer r, the negative binomial distribution can be recognized as the distribution for the number of heads before the rth tail in biased coin-tossing, but it is a valid distribution for all $r > 0$. In engineering, it is sometimes called the Pólya distribution in the case where r is not integer.

Derivation of the Expectation/Conditional Maximixation (ECM) Algorithm in Section 4.3.4

Treating $\lambda = \lambda_{iq}$ for all i and $q = 2$ to Q as missing data and $y = y_{ijq}$ for all i, j and q as observed data, the complete data is (y_{ijq}, λ). Denote $\theta = (\gamma, (\beta_q)_{q=2}^Q)$, where γ includes the incidental parameters $\log(\delta_{ij})$ for all i and j. The complete data log-likeliood $\ell(\theta|y,\lambda)$ is

$$-\sum_i\sum_{q\neq 1}\lambda_{iq}\left(\sum_j e^{x_{ijq}^T\gamma}\right)+\sum_i\sum_{q\neq 1}y_{i+q}\log\lambda_{iq}+\sum_i\sum_j\sum_q x_{ijq}^T\gamma y_{ijq}+\sum_i\sum_j\sum_q\log(y_{ijq}!)+$$
$$\sum_i\sum_{q\neq 1}(1/\beta_q-1)\log\lambda_{iq}-\sum_i\sum_{q\neq 1}\lambda_{iq}/\beta_q-\sum_i\sum_{q\neq 1}\log\beta_q/\beta_q-\sum_i\sum_{q\neq 1}\log\Gamma(1/\beta_q). \tag{4.23}$$

The $(t+1)$th E-step involves finding the conditional expectation of the complete data log-likelihood with respect to to the conditional distribution of λ given y and the current estimated parameter $\theta^{(t)}$. Straightforward algebra establishes that

$$\lambda_{iq}|y_{ijq},\theta^{(t)}\sim\mathcal{G}\left(y_{i+q}+1/\beta_q^{(t)},\left(\sum_j e^{x_{ijq}\gamma^{(t)}}+1/\beta_q^{(t)}\right)^{-1}\right), \tag{4.24}$$

independently for each i and q, where the gamma distribution is parameterized in terms of scale parameter. It follows that

$$\hat{\lambda}_{iq}^{(t+1)}=\mathrm{E}\left(\lambda_{iq}\Big|(y_{ijq})_j,\theta^{(t)}\right)=\frac{y_{i+q}+1/\beta_q^{(t)}}{\sum_j e^{x_{ijq}\gamma^{(t)}}+1/\beta_q^{(t)}} \tag{4.25}$$

$$\hat{\chi}_{iq}^{(t+1)}=\mathrm{E}\left(\log(\lambda_{iq})\Big|(y_{ijq})_j,\theta^{(t)}\right)=\psi\left(y_{i+q}+1/\beta_q^{(t)}\right)-\log\left(\sum_j e^{x_{ijq}\gamma^{(t)}}+1/\beta_q^{(t)}\right). \tag{4.26}$$

Thus, in the $(t+1)$th E-step, we replace λ_{iq} and $\chi_{iq}=\log(\lambda_{iq})$ in Equation 4.23 with $\hat{\lambda}_{iq}^{(t+1)}$ and $\hat{\chi}_{iq}^{(t+1)}$, giving $Q(\theta|\theta^{(t)})$. The $(t+1)$th CM-step then finds $\theta^{(t+1)}$ to maximize $Q(\theta|\theta^{(t)})$ via a sequence of conditional maximization steps, each of which maximizes the Q function over a subset of θ, with the rest fixed at its previous value. In our application, it is natural to partition θ into γ and β_q for each $q=2$ to Q. Differentiating Equation 4.23 with respect to γ, we obtain

$$-\sum_i\sum_j\sum_q\lambda_{iq}x_{ijq}e^{x_{ijq}^T\gamma}+\sum_i\sum_j\sum_q x_{ijq}y_{ijq}, \tag{4.27}$$

which is the score equation of the Poisson log-linear model (McCullagh and Nelder, 1989) with an additional offset λ_{iq}. This allows us to leverage existing functions for fitting generalized linear models available in most statistical software packages for maximizing γ in the CM step. This is an important feature as γ often contains a huge amount of parameters in our applications, due to the inclusion the incidental parameter $\log(\delta_{ij})$ for every unique combination of covariates. Existing functions for fitting generalized linear models are typically stable and heavily optimized, even for a large number of parameters. Maximizing β_q in the CM step for each q is straightforward, as it only involves univariate optimization.

Chapter 5

Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era

Summary

The big data era demands new statistical analysis paradigms, since traditional methods often break down when datasets are too large to fit on a single desktop computer. Divide and Recombine (D&R) is becoming a popular approach for big data analysis, where results are combined over subanalyses performed in separate data subsets. In this chapter, we consider situations where unit record data cannot be made available by data custodians due to privacy concerns, and explore the concept of statistical sufficiency and summary statistics for model fitting. The resulting approach represents a type of D&R strategy, which we refer to as *summary statistics D&R*; as opposed to the standard approach, which we refer to as *horizontal D&R*. We demonstrate the concept via an extended Gamma-Poisson model, where summary statistics are extracted from different databases and incorporated directly into the fitting algorithm without having to combine unit record data. By exploiting the natural hierarchy of data, our approach has major benefits in terms of privacy protection. Incorporating the proposed modeling framework into data extraction tools such as TableBuilder by the Australian Bureau of Statistics allows for potential analysis at a finer geographical level, which we illustrate with a multilevel analysis of the Australian unemployment data.

Keywords: Big Data; Distributed Database; Divide and Recombine; Generalized Linear Mixed Model; Multilevel Model; Privacy.

5.1 Introduction

The advent of big data has created a new research paradigm, with increasing reliance on large-scale administrative data from both public and private sectors (Einav and Levin, 2014). These changes have had concomitant impact on statistical analysis. The traditional practice of performing statistical analysis using a single combined dataset is often infeasible due to memory and storage limitations of standard computers. Adding to these issues are privacy concerns, which often render data custodians reluctant to release unit record data. These issues combine to limit the ability of analysts to fully unlock the actionable information in big data.

As a solution to the memory and storage limitations problem, Divide and Recombine (D&R) has been proposed as an effective, generic approach to statistical analysis of big data (Guha et al., 2012). D&R involves (i) dividing data into manageable subsets, (ii) performing statistical analysis independently on each subsets, and then (iii) combining the results, typically via some form of averaging (Figure 5.1). Data are typically divided via either *replicate* division or *conditioning variable* division (Bühlmann et al., 2016, chap. 3). Replicate division divides the data based on random sampling without replacement, whereas conditioning variable division stratifies the data according to one or more variables in the data. An example of conditioning variable division is to partition disease incidence data by postal areas. The recombination method is chosen in a way that results in the least discrepancy compared to the *all data estimate*, that is, estimate obtained by using the entire dataset. Except in very simple cases, D&R results are approximate.

DeltaRho (formerly Tessera) (Bühlmann et al., 2016, chap. 3) is an open source implementation of D&R that combines the R statistical programming environment (R

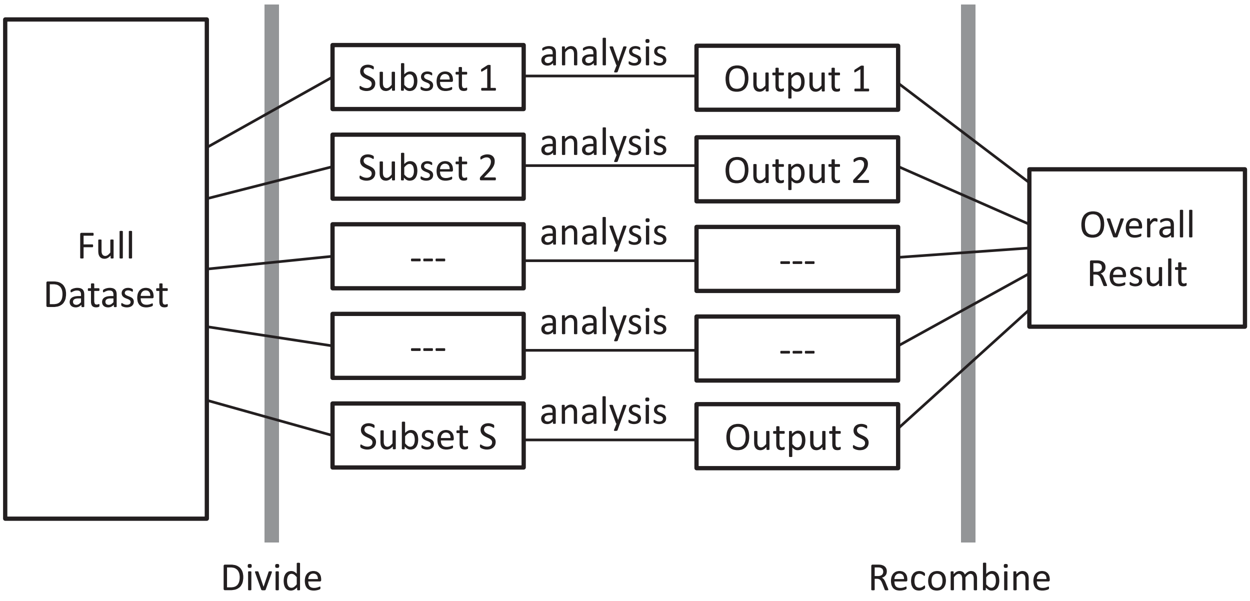


Figure 5.1 : The Divide and Recombine (D&R) framework.

Development Core Team, 2017) at the front end with various back end options such as Hadoop (White, 2009) and Spark (Zaharia et al., 2010, 2012). This allows users to scalably leverage all of the statistical methods readily available in R while abstracting the technical programming details, making D&R more accessible to the general statistical community. The emergence of these systems has sparked research interest in the D&R algorithm. A selection of recent examples include Boyd et al. (2011); Chu et al. (2013); Lubell-Doughtie and Sondag (2013); Scott et al. (2016); Chen and Xie (2014); Kleiner et al. (2014); Minsker et al. (2014); Neiswanger et al. (2014); Xu et al. (2014); Perry (2017); and Miroshnikov et al. (2015). In typical D&R applications, we have unit record data where analysis on a single machine is not feasible, because the data are either too large to store, or of moderate size but the statistical method being used is very computationally intensive. The dataset is divided into subsets of similar structure and the intended analysis is performed on each of the subsets. We refer to this kind of division as *horizontal D&R*, where unit level data are partitioned in such a way that each subset holds the same variables but for different cases.

In this chapter, we consider situations where unit record data cannot be made available due to privacy reasons, even after personal identifiers such as name, address, date of birth, and ID number have been removed. This situation arises often in practice because the presence of rich information, when combined with the use of sophisticated data mining tools, renders privacy breaching a major threat (Fienberg, 2006). This is true even after statistical disclosure control methods (Hundepool et al., 2012) have been applied to safeguard the confidentiality of data (Sweeney, 2002; Coull et al., 2007; Homer et al., 2008; Narayanan and Shmatikov, 2008).

As a solution, we explore the concept of statistical sufficiency and summary statistics for model fitting. Sufficiency is a concept taught in every introductory mathematical statistics course, but it has not been actively utilized for practical model fitting because the need has not been there. The aim is to compress the raw data in each subset into low dimensional summary statistics for model fitting. We refer to this as *summary statistics D&R*, emphasizing the fact that unit record data cannot be made available, as opposed to horizontal D&R. We illustrate the concept via a multilevel model (Gelman and Hill, 2007; Goldstein, 2011) based on an extension of the Gamma-Poisson model by Christiansen and Morris (1997). In this context, the use of summary statistics exploits the natural grouping structure in the data and allows the direct modeling of data from multiple sources using summary information, without the need to combine them into a single file, thus is privacy preserving. We apply the model to publicly available unemployment data from the Australian Bureau of Statistics and explain the benefit of our framework in terms of allowing analysis at a finer geographical level.

The chapter is organized as follows. In Section 5.2, we motivate the distinction between summary statistics D&R and horizontal D&R using simple linear regression as an example. We then describe the proposed extended Gamma-Poisson model in Section 5.3. In Section 5.4, the model is applied to the Australian unemployment data. We close with some concluding remarks in Section 5.5.

5.2 Illustrative Example

A multiple linear regression takes the general form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right),$$

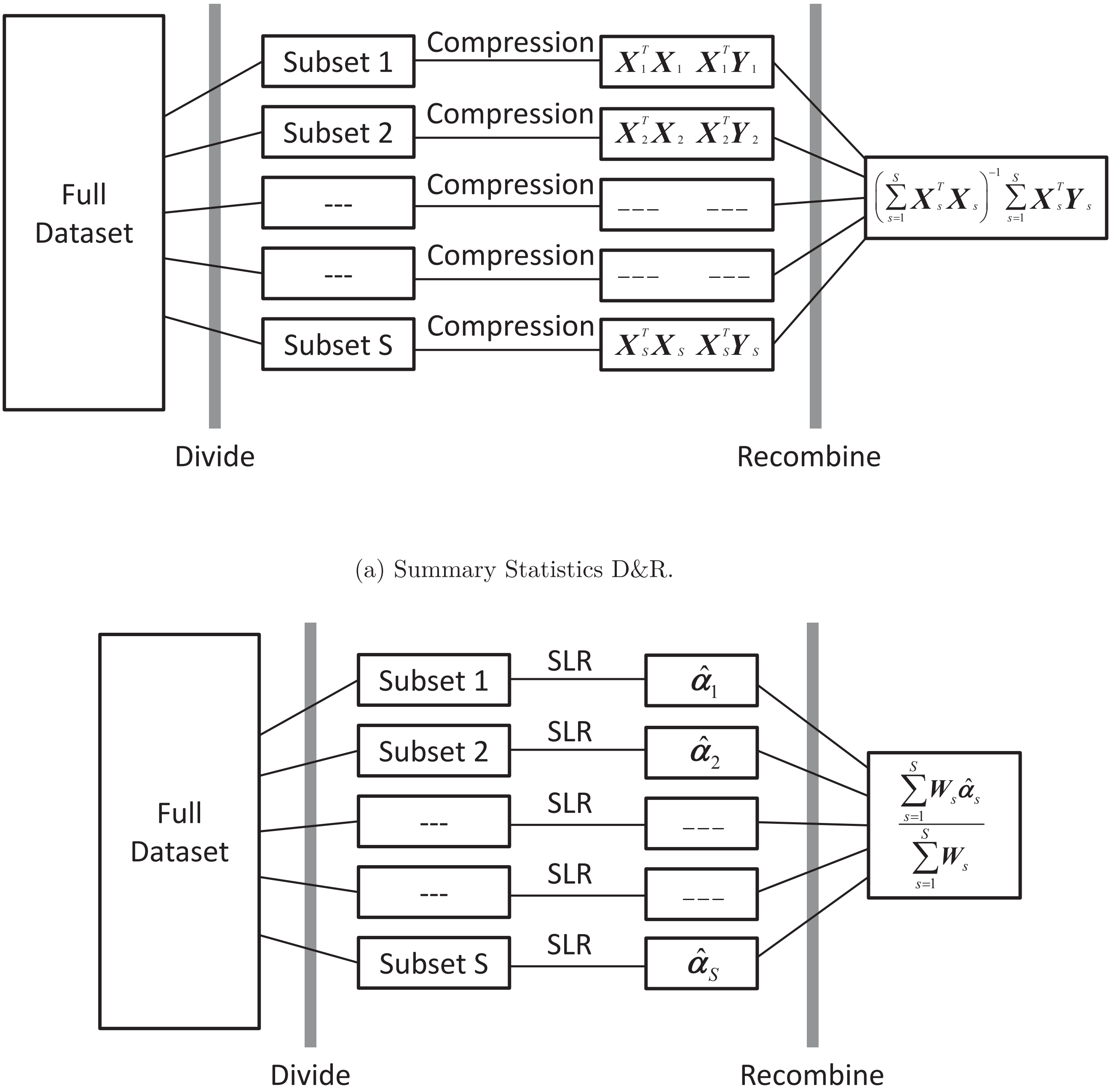
where $\boldsymbol{Y}$ is a $n \times 1$ vector of response variables, $\boldsymbol{X}$ is a $n \times p$ design matrix, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of regression parameter, $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of independent errors, and $\boldsymbol{\Sigma}$ is a $n \times n$ diagonal matrix, with common diagonal elements σ^2. Standard least squares and maximum likelihood estimates give the *all data estimate* $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. When data are too large to fit into a single machine, we can resort to two different approaches: (i) summary statistics D&R and (ii) horizontal D&R.

Summary statistics D&R (Figure 5.2a) includes the following steps: (i) divide the data into S subsets of similar structure, with $\boldsymbol{Y}_s$ denoting the vector of responses in subset s and $\boldsymbol{X}_s$ the corresponding design matrix, (ii) calculate two sets of summary data for each of the subsets, i.e. $\boldsymbol{X}_s^T\boldsymbol{X}_s$ and $\boldsymbol{X}_s^T\boldsymbol{Y}_s$, then (iii) combine via $\left(\sum_s \boldsymbol{X}_s^T\boldsymbol{X}_s\right)^{-1}\sum_s \boldsymbol{X}_s^T\boldsymbol{Y}_s$. Chen et al. (2006) refer to this as the *regression cube* technique. The resulting *aggregated estimate* is exactly equivalent to the all data estimate due to the matrix properties $\boldsymbol{X}^T\boldsymbol{X} = \sum_s \boldsymbol{X}_s^T\boldsymbol{X}_s$ and $\boldsymbol{X}^T\boldsymbol{y} = \sum_s \boldsymbol{X}_s^T\boldsymbol{y}_s$.

Horizontal D&R (Figure 5.2b) includes the following steps: (i) divide the data into S subsets of similar structure, (ii) perform independent least squares regression on each subset to obtain $\hat{\boldsymbol{\alpha}}_s = (\boldsymbol{X}_s^T \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \boldsymbol{Y}_s$ and then (iii) weight the results to obtain the aggregated estimate $\sum_s \boldsymbol{W}_s \hat{\boldsymbol{\alpha}}_s / \sum_s \boldsymbol{W}_s$. The optimal weight is obtained using $\boldsymbol{W}_s = \boldsymbol{X}_s^T \boldsymbol{X}_s$, which is proportional to the inverse variance-covariance matrix of the regression parameter. This results in an aggregated estimate that is exactly equivalent to the one obtained via the summary statistics approach. Intuitively, using the inverse variance-covariance matrix as weight makes sense as we are giving larger credibility to subsets with lower variability.

Summary statistics D&R differs from horizontal D&R. For summary statistics D&R, we extract the relevant summary statistics that best summarizes data in each subset, so that the resulting aggregated estimate is as “close” as possible to the all data estimate. Only summary data are extracted and made available to the analyst, rendering unit record data unnecessary. For horizontal D&R, we perform the intended statistical analysis (linear regression in this case) independently on each subset and choose an aggregate estimate to minimize the error. Unit record data are typically required so that the intended analysis can be done on each subset.

For the simple case of linear regression, horizontal D&R is not materially any different from summary statistics D&R assuming optimal weights have been used, as both aggregated estimates are exactly equivalent to the all data estimate due to the linearity of the estimating equation in $\boldsymbol{\alpha}$. However, this is generally not true for more complicated models such as logistic regression (Xi et al., 2009) and nonlinear estimating equations (Lin and Xi, 2011), where we can only hope to find aggregated estimators that are consistent. We shall see in the next section that summary statistics D&R allows the exact reproduction of regression estimates for the extended Gamma-Poisson model, as opposed to horizontal D&R.


Full Dataset
Divide
Subset 1
Subset 2

Subset S
Compression
$X_1^T X_1 \;\; X_1^T Y_1$
$X_2^T X_2 \;\; X_2^T Y_2$
$X_S^T X_S \;\; X_S^T Y_S$
Recombine
$\left(\sum_{s=1}^{S} X_s^T X_s\right)^{-1} \sum_{s=1}^{S} X_s^T Y_s$
SLR
$\hat{\alpha}_1$
$\hat{\alpha}_2$
$\hat{\alpha}_S$
$\frac{\sum_{s=1}^{S} W_s \hat{\alpha}_s}{\sum_{s=1}^{S} W_s}$

(a) Summary Statistics D&R.

(b) Horizontal D&R.

Figure 5.2 : Linear Regression via Divide and Recombine (D&R).

5.3 Extended Gamma-Poisson Model

In this section, we propose a model for correlated data that characterizes individual level event rates as a function of both individual and area level covariates, and show that it can be fitted using sufficient and summary statistics. Our approach extends the Gamma-Poisson model of Christiansen and Morris (1997) to include both individual and area level predictors. The Poisson component characterizes the effect of individual level variables on the event rate. The Gamma component incorporates the effects of area level covariates as well as a random component that reflects area to area variation that is not captured by area level covariates.

5.3.1 Model Formulation

Let Y_{ij} denotes whether or not the ith individual in the jth area experienced the event of interest and $\boldsymbol{x}_{ij}$ be the $p \times 1$ vector of covariates measured on this individual, with $x_{ij,1} = 1$ to allow for the intercept. Given an area specific random effect b_j, we assume that Y_{ij} is independently Poisson distributed with mean $\lambda_{ij} = b_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown regression coefficients to be estimated. The random effect b_j is assumed to follow a Gamma distribution with parameters such that the mean μ_j is $\exp(\boldsymbol{u}_j^T \boldsymbol{\gamma})$ and the variance is $\kappa\mu_j$, where $\boldsymbol{u}_j$ is a vector of area level covariates, $\boldsymbol{\gamma}$ is the corresponding vector of unknown regression parameters to be estimated and κ is a dispersion parameter to be estimated. That is,

$$\begin{aligned} b_j &\overset{ind.}{\sim} \text{Gamma}\left(\frac{\mu_j}{\kappa}, \kappa\right) \\ &\overset{ind.}{\sim} \text{Gamma}\left[\mu_j, \kappa\mu_j\right]. \end{aligned}$$

The random effect b_j captures the deviation of the area specific rates from the mean outcome, taking into account area level variables as well as any remaining unexplained variation. In our formulation, $b_j > 1$ corresponds to a positive deviation of area j from the mean rate, whereas $0 < b_j < 1$ corresponds to a negative deviation. Note that we parameterize the Gamma distribution in terms of an area specific mean parameter $\mu_j > 0$ and a constant scale parameter $\kappa > 0$. The round parentheses indicate the standard shape and scale parameterization of the Gamma distribution, whereas the square brackets indicate the mean and variance formulation.

We choose to model the binary response variable with a Poisson model, as the Poisson distribution is a good approximation to the Binomial distribution when dealing with relatively rare events (events where the chance of a success on any particular trial is small) such as unemployment and heart disease. The Poisson model is often preferred because the covariate effects can be directly interpreted as risk ratios due to the log canonical link. Although we focus on the case where Y_{ij} is a binary 0 or 1 variable, extension to the more general case where Y_{ij} can be any integer counts is straightforward.

5.3.2 Model Fitting using Summary Data

With some straightforward algebra, the log-likelihood function $\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y})$ can be written as

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y}) = \sum_j \left(\sum_{i:Y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right) +$$

$$\sum_j \left\{ -\frac{\mu_j}{\kappa} \log(\kappa) - \log\Gamma\left(\frac{\mu_j}{\kappa}\right) + \log\Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) - \left(\omega_j + \frac{\mu_j}{\kappa}\right) \log\left[\sum_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right] \right\}.$$

The log-likelihood is a function of a few sufficient and summary statistics from various databases. Figure 5.3 presents a summary diagram of the data requirement to obtain the maximum likelihood estimates via the Newton-Raphson algorithm. For individuals who experienced the event, the summary data required are:

1. For categorical variables, the number of people who had the event within every level of individual level variables of interest; for continuous variables, the sum of individual level variables of interest across all individuals who experienced the event. For the gender (categorical) and age (continuous) variables, this translates into the number of events among either males or females and the sum of ages for all individuals who experienced the event. Technically we write this as $\boldsymbol{X}_{\text{event}}^{T}\mathbf{1}$, where $\boldsymbol{X}_{\text{event}}$ is a $M \times p$ matrix where each row comprises a row vector of covariates for one of the M individuals who experienced the event of interest. Hence $\boldsymbol{X}_{\text{event}}^{T}\mathbf{1}$ is simply the column totals of $\boldsymbol{X}_{\text{event}}$.

2. The number of subjects who had the event in each area, defined as ω_j for area j.

The summary data above only need to be computed once, since they do not depend on any unknown model parameters being estimated. Technically these are sufficient statistics (Casella and Berger, 2002). For the dataset on the population at risk, three sets of summary statistics are required from each area j. These involve summations over all individuals living in the common area, rendering individual level data unnecessary. These are not sufficient statistics since they involve the unknown parameter $\boldsymbol{\alpha}$. A similar process applies for the area dataset. At each iteration of the algorithm, likelihood contributions are computed as functions of these summary statistics, leading to an improved value of the unknown parameters. The process repeats until convergence.

As an illustration, the framework can be applied to hospital variation studies whose goal is to quantify variation in hospital admission rates as a function of a variety of individual level factors such as age, gender, medical history, as well as hospital level factors such as proportion of interns, ratio of residents to beds, hospital

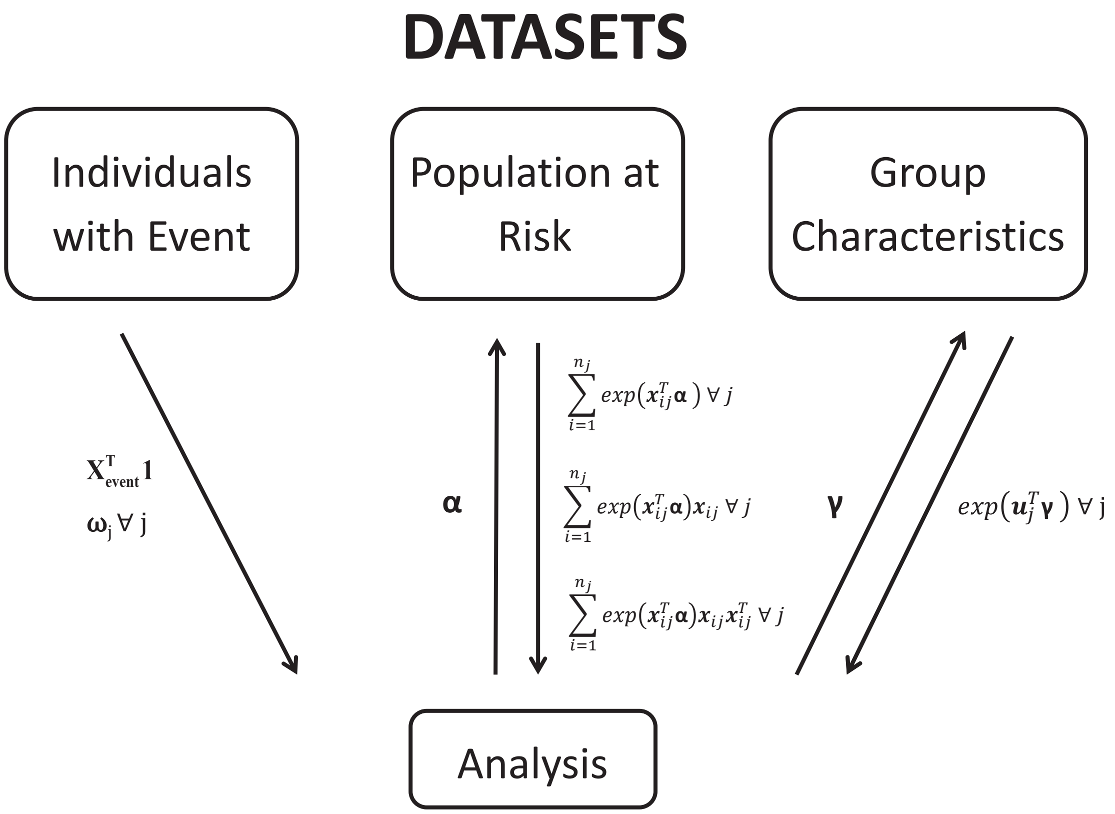

DATASETS
Individuals with Event
Population at Risk
Group Characteristics
$\mathbf{X}_{\mathbf{event}}^{\mathbf{T}}\mathbf{1}$
$\boldsymbol{\omega}_j \, \forall \, j$
$\boldsymbol{\alpha}$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}) \, \forall \, j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij} \, \forall \, j$
$\sum_{i=1}^{n_j} exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T \, \forall \, j$
$\boldsymbol{\gamma}$
$exp(\boldsymbol{u}_j^T\boldsymbol{\gamma}) \, \forall \, j$
Analysis

Figure 5.3 : Data requirement to obtain the maximum likelihood estimates of the extended Gamma-Poisson model via the Newton-Raphson algorithm. Only a few sufficient and summary statistics are required without needing to access the unit record data.

resources and area level socio-economic advantage. Patients' data are confidential and hospitals are obliged to protect them. De-identifying individual level data is not sufficient to prevent disclosure, and analysts who wish to obtain the data have to go through an ethics application, which can be time consuming if not impossible. However, hospitals may be quite willing to provide the sufficient and summary statistics required by the proposed framework. These summaries are then passed to an analysis computer via a network, which returns an improved value of the unknown parameters and passed back to the hospitals to compute a new set of summary statistics. The process iterates until convergence.

5.3.3 Parameter Initialization

A good starting value is essential to ensure proper convergence. Here, we propose an initialization process for the regression coefficients using summary data. Assuming an independence structure, the parameters of standard log-linear Poisson models can be estimated via the Newton-Raphson algorithm

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $\boldsymbol{A} = \mathrm{diag}\left(\mathrm{Var}(Y_{ij}) \frac{\mathrm{d}\eta_{ij}}{\mathrm{d}\mu_{ij}}\right)$ and $\boldsymbol{W} = \boldsymbol{A}\left(\frac{\mathrm{d}\eta_{ij}}{\mathrm{d}\mu_{ij}}\right)^{-1}$, with η being the link function.

For log-linear models where the outcome is either a success or a failure, $\boldsymbol{A} = \mathrm{diag}(1)$ and $Y = 0, 1$. Thus, the equation above can be rewritten as:

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + (\boldsymbol{X}_{\text{pop}}^T \boldsymbol{W} \boldsymbol{X}_{\text{pop}})^{-1} (\boldsymbol{X}_{\text{event}}^T \mathbf{1} - \boldsymbol{X}_{\text{pop}}^T \boldsymbol{\mu}),$$

where $\boldsymbol{X}_{\text{event}}$ and $\boldsymbol{X}_{\text{pop}}$ are the design matrices for individuals with event and the entire population at risk respectively, $\boldsymbol{\mu} = \exp(\boldsymbol{X}_{\text{pop}} \boldsymbol{\alpha})$ and $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{\mu})$. Thus, we only require a set of sufficient statistics for all individuals with event $\left(\boldsymbol{X}_{\text{event}}^T \mathbf{1}\right)$ and two sets of summary statistics for the population at risk for each area $\left(\boldsymbol{X}_j^T \boldsymbol{W}_j \boldsymbol{X}_j \ \forall j \text{ and } \boldsymbol{X}_j^T \boldsymbol{\mu}_j \ \forall j\right)$, due to the matrix properties $\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} = \sum_j \boldsymbol{X}_j^T \boldsymbol{W}_j \boldsymbol{X}_j$ and $\boldsymbol{X}^T \boldsymbol{\mu} = \sum_j \boldsymbol{X}_j^T \boldsymbol{\mu}_j$.

These sufficient and summary statistics coincide with those in Figure 5.3. In other words, the same set of sufficient and summary statistics can be used for parameter initialization and model fitting.

5.4 Application: Australian Unemployment Data

The dependence of the extended Gamma-Poisson model on only a few sufficient and summary statistics allows the fitting of detailed statistical models without actually having access to unit record data. This has important applications in terms of privacy protection in distributed databases, which we illustrated using the hypothetical hospital variation studies example in Section 5.3.2. This section aims to demonstrate the benefit of incorporating the proposed modeling framework into data extraction tools such as TableBuilder by the Australian Bureau of Statistics (ABS).

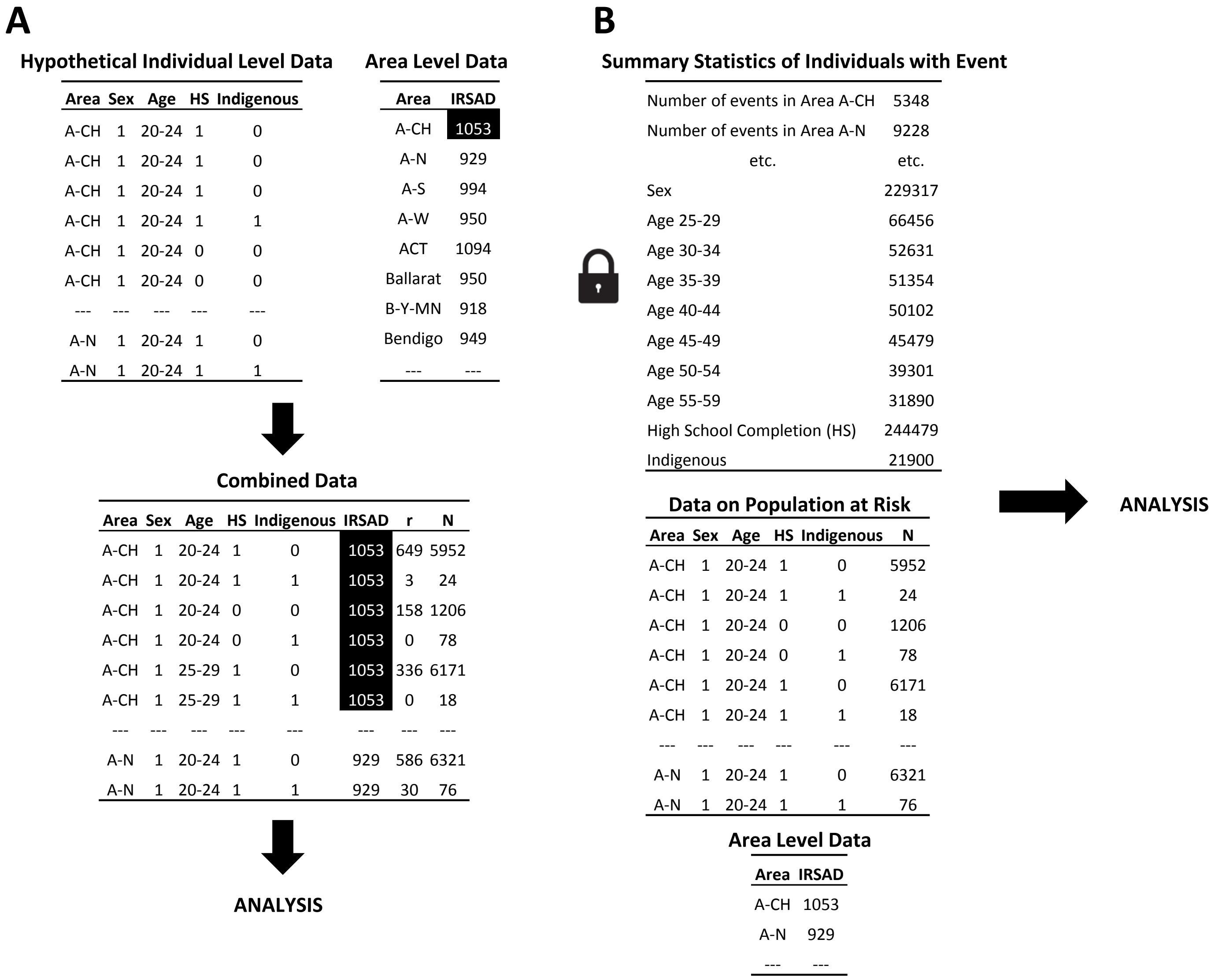
We obtain the Australian unemployment data from TableBuilder, an online tool by the ABS whereby users can build cross-classifications of census variables for geographical areas as defined in the Australian Statistical Geography Standard (ASGS) (Australian Bureau of Statistics, 2012). We wish to explore whether individuals living in areas in 2011 that had higher socio-economic advantage prior to the Global Financial Crisis (2007-8) have more resilience to unemployment, adjusting for individual level variables such as sex, age, high school completion and identifying as an indigenous Australian. The individual level variables sex, age, high school completion and indigenous status in 2011 are cross-classified according to Statistical Area Level 4 (SA4). SA4 is chosen in this context as it was originally designed for the outputs of the Australian Labour Force Survey. For a measure of socio-economic advantage, we use the Index of Relative Advantage and Disadvantage (IRSAD) (Australian Bureau of Statistics, 2008) that can be obtained from the ABS in a separate database. High values of IRSAD indicate high social advantage, and vice versa. We use IRSAD values from the 2006 Census, the most recent Census preceeding 2011.

We now have data on individuals with events and population at risk from Table-

Builder, and data on area level covariates from a separate ABS database. Available packages for fitting multilevel models in standard statistical softwares such as lme4 (Bates et al., 2015) in R (R Development Core Team, 2017) require data to be combined (Figure 5.4a). This results in unnecessary repetition of area level covariates over multiple rows (see the shaded cells in Figure 5.4a). With the extended Gamma-Poisson model, we can avoid the step of combining these data into a single large design matrix by computing just a few summary statistics directly from each dataset (Figure 5.4b).

Moreover, privacy legislation precludes anyone apart from ABS employees from having access to individual level census data. Only confidentialized tabular data can be made available to researchers, whereby the counts are randomly adjusted to reduce the risk of disclosure (O'Keefe, 2008; Leaver, 2009). This means that we need to be cautious when using data from tables with small cell counts, since they are likely to be unreliable. This is especially true for the event counts (denoted by r in Figure 5.4a), even more when the event under consideration is rare or when there are many cross-classifications. In our application, it would be ideal to model at the finer SA3 level as it provides a more detailed analysis compared to the Australian Labor Force Survey. However, small cell counts prevents us from doing so. In this regard, extending tools such as TableBuilder to only output the sufficient statistics of individuals with event required by the Gamma-Poisson model would be useful (Figure 5.4b).

We fit the extended Gamma-Poisson model to the unemployment data, and compare the results with the Normal-Poisson model fitted on the combined data using the lme4 package (Bates et al., 2015) in R (R Development Core Team, 2017). The estimates and standard errors produced by both models are very similar. The results,



Hypothetical Individual Level Data

Area	Sex	Age	HS	Indigenous
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	0
A-CH	1	20-24	1	1
A-CH	1	20-24	0	0
A-CH	1	20-24	0	0
---	---	---	---	---
A-N	1	20-24	1	0
A-N	1	20-24	1	1

Area Level Data

Area	IRSAD
A-CH	1053
A-N	929
A-S	994
A-W	950
ACT	1094
Ballarat	950
B-Y-MN	918
Bendigo	949
---	---

Combined Data

Area	Sex	Age	HS	Indigenous	IRSAD	r	N
A-CH	1	20-24	1	0	1053	649	5952
A-CH	1	20-24	1	1	1053	3	24
A-CH	1	20-24	0	0	1053	158	1206
A-CH	1	20-24	0	1	1053	0	78
A-CH	1	25-29	1	0	1053	336	6171
A-CH	1	25-29	1	1	1053	0	18
---	---	---	---	---	---	---	---
A-N	1	20-24	1	0	929	586	6321
A-N	1	20-24	1	1	929	30	76

Summary Statistics of Individuals with Event

Number of events in Area A-CH	5348
Number of events in Area A-N	9228
etc.	etc.
Sex	229317
Age 25-29	66456
Age 30-34	52631
Age 35-39	51354
Age 40-44	50102
Age 45-49	45479
Age 50-54	39301
Age 55-59	31890
High School Completion (HS)	244479
Indigenous	21900

Data on Population at Risk

Area	Sex	Age	HS	Indigenous	N
A-CH	1	20-24	1	0	5952
A-CH	1	20-24	1	1	24
A-CH	1	20-24	0	0	1206
A-CH	1	20-24	0	1	78
A-CH	1	20-24	1	0	6171
A-CH	1	20-24	1	1	18
---	---	---	---	---	---
A-N	1	20-24	1	0	6321
A-N	1	20-24	1	1	76

Area Level Data

Area	IRSAD
A-CH	1053
A-N	929
---	---

Figure 5.4 : Data requirement. (A) Existing software packages for fitting multilevel models require the individual and area level data to be combined into a single file before performing analysis. This results in the unnecessary repetition of area level variables as indicated by the shaded cells. (B) By using the extended Gamma-Poisson model, datasets are analyzed directly without the need to combine them. In addition, for individuals who experienced the event, the model only requires sufficient statistics instead of the full dataset.

summarized in Table 5.1, reveal some interesting patterns. Unemployment rates are lower for males ($p < 0.001$) and tend to decrease with age. The unemployment rate of a person who completed high school is 36.87% lower than for a person who did not complete high school ($p < 0.001$). The unemployment rate of indigenous Australians is 115.98% higher than non-indigenous Australians, even after controlling for education ($p < 0.001$), demonstrating the continued need to reduce economic disadvantage for this community. After adjusting for individual level characteristics, area level measures of social advantage have only a modest impact ($p = 0.138$) on unemployment rates. However, there remains significant area-to-area variation in unemployment rates.

5.5 Concluding Remarks

The chapter argues that statistical sufficiency and summary statistics offer an attractive framework in the big data era, especially in the setting of large-scale administrative databases where privacy concerns prevent general access to unit record data. The concept is illustrated via an extended Gamma-Poisson multilevel model. The model works by gathering relevant pieces of summary information required for construction of the log-likelihood directly from the separate data sources. This is a natural solution since the relevant data are often drawn from different sources anyway. For example, epidemiologists often augment their study populations with information about the communities in which their study participants live. Such community-level variables might be obtained from a national census or from other surveys. As another example, information about the study population may come from different hospital administrative databases, held locally at the respective hospitals. Sharing these databases among hospitals might not be possible due to privacy reasons. Aside from offering benefits in terms of privacy protection, incorporating the model into data extraction tools such as TableBuilder allows for potential analysis at a finer geographical level.

	Normal-Poisson		Gamma-Poisson	
Parameter	Est	SE	Est	SE
α_o (Intercept)	-2.06*	0.028	-2.03*	0.027
α_1 (Female)		*Reference Group*		
α_1 (Male)	-0.06*	0.003	-0.06*	0.003
α_2 (age 20 to 24)		*Reference Group*		
α_2 (age 25 to 29)	-0.51*	0.005	-0.51*	0.005
α_3 (age 30 to 34)	-0.70*	0.005	-0.70*	0.005
α_4 (age 35 to 39)	-0.80*	0.005	-0.80*	0.005
α_5 (age 40 to 44)	-0.92*	0.006	-0.92*	0.006
α_6 (age 45 to 49)	-1.03*	0.006	-1.03*	0.006
α_7 (age 50 to 54)	-1.11*	0.006	-1.11*	0.006
α_8 (age 55 to 59)	-1.10*	0.007	-1.10*	0.007
α_9 (Not completed High School)		*Reference Group*		
α_9 (Completed High School)	-0.46*	0.003	-0.46*	0.003
α_{10} (Not Indigenous)		*Reference Group*		
α_{10} (Indigenous)	0.77*	0.007	0.77*	0.007
γ (IRSAD)	-0.04	0.027	-0.04	0.027
σ^2	0.07	N/A	N/A	N/A
κ	N/A	N/A	0.06*	0.010

*Significant at $p < 0.001$.

Table 5.1 : Estimates and standard errors based on the Australian unemployment data, fitted using the Normal-Poisson model and the extended Gamma-Poisson model. “Est” and “SE” correspond to the estimates and standard errors, respectively.

The ideas discussed in this chapter bear some connection to *symbolic data analysis* (Billard and Diday, 2006), where data are compressed into distributions such as hyperrectangles or histograms, rather than a single summary. The main difference is that in symbolic data analysis, exact statistical analysis is performed using approximate data (e.g., loss of information when summarizing individual data points into histograms); whereas in our Gamma-Poisson model, exact analysis is performed using exact data (using summary statistics does not result in any loss of information under the Gamma-Poisson model, compared to using unit record data).

The proposed model has the potential to become a valuable addition to the statistician's toolbox in the quest to make better use of the ever increasing volumes of data being generated in the big data era (Einav and Levin, 2014). More generally, the model and analysis we developed and implemented in this chapter are examples of rethinking classic statistical ideas for model fitting in the big data era. There is great potential for developing new algorithms that can be used in the analysis of large administrative databases. For example, in the case of *vertical D&R*, where data are partitioned in such a way that each partition hold a subset of the variables for the common individuals, there is still considerable methodological work to be done. It would be good to see more of these developments happening in the statistics literature.

Supplementary Materials

Derivation of the fitting algorithm for the extended Gamma-Poisson model

In this section, we derive the fitting algorithm for the extended Gamma-Poisson model proposed in Section 5.3. The likelihood function $\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y})$ is obtained by integrating out the random effects. Under our proposed model, the likelihood can be solved explicitly. Straightforward algebra establishes that the likelihood equation is a function of a few sufficient and summary statistics.

$$
\begin{aligned}
&\prod_j \left\{ \int \left[\left(\prod_{i=1}^{n_j} f(y_{ij}|b_j) \right) f(b_j) \right] \mathrm{d}b_j \right\} \\
&\quad = \prod_j \left\{ \int \left[\left(\prod_{i=1}^{n_j} \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!} \right) f(b_j) \right] \mathrm{d}b_j \right\} \\
&\quad = \prod_j \left\{ \int \left[\left(\prod_{i=1}^{n_j} e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}} \right) f(b_j) \right] \mathrm{d}b_j \right\} \\
&\quad = \prod_j \left\{ \int \left[\left(\prod_{i=1}^{n_j} e^{-b_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})} b_j^{y_{ij}} e^{(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}) y_{ij}} \right) f(b_j) \right] \mathrm{d}b_j \right\} \\
&\quad = \prod_j \left\{ \prod_{i=1}^{n_j} e^{(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}) y_{ij}} \int \left[\left(\prod_{i=1}^{n_j} e^{-b_j \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})} b_j^{y_{ij}} \right) f(b_j) \right] \mathrm{d}b_j \right\} \\
&\quad = \prod_j \left\{ \prod_{i:Y_{ij}=1} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} \int \left[\left(e^{-b_j \sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})} b_j^{\omega_j} \right) f(b_j) \right] \mathrm{d}b_j \right\},
\end{aligned}
$$

where the second equality follows from the fact that Y_{ij} is either 0 or 1. $i : Y_{ij} = 1$ and ω_j represents individuals who had the event of interest in area j and the total number of subjects who had events ($Y_{ij} = 1$) in area j. Since $f(b_j)$ is Gamma distributed, the integral term in the likelihood function can be expressed as

$$\begin{aligned}
&\int \left[e^{-b_j \sum\limits_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha})} b_j^{\omega_j} \frac{1}{\kappa^{\frac{\mu_j}{\kappa}} \Gamma(\frac{\mu_j}{\kappa})} b_j^{\frac{\mu_j}{\kappa}-1} e^{-\frac{b_j}{\kappa}} \right] \mathrm{d}b_j \\
&= \frac{1}{\kappa^{\frac{\mu_j}{\kappa}} \Gamma(\frac{\mu_j}{\kappa})} \int \left[b_j^{(\omega_j + \frac{\mu_j}{\kappa})-1} e^{-\left[\sum\limits_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}) + \frac{1}{\kappa}\right] b_j} \right] \mathrm{d}b_j \\
&= \frac{1}{\kappa^{\frac{\mu_j}{\kappa}} \Gamma(\frac{\mu_j}{\kappa})} \Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) \left[\sum_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right]^{-\left(\omega_j + \frac{\mu_j}{\kappa}\right)},
\end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function given by $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$. The integral in the second line equals to one when scaled appropriately, as the integrand represents the kernel of a Gamma distribution with shape parameter $(\omega_j + \frac{\mu_j}{\kappa})$ and scale parameter $\left[\sum\limits_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right]^{-1}$.

It follows that the likelihood $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y})$ is

$$\prod_j \left\{ \left(\prod_{i:Y_{ij}=1} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} \right) \frac{1}{\kappa^{\frac{\mu_j}{\kappa}} \Gamma(\frac{\mu_j}{\kappa})} \Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) \left[\sum_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right]^{-\left(\omega_j + \frac{\mu_j}{\kappa}\right)} \right\}$$

and the log-likelihood $\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y})$ is

$$\begin{aligned}
\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y}) &= \sum_j \left(\sum_{i:Y_{ij}=1} \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} \right) \\
+ \sum_j \Bigg\{ -\frac{\mu_j}{\kappa} \log(\kappa) - \log\Gamma\left(\frac{\mu_j}{\kappa}\right) &+ \log\Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) - \left(\omega_j + \frac{\mu_j}{\kappa}\right) \log\left[\sum_{i=1}^{n_j} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} + \frac{1}{\kappa}\right] \Bigg\}.
\end{aligned}$$

Maximum likelihood estimates (MLE) of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and κ can be obtained by differentiating the log-likelihood with respect to each of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and κ and set the derivatives simultaneously to zero. First we define some symbols that will ease our calculation.

We already defined $\mu_j(\boldsymbol{\gamma}) = \exp(\boldsymbol{u}_j^T\boldsymbol{\gamma})$. Now denote

$$c_j(\boldsymbol{\gamma}, \kappa) = \omega_j + \frac{\mu_j}{\kappa}$$

$$b_j(\boldsymbol{\alpha}, \kappa) = \sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}) + \frac{1}{\kappa}$$

$$\boldsymbol{h}_j(\boldsymbol{\alpha}) = \sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij} \qquad \boldsymbol{h}_j^T(\boldsymbol{\alpha}) = \sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}^T$$

$$\boldsymbol{G}_j(\boldsymbol{\alpha}) = \sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T.$$

Using the definitions above, the log-likelihood $\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y})$ can be re-written as

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa; \boldsymbol{y}) \propto \boldsymbol{1}^T\boldsymbol{X}_{\text{event}}\boldsymbol{\alpha} + \sum_j \left\{ -\frac{\mu_j}{\kappa}\log(\kappa) - \log\Gamma\left(\frac{\mu_j}{\kappa}\right) + \log\Gamma\left(c_j\right) - c_j\log(b_j) \right\},$$

where the arguments of the newly defined functions are left out on purpose for simplicity. The score vector $\boldsymbol{S}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ (vector of the first partial derivatives of the log-likelihood function) is

$$\begin{pmatrix} \boldsymbol{S}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \\ \boldsymbol{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \\ \boldsymbol{S}_{\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \end{pmatrix},$$

and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ (matrix of the second partial derivatives of the log-likelihood function) is

$$\begin{pmatrix} \boldsymbol{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\boldsymbol{\alpha}\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\boldsymbol{\alpha}\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \\ \boldsymbol{H}_{\boldsymbol{\gamma}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\boldsymbol{\gamma}\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \\ \boldsymbol{H}_{\kappa\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\kappa\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) & \boldsymbol{H}_{\kappa\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa) \end{pmatrix}.$$

The components of the score vector and the Hessian matrix are derived in the following.

Score and Hessian Contribution from $\boldsymbol{\alpha}$

Relevant terms of the log-likelihood are

$$\mathbf{1}^T \boldsymbol{X}_{\text{event}} \boldsymbol{\alpha} - \sum_{j=1}^{q} c_j \log(b_j).$$

It follows that the score equation, $\boldsymbol{S}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\begin{aligned} &\boldsymbol{X}_{event}^T \mathbf{1} - \sum_{j=1}^{q} \left\{ c_j \, \frac{\sum\limits_{i=1}^{n_j} (e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} \boldsymbol{x}_{ij})}{b_j} \right\} \\ &= \boldsymbol{X}_{event}^T \mathbf{1} - \sum_{j=1}^{q} \frac{c_j \boldsymbol{h}_j}{b_j}. \end{aligned}$$

The $\boldsymbol{\alpha}$ component of Hessian matrix, $\boldsymbol{H}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\begin{aligned} &- \sum_{j=1}^{q} \left\{ c_j \times \frac{b_j \sum\limits_{i=1}^{n_j} (e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T) - \boldsymbol{h}_j \left[\sum\limits_{i=1}^{n_j} (e^{\boldsymbol{x}_{ij}^T \boldsymbol{\alpha}} \boldsymbol{x}_{ij}^T) \right]}{b_j^2} \right\} \\ &= \sum_{j=1}^{q} \left\{ c_j \times \frac{\boldsymbol{h}_j \boldsymbol{h}_j^T - b_j \boldsymbol{G}_j}{b_j^2} \right\}. \end{aligned}$$

Score and Hessian Contribution from $\boldsymbol{\gamma}$

Relevant terms of the log-likelihood are

$$\sum_{j=1}^{q} \left\{ -\frac{\mu_j}{\kappa} \log(\kappa) - \log \Gamma\left(\frac{\mu_j}{\kappa}\right) + \log \Gamma(c_j) - \frac{\mu_j}{\kappa} \log(b_j) \right\}.$$

Denote the Digamma and Trigamma functions by $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $\psi'(x) = \frac{\mathrm{d}}{\mathrm{d}x} \psi(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \Gamma(x)$ respectively. It follows that the score equation,

$\boldsymbol{S}_\gamma(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\sum_{j=1}^{q}\left\{-\frac{\mu_j}{\kappa}\log(\kappa)\boldsymbol{u}_j - \frac{\mu_j}{\kappa}\boldsymbol{u}_j\psi\left(\frac{\mu_j}{\kappa}\right) + \frac{\mu_j}{\kappa}\boldsymbol{u}_j\psi(c_j) - \frac{\mu_j}{\kappa}\boldsymbol{u}_j\log(b_j)\right\}$$
$$= \sum_{j=1}^{q}\left\{\frac{\mu_j}{\kappa}\boldsymbol{u}_j\left[-\log(\kappa) - \psi\left(\frac{\mu_j}{\kappa}\right) + \psi(c_j) - \log(b_j)\right]\right\}.$$

The $\boldsymbol{\gamma}$ component of Hessian matrix, $\boldsymbol{H}_{\gamma\gamma}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\sum_{j=1}^{q}\left\{\frac{\mu_j}{\kappa}\boldsymbol{u}_j\left[-\frac{\mu_j}{\kappa}\boldsymbol{u}_j^T\psi'\left(\frac{\mu_j}{\kappa}\right) + \frac{\mu_j}{\kappa}\boldsymbol{u}_j^T\psi'(c_j)\right] + \frac{\mu_j}{\kappa}\boldsymbol{u}_j\boldsymbol{u}_j^T\left[-\log(\kappa) - \psi\left(\frac{\mu_j}{\kappa}\right) + \right.\right.$$
$$\left.\left.\psi(c_j) - \log(b_j)\right]\right\}$$
$$= \sum_{j=1}^{q}\left\{\frac{\mu_j}{\kappa}\boldsymbol{u}_j\boldsymbol{u}_j^T\left[\frac{\mu_j}{\kappa}\left(-\psi'\left(\frac{\mu_j}{\kappa}\right) + \psi'(c_j)\right) - \log(\kappa) - \psi\left(\frac{\mu_j}{\kappa}\right) + \psi(c_j) - \log(b_j)\right]\right\}.$$

Score and Hessian Contribution from κ

Relevant terms of the log-likelihood are

$$\sum_{j=1}^{q}\left\{-\frac{\mu_j}{\kappa}\log(\kappa) - \log\Gamma\left(\frac{\mu_j}{\kappa}\right) + \log\Gamma(c_j) - c_j\log(b_j)\right\}.$$

It follows that the score equation, $S_\kappa(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\sum_{j=1}^{q}\left\{-\mu_j\left[\frac{1}{\kappa^2} - \frac{1}{\kappa^2}\log(\kappa)\right] - \left(-\frac{\mu_j}{\kappa^2}\right)\psi\left(\frac{\mu_j}{\kappa}\right) + \left(-\frac{\mu_j}{\kappa^2}\right)\psi(c_j) - \left[\frac{c_j}{b_j}\left(-\frac{1}{\kappa^2}\right) - \right.\right.$$
$$\left.\left.\frac{\mu_j}{\kappa^2}\log(b_j)\right]\right\}$$
$$= \sum_{j=1}^{q}\left\{\frac{\mu_j}{\kappa^2}\left[-1 + \log(\kappa) + \psi\left(\frac{\mu_j}{\kappa}\right) - \psi(c_j) + \log(b_j)\right] + \frac{c_j}{b_j\kappa^2}\right\}.$$

The κ component of Hessian matrix, $\boldsymbol{H}_{\kappa\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is

$$\sum_{j=1}^{q}\left\{\frac{\mu_j}{\kappa^2}\left[\frac{1}{\kappa} - \frac{\mu_j}{\kappa^2}\psi'\left(\frac{\mu_j}{\kappa}\right) + \frac{\mu_j}{\kappa^2}\psi'(c_j) - \frac{1}{\kappa^2 b_j}\right] - \right.$$
$$\left.\frac{2\mu_j}{\kappa^3}\left[-1 + \log(\kappa) + \psi\left(\frac{\mu_j}{\kappa}\right) - \psi(c_j) + \log(b_j)\right] - \frac{b_j\mu_j + c_j(2b_j\kappa - 1)}{b_j^2\kappa^4}\right\}$$

$$= \sum_{j=1}^{q} \left\{ \frac{\mu_j}{\kappa^3} \Bigg[3 - \frac{\mu_j}{\kappa}\psi'\left(\frac{\mu_j}{\kappa}\right) + \frac{\mu_j}{\kappa}\psi'(c_j) - \frac{1}{\kappa b_j} + 2\Bigg(-\log(\kappa) - \psi\left(\frac{\mu_j}{\kappa}\right) + \psi(c_j) - \log(b_j) \Bigg) \Bigg] \frac{b_j\mu_j + c_j(2b_j\kappa - 1)}{b_j^2\kappa^4} \right\} \Bigg\}.$$

Hessian Contribution from Off Diagonal Elements

$\boldsymbol{H}_{\alpha\gamma}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is given by

$$-\sum_{j=1}^{q} \left\{ \frac{\boldsymbol{h}_j}{b_j} \frac{\mu_j \boldsymbol{u}_j^T}{\kappa} \right\}.$$

$\boldsymbol{H}_{\alpha\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is given by

$$-\sum_{j=1}^{q} \left\{ \boldsymbol{h}_j \frac{b_j(-\frac{\mu_j}{\kappa^2}) - c_j(\frac{-1}{\kappa^2})}{b_j^2} \right\}$$

$$= -\sum_{j=1}^{q} \left\{ \frac{\boldsymbol{h}_j(b_j\mu_j - c_j)}{\kappa^2 b_j^2} \right\}.$$

$\boldsymbol{H}_{\gamma\kappa}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \kappa)$ is given by

$$-\sum_{j=1}^{q} \left\{ \frac{\mu_j \boldsymbol{u}_j}{\kappa} \left[-\frac{1}{\kappa} + \frac{\mu_j}{\kappa^2}\psi'\left(\frac{\mu_j}{\kappa}\right) - \frac{\mu_j}{\kappa^2}\psi'(c_j) + \frac{\frac{1}{\kappa^2}}{b_j} \right] - \frac{\mu_j \boldsymbol{u}_j}{\kappa^2} \left[-\log(\kappa) - \psi\left(\frac{\mu_j}{\kappa}\right) + \psi(c_j) - \log(b_j) \right] \right\}$$

$$= -\sum_{j=1}^{q} \left\{ \frac{\mu_j \boldsymbol{u}_j}{\kappa^2} \left[-1 + \frac{\mu_j}{\kappa}\psi'\left(\frac{\mu_j}{\kappa}\right) - \frac{\mu_j}{\kappa}\psi'(c_j) + \frac{1}{\kappa b_j} + \log(\kappa) + \psi\left(\frac{\mu_j}{\kappa}\right) - \psi(c_j) + \log(b_j) \right] \right\}.$$

Note that all components of the score vector and Hessian matrix are functions of all parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and κ. Although explicit solutions cannot be found when we set the score vector to zero, we have derived analytic expression for the first and second derivatives. Thus, the equations can be solved numerically. Newton-Raphson method is chosen as it is a popular root-finding algorithm.

Denote $\boldsymbol{\theta}$ to be

$$\begin{pmatrix} \boldsymbol{\alpha} \\ \gamma \\ \kappa \end{pmatrix}.$$

The iterative scheme for finding the maximum likelihood estimate of $\boldsymbol{\theta}$ is given in Algorithm 2. Standard error of the MLE of $\boldsymbol{\theta}$ can then be obtained by taking the square root of the corresponding diagonal elements of the inverse Hessian matrix $\boldsymbol{H}^{-1}$. We also incorporate the step-halving procedure (in the second **while** statement) in Algorithm 2 to facilitate convergence. If a Newton-Raphson step leads to a decrease in the log-likelihood, the change in parameter estimates is repeatedly halved until the updated estimates result in an increase in the log-likelihood. Step-halving is widely implemented in many statistical routines to alleviate convergence issues, such as within the *glm* function in the **stats** package and within the *glm2* function in the **glm2** package (Marschner, 2014) in R.

```
Initialize $\boldsymbol{\theta}$.
Input: 2 sets of sufficient statistics for individuals with event $\left(\boldsymbol{X}_{\text{event}}^T\mathbf{1};\ \ \omega_j\ \forall j\right)$.
Cycle:
while relative differences in $\ell(\boldsymbol{\theta})$ is not negligible do
    Input: 3 sets of summary statistics for population at risk for each area:
    $\sum_{i=1}^n \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\ \forall j;\ \ \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\ \forall j;\ \ \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\ \forall j;$
    1 summary statistics for each area: $\exp(\boldsymbol{u}_j^T\boldsymbol{\gamma})\ \forall j;$
    $\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} - \{\boldsymbol{H}(\boldsymbol{\theta}_{m-1})\}^{-1}\{\boldsymbol{S}(\boldsymbol{\theta}_{m-1})\};$
    if $\kappa < 0$ then
        set $\kappa =$ very small value (we use 0.001);
    end
    while $\ell(\boldsymbol{\theta}_m) - \ell(\boldsymbol{\theta}_{m-1}) < 0$ do
        set $\boldsymbol{\theta}_m = (\boldsymbol{\theta}_{m-1} + \boldsymbol{\theta}_m)/2;$
    end
end
```

Algorithm 2: *Iterative scheme for obtaining the maximum likelihood estimates for the extended Gamma-Poisson Model.*

Chapter 6

Outlook

In this thesis we have proposed a class of modelling framework for *generalized linear mixed models* (GLMMs) that is able to incorporate unit-level covariates while maintaining a closed-form representation of the marginal likelihood. We focus primarily on two-level data where the random effects are mapped uniquely onto the grouping structure and are independent between groups. We refer to the proposed framework as *conjugate generalized linear mixed models* (CGLMMs). For multinomial mixed models that do not belong to the CGLMMs framework, we developed an approximating approach based on Poisson CGLMMs and derived an estimation procedure that exploit existing functions for fitting generalized linear models. The proposed CGLMMs framework is applied to discrete choice models and privacy preservation in large-scale administrative databases. We also compared the performance of GLMMs (with normal random effects) vs. CGLMMs (with log-gamma random effects) in terms of estimation of fixed effects and prediction of random effects. Under a Poisson distribution, the performance of GLMMs vs. CGLMMs is shown to be quite comparable.

Here we provide a high level summary of the opportunities for further research in this field:

- *Bayesian approach*: This thesis restricted attention to the frequentist paradigm, but it is arguable that the methodologies developed can also be applied to the Bayesian setting.
- *Multinomial mixed models vs. supervised classification methods*: Compare the

classification performance of multinomial mixed models vs. supervised learning methods such as linear discriminant analysis, support vector machines, tree classifiers, random forests and nearest neighbour classifiers, for the case of correlated responses.

- *Privacy*: Derive methodologies for fitting GLMMs via sufficient and summary statistics, where the marginal likelihood is approximated via Laplace approximation or Penalized Quasi-Likelihood.
- *Small area estimation*: Extend the results of Chapter 5 for complex surveys using a pseudolikelihood approach to accommodate inverse probability weights.

Overall, this thesis has made significant advances in inferential tools for correlated data. The proposed CGLMMs framework adds to the existing models for correlated data, and can be a good alternative when dealing with a large amount of data and/or privacy is of a concern.

Bibliography

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

Australian Bureau of Statistics (2008). Information Paper: An Introduction to Socio-Economic Indexes for Areas (SEIFA). *ABS Catalogue No. 2039.0*.

Australian Bureau of Statistics (2012). Australian Statistical Geography Standard (ASGS): Correspondences. *ABS Catalogue No. 1270.0.55.006*.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D 43*(4), 495–504.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*(1), 1–122.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Bühlmann, P., P. Drineas, M. Kane, and M. van der Laan (Eds.) (2016). *Handbook of Big Data*. CRC Press.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. Brooks/Cole.

Chambers, R. and R. Clark (2012). *An Introduction to Model-Based Survey Sampling with Applications.* Oxford University Press.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*(4), 1655–1684.

Chen, Y., G. Dong, J. Han, J. Pei, B. W. Wah, and J. Wang (2006). Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering 18*(12), 1585–1599.

Chen, Z. and L. Kuo (2001). A note on the estimation of the multinomial logit model with random effects. *The American Statistician 55*(2), 89–95.

Christiansen, C. L. and C. N. Morris (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association 92*(438), 618–632.

Chu, E., A. Keshavarz, and S. Boyd (2013). A distributed algorithm for fitting generalized additive models. *Optimization and Engineering 14*(2), 213–224.

Consonni, G. and P. Veronese (1992). Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association 87*(420), 1123–1127.

Coull, S., M. Collins, C. Wright, F. Monrose, and M. Reiter (2007). On web browsing privacy in anonymized NetFlows. *Proceedings of 16th USENIX Security Symposium.*

Croissant, Y. (2013). *Estimation of multinomial logit models in* R*: The mlogit Packages.* R package version 0.2-4.

Crowder, M. (1978). Beta-binomial ANOVA for proportions. *Journal of the Royal Statistical Society: Series C 27*(1), 34–37.

Daniels, M. J. and C. Gatsonis (1997). Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine 16*(20), 2311–2325.

de Rooij, M. and H. M. Worku (2012). A warning concerning the estimation of multinomial logistic models with correlated response in SAS. *Computer Methods and Programs in Biomedicine 107*(2), 341–346.

Demidenko, E. (2013). *Mixed models: theory and applications with R.* John Wiley & Sons.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data.* Oxford.

Einav, L. and J. Levin (2014). Economics in the age of big data. *Science 346*(6210), 479–480.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), 799–815.

Fienberg, S. E. (2006). Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science 21*(2), 143–154.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). *Applied Longitudinal Analysis.* Wiley.

Fong, Y., H. Rue, and J. Wakefield (2010). Bayesian inference for generalized linear mixed models. *Biostatistics 11*(3), 397–412.

Fornaroli, R., R. Cabrini, L. Sartori, F. Marazzi, D. Vracevic, V. Mezzanotte, M. Annala, and S. Canobbio (2015). Predicting the constraint effect of environmental characteristics on macroinvertebrate density and diversity using quantile regression mixed model. *Hydrobiologia 742*(1), 153–167.

Fullerton, A. S. and J. Xu (2016). *Ordered Regression Models: Parallel, Partial, and Non-Parallel Alternatives*. CRC Press.

Gelman, A. and J. Hill (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Goldstein, H. (2011). *Multilevel Statistical Models*. Wiley.

Gong, X., A. van Soest, and E. Villagomez (2004). Mobility in the urban labor market: A panel data analysis for Mexico. *Economic Development and Cultural Change 53*(1), 1–36.

Guha, S., R. Hafen, J. Rounds, J. Xia, J. Li, B. Xi, and W. S. Cleveland (2012). Large complex data: Divide and recombine (D&R) with RHIPE. *Stat 1*(1), 53–67.

Gutiérrez-Pẽna, E. and A. Smith (1995). Conjugate parameterizations for natural exponential families. *Journal of the American Statistical Association 90*(432), 1347–1356.

Hann, P. and A. Uhlendorff (2006). Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. *The Stata Journal 6*(2), 229–245.

Hartzel, J., A. Agresti, and B. Caffo (2001). Multinomial logit random effects models. *Statistical Modelling 1*(2), 81–102.

Harville, D. A. (1977). Maximum likelihood apapproach to variance component estimation and to related problems. *Journal of the American Statist 72*(358), 320–338.

He, Z. and D. Sun (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics 5*(3), 223–236.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine 22*(9), 1433–1446.

Hedeker, D. and R. D. Gibbons (2006). *Longitudinal data analysis.*

Hensher, D. A., J. M. Rose, and W. H. Greene (2015). *Applied Choice Analysis.* Cambridge.

Hodges, J. S. (2013). *Richly parameterized linear models: additive, time series and spatial models using random effects.* CRC Press.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson, and D. Craig (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics 4*(8), e1000167.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P. P. De Wolf (2012). *Statistical Disclosure Control.* John Wiley & Sons.

Jain, D. C., N. J. Vilcassim, and P. K. Chintagunta (1994). A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics 12*(3), 317–328.

Jennrich, R. I. and M. D. Schluchter (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics 42*(4), 805–820.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B 76*(4), 795–816.

Kleinman, J. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association 68*(341), 46–54.

Kuss, O. and D. McLerran (2007). A note on the estimation of the multinomial logistic model with correlated response in SAS. *Computer Methods and Programs in Biomedicine 87*(3), 262–269.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Leaver, V. (2009). Implementing a method for automatically protecting user-defined census tables. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

Lee, J. Y. L., J. J. Brown, and L. M. Ryan (2017). Sufficiency revisited: Rethinking statistical algorithms in the big data era. *The American Statistician 71*(3), xxx–xxx (page numbers pending).

Lee, J. Y. L., P. J. Green, and L. M. Ryan (2017a). Conjugate generalized linear mixed models for clustered data. *arXiv preprint* (1709.06288).

Lee, J. Y. L., P. J. Green, and L. M. Ryan (2017b). On the "Poisson trick" and its extensions for fitting multinomial regression models. *arXiv preprint* (1706.09523).

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear model. *Journal of the Royal Statistical Society: Series B 58*(4), 619–678.

Lin, N. and R. Xi (2011). Aggregated estimating equation estimation. *Statistics and Its Interface 4*(1), 73–83.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*(435), 1007–1016.

Lubell-Doughtie, P. and J. Sondag (2013). Practical distributed classification using the alternating direction method of multipliers algorithm. In *Big Data, 2013 IEEE International Conference on*, pp. 773–776.

Luts, J., T. Broderick, and M. Wand (2014). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics 23*(3), 589–615.

Ma, R. and B. Jorgensen (2007). Nested generalized linear mixed models: an orthodox best linear unbiased predictor approach. *Journal of the Royal Statistical Society: Series B 69*(4), 625–641.

Malchow-Møller, N. and M. Svarer (2003). Estimation of the multinomial logit model with random effects. *Applied Economics Letters 10*(7), 389–392.

Marks, H. M. and S. M. Printy (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly 39*(37), 370–397.

Marschner, I. C. (2014). *glm2: Fitting generalized linear models.* R package version 1.1.2.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* Chapman & Hall.

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*(437), 162–170.

McCulloch, C. E. and J. M. Neuhaus (2011a). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science 26*(3), 388–402.

McCulloch, C. E. and J. M. Neuhaus (2011b). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics 67*, 270–279.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus (2008). *Generalized, linear and mixed models*. John Wiley & Sons.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*(2), 267–278.

Minsker, S., S. Srivastava, L. Lin, and D. Dunson (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31th International Conference on Machine Learning (ICML-14)*, pp. 1656–1664.

Miroshnikov, A., Z. Wei, and E. M. Conlon (2015). Parallel Markov Chain Monte Carlo for non-Gaussian posterior distributions. *Stat 4*(1), 304–319.

Molenberghs, G., G. Verbeke, C. G. Demétrio, and A. Vieira (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science 25*(3), 325–347.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Morrell, C. H., L. J. Brant, S. Sheng, and E. J. Metter (2012). Screening for prostate cancer using multivariate mixed-effects models. *Journal of Applied Statistics 39*(6), 1151–1175.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics 11*(2), 515–529.

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses.

Narayanan, A. and V. Shmatikov (2008). Robust de-anonymization of large sparse datasets. *Proceedings of the IEEE Symposium on Security and Privacy*, 111–125.

Neiswanger, W., C. Wang, and E. Xing (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Unvertainty in Artificial Intelligence (UAI-14)*, pp. 623–632.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A 135*, 370–384.

Neuhaus, J. M., C. E. McCulloch, and R. Boylan (2012). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in Medicine*.

O'Keefe, C. M. (2008). Privacy and the use of health data-reducing disclosure risk. *Electronic Journal of Health Informatics 3*(1), e5.

Perry, P. O. (2017). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B 79*(1), 267–291.

Pinheiro, J. C. and D. M. Bates (1978). *Mixed-Effects Models in S and S-Plus*. Springer.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabe-Hesketh, S. and A. Skrondal (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A 169*(4), 805–827.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal 2*(1), 1–21.

Rao, J. and I. Molina (2015). *Small area estimation*. Wiley.

Rao, P. S. (1997). *Variance components: mixed models, methodologies and applications.* CRC Press.

Richards, F. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society: Series B 23*(2), 469–475.

Ripley, B. and W. Venables (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models.* R package version 7.3-12.

Ronquist, F. and J. P. Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*(12), 1572–1574.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society: Series B 71*(2), 319–392.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management 11*(2), 78–88.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*(21), 2688–2690.

Stroup, W. W. (2013). *Generalized Linear Mixed Models.* CRC Press.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems 10*(5), 557–570.

Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*(3), 657–671.

Tierney, L. and J. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Train, K. E. (2009). *Discrete Choice Methods with Simulation.* Cambridge.

Tzavidis, N., M. G. Ranalli, N. Salvati, E. Dreassi, and R. Chambers (2015). Robust small area prediction for counts. *Statistical methods in medical research 24*(3), 373–395.

Venables, W. and B. Ripley (2002). *Modern Applied Statistics with S.* Springer.

Verbeke, G. and E. Lesaffre (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis 23*, 541–556.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data.* Springer.

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative pools. *International Journal of Forecasting 31*(3), 980–991.

White, T. (2009). *Hadoop: The definitive guide.* O'Reilly Media.

Wright, S. (1998). Multivariate analysis using the MIXED procedure (paper 229-23). In *Proceedings of the 23rd Annual SAS Users Group (SUGI) International Conference.*

Wu, L. (2010). *Mixed Effects Models for Complex Data.* CRC Press.

Xi, R., N. Lin, and Y. Chen (2009). Compression and aggregation for logistic regression analysis in data cubes. *IEEE Transactions on Knowledge and Data Engineering 21*(14), 479–492.

Xu, M., B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang (2014). Distributed Bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*, pp. 3356–3364.

Zaharia, M., M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica (2012). Resilient distributed datasets: A fault-tolerent abstraction for in-memory cluster computing. *9th USENIX Symposium on Networked Systems Design and Implementation.*

Zaharia, M., M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica (2010). Spark: Cluster computing with working sets. *2nd USENIX Workshop on Hot Topics in Cloud Computing.*

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distribution of random effects for longitudinal data. *Biometrics 57*, 795–801.

Zhang, P., P. X.-K. Song, A. Qu, and T. Greene (2008). Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. *Biometrics 64*, 29–38.

Zhang, Y. and J. Koren (2007). Efficient Bayesian hierarchical user modeling for recommendation systems. In *Proc. 30th A. Int. Association for Computing Machinery Special Interest Group on Information Retrieval Conf. Research and Development in Information Retrieval*, pp. 47–53. New York: Association for Computing Machinery.

Ziegler, A. (2011). *Generalized estimating equations.* Springer.