Generalized Linear Models

A Bayesian Perspective

$$
\begin{aligned}
f(y|\mu) &\propto \exp\{y\mu - \psi(\mu)\} \\
g(\mu) &= x^t\theta \\
\pi(\theta|y) &\propto f(y|\theta)\pi(\theta)
\end{aligned}
$$

edited by
Dipak K. Dey
Sujit K. Ghosh
Bani K. Mallick

GENERALIZED LINEAR MODELS

Biostatistics: A Series of References and Textbooks

Series Editor

Shein-Chung Chow

President, U.S. Operations
StatPlus, Inc.
Yardley, Pennsylvania

Adjunct Professor
Temple University
Philadelphia, Pennsylvania

1. *Design and Analysis of Animal Studies in Pharmaceutical Development*, edited by Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications*, James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded*, Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy*, edited by Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective*, edited by Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick

ADDITIONAL VOLUMES IN PREPARATION

Medical Biostatistics, Abhaya Indrayan and S.B. Sarmukaddam

GENERALIZED LINEAR MODELS

A BAYESIAN PERSPECTIVE

edited by

Dipak K. Dey
The University of Connecticut
Storrs, Connecticut

Sujit K. Ghosh
North Carolina State University
Raleigh, North Carolina

Bani K. Mallick
Texas A&M University
College Station, Texas

Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

CRC is an imprint of the Taylor & Francis Group,
an informa business

Library of Congress Cataloging-in-Publication Data

Dey, Dipak.
Generalized linear models : a Bayesian perspective / Dipak K. Dey, Sujit K. Ghosh, Bani K. Mallick.
p. cm. — (Biostatistics ; 5)
Includes index.
ISBN 0-8247-9034-0 (alk. paper)
1.Linear models (Statistics) 2. Bayesian statistical decision theory. I. Ghosh, Sujit K., 1970- II. Mallick, Bani K., 1965- III. Title. IV. Biostatistics (New York, N.Y.) ; 5.

QA276 .D485 2000
519.5'35—dc21 00-024055

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fax: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit)
10 9 8 7 6

Series Introduction

The primary objectives of the Biostatistics series are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and/or graduate courses in the area of biostatistics. This series will provide comprehensive and unified presentations of statistical designs, analyses, and interpretations of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary will be given of current and recently developed statistical methods and interpretations for both biostatisticians and researchers/scientists with minimal statistical knowledge who are engaged in applied biostatistics. The series is committed to providing easy-to-understand state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies will be illustrated through real examples.

Generalized linear models (GLMs) have been frequently used in pharmaceutical research and development, especially in clinical research and development for demonstration of the safety and efficacy of a pharmaceutical compound under investigation. However, the concept for the analysis of GLMs with mixed effects for categorical and/or longitudinal data is often misused or misinterpreted due to its complexity. This volume provides a comprehensive overview of key statistical concepts and methodologies including the Bayesian approach for analysis of GLMs including logistic regression and log-linear models from both a theoretical and a practical point of view. In addition, it includes important issues related to model diagnostics and variable selection in GLMs.

This volume serves as an intersection for biostatisticians, practitioners, and researchers/scientists by providing a good understanding of key statistical concepts and methodologies for analysis and interpretation of GLMs and GLMs with mixed effects as well. This volume is in compliance with good statistics practice (GSP) standards for good clinical practice (GCP) as required by most regulatory agencies for pharmaceutical research and development.

Shein-Chung Chow

Preface

Generalized Linear Models (GLMs) are widely used as flexible models in which a function of the mean response is "linked" to covariates through a linear predictor and in which variability is described by a distribution in an exponential dispersion family. These models include logistic regression and log-linear models for binomial and Poisson counts as well as normal, gamma and inverse-Gaussian models for continuous responses. Standard techniques for analyzing censored survival data such as Cox regression can also be handled within the GLM framework. Other topics closely related to GLMs include conditionally independent hierarchical models, graphical models, generalized linear mixed models (GLMMs) for estimating subject-specific effects, semi-parametric smoothing methods, pharmokinetic models and spatio-temporal models.

GLMs thus provide a versatile statistical modeling framework for medical and industrial applications, but questions remain about how the power of these models can be safely exploited when training data are limited. This volume demonstrates how Bayesian methodology allows complex models (ranging from simple logistic regression models to semi-parametric survival models for censored data) to be used without fear of the "over-fitting" that can occur with traditional GLM methods which are usually based on normal approximation theory. Insight into the nature of these complex Bayesian models is provided by theoretical investigations and practical implementations. Presupposing only basic knowledge of probability and statistics, this volume should be of interest to researchers in statistics, engineering and medicine.

This volume will serve as a comprehensive reference book for practitioners and researchers. Each part in the volume has chapters written by an expert in that particular topic, and the chapters are carefully edited to ensure that a uniform style of notation and presentation is used throughout. As a result, all researchers whose work uses GLM theory will find this an indispensable companion to their work and it will be the reference volume for this subject for many years to come. In particular, each chapter describes how to conceptualize, perform and criticize traditional GLMs from a **Bayesian perspective**. In addition, how to use modern computational methods to summarize inferences using simulation is elucidated.

The primary users of this volume include professionals in statistics and other related disciplines who work in the pharmaceutical industry, medical centers (including public health and epidemiology) and public and private research and academic institutions.

Our hope is that this volume will also help researchers identify areas of important future research and open new applications of generalized linear models using Bayesian approaches.

The papers in this volume are divided into six parts: General overview, extension of the GLMs, categorical and longitudinal data, semiparametric and nonparametric approaches, model diagnostics and variable selection and challenging problems.

In part I, Gelfand and Ghosh introduce Bayesian analysis of generalized linear models from its developments. Sun, Speckman and Tsutakawa describe random effects in generalized linear mixed model with fully explained examples. Ibrahim and Chen develop methods of prior elicitation and variable selection for generalized

linear mixed models with an example of pediatric pain data.

Chapters in Part II of the volume describe several extensions of GLMs. Ferreira and Gamerman introduce dynamic modeling approach for GLMs. They also lay out computational steps with two applications. Dey and Ravishanker extend GLMs in the presence of overdispersion. Both parametric and nonparametric approaches to overdispersed GLMs are considered. Nandram proposes Bayesian GLMs for inference about small areas and describes an application with mortality data of U.S.A.

Part III concerns modeling categorical and longitudinal data. Modeling dichotomous, polychotomous and count data are quite useful and challenging in the presence of correlation. In this part, first Chib describes methods for the analysis of correlated binary data using latent variables. He also describes three algorithms for implementation. Chen and Dey extend this to correlated ordinal data and propose algorithms for analysis of such data. Bayesian methods for time series count data are described by Ibrahim and Chen with an application to the analysis of pollen count. Albert and Ghosh propose and analyze item response modeling for categorical data. This part concludes with a case study using Bayesian probit and logit models by Landrum and Normand.

Part IV describes GLMs using rich classes of nonparametric and semiparametric approaches. Semiparametric GLMs are considered by Mallick, Denison and Smith using Bayesian approaches. The chapter by Basu and Mukhopadhyay presents a semiparametric method to model link functions for the binary response data. Next, Haro-López, Mallick and Smith develop a data adaptive robust link function. In the last chapter of Part IV, Kuo and Peng present a mixture-model approach to the analysis of survival data.

Part V deals with important issues relating to model diagnostics and variable selection in GLMs. In this part, the chapter by Dellaportas, Forster and Ntzoufras presents Bayesian variable selection in using Gibbs sampler. Next, Ibrahim and Chen describe variable selection methods for Cox models. This part is concluded by Dey and Chen on Bayesian model diagnostics for correlated binary data.

Part VI concludes the volume with challenging problems. Wakefield and Stephens develop a case study by incorporating errors-in-variable modeling. Iyengar and Dey review parametric and semiparametric approaches for the analysis of compositional data. Denison and Mallick describe classification trees from a Bayesian perspective and apply the algorithm on a case study problem. In the next chapter, Gelfand, Ravishanker and Ecker develop a new modeling and inference method for point-referenced binary spatial data. The part closes with the chapter by Best and Thomas on graphical models and software for GLMs.

The cooperation of all contributors in the timely preparation of their manuscripts is greatly appreciated. We decided early on that it was important to referee and critically evaluate the papers which were submitted for inclusion in this volume. For this substantial task, we relied on the service of numerous referees to whom we are most indebted. Among those whom we wish to acknowledge are Sudipto Banerjee, Pabak Mukerjee and Kaushik Patra.

Finally we thank the editors at Marcel Dekker, Inc. for considering our proposal. Our special thanks go to Debosri, Swagata and Mou for their encouragements in this project .

Dipak K. Dey, Storrs, CT, USA
Sujit K. Ghosh Raleigh, NC, USA
and
Bani K. Mallick College Station, TX, USA

Contents

Contributors

James Albert — Department of Mathematics and Statistics, Bowling Green State University, U.S.A
albert@bgnet.bgsu.edu

Sanjib Basu — Division of Statistics, Northern Illinois University, U.S.A
basu@math.niu.edu

Nicky Best — School of Medicine at St. Mary's, Imperial College, U.K.
n.best@ic.ac.uk

Ming-Hui Chen — Department of Mathematical Sciences, Worchester Polytechnic Institute, U.S.A
mhchen@wpi.edu

Siddhartha Chib — John M. Olin School of Business, Washington University, U.S.A
chib@simon.wustl.edu

Petros Dellaportas — Department of Statistics, Athens University of Economics and Business, Greece
petros@aueb.gr

David D.T. Dension — Department of Mathematics, Imperial College, U.K
d.denison@ma.ic.ac.uk

Dipak K. Dey — Department of Statistics, University of Connecticut, U.S.A
dey@stat.uconn.edu

Mark D. Ecker — Department of Mathematics, University of North Iowa, U.S.A
ecker@math.uni.edu

Marco A. R. Ferreira — Instituto de Matematica, Federal University of Rio de Janeiro, Brazil
marco@dme.ufrj.br

Jonathan J. Forster — Department of Statistics, Athens University of Economics and Business, Greece
forster@aueb.gr

Dani Gamerman Instituto de Matematica,
Federal University of Rio de Janeiro, Brazil
dani@dme.ufrj.br

Alan E. Gelfand Department of Statistics,
University of Connecticut, U.S.A
gelfand@uconnvm.uconn.edu

Malay Ghosh Department of Statistics,
University of Florida, U.S.A
ghoshm@stat.ufl.edu

Rubén A. Haro-López Department of Mathematics,
Imperial College, U.K
haro@gauss.rhon.itam.mx

Joseph G. Ibrahim Department of Biostatistics,
Harvard University, U.S.A
d.denison@ma.ic.ac.uk

Malini Iyengar Clinical Biostatistics,
Smith Kline Beecham Pharmaceuticals, U.S.A
malini@stat.uconn.edu

Lynn Kuo Department of Statistics,
University of Connecticut, U.S.A
lynn@stat.uconn.edu

Mary Beth Landrum Department of Health Care Policy,
Harvard University, U.S.A
landrum@hcp.med.harvard.edu

Bani K. Mallick Department of Statistics,
Texas A & M University, U.S.A
bmallick@stat.tamu.edu

Saurabh Mukhopadhyay Clinical Biostatistics,
Merck & Co., Inc., U.S.A
saurabh_mukhopadhyay@merck.com

Balgobin Nandram Department of Mathematical Sciences,
Worchester Polytechnic Institute, U.S.A
balnan@wpi.edu

Sharon Lise-Normand Department of Health care policy,
Harvard University, U.S.A
sharon@hcp.med.harvard.edu

Ioannis Ntzoufras Department of Statistics,
Athens University of Economics and Business, Greece
ioannis@aueb.gr

Fengchun Peng Credit Marketing,
Sears, Roebuck and Co., U.S.A
peng@stat.uconn.edu

Nalini Ravishanker Department of Statistics,
University of Connecticut, U.S.A
nalini@stat.uconn.edu

Adrian F.M. Smith Queen Mary and Westfield College,
University of London, U.K
AFMS.BSU_DOM

David Stephens Department of Mathematics,
Imperial College, U.K
d.stephens@ma.ic.ac.uk

Dong Chu Sun Department of Statistics,
University of Missouri Columbia, U.S.A
sun@stat.missouri.edu

Paul L. Speckman Department of Statistics,
University of Missouri Columbia, U.S.A
speckman@stat.missouri.edu

Andrew Thomas School of Medicine at St. Mary's,
Imperial College, U.K.
a.thomas@ic.ac.uk

Robert K. Tsutakawa Department of Statistics,
University of Missouri Columbia, U.S.A.
tsutakawa@stat.missouri.edu

Jon Wakefield School of Medicine at St. Mary's
Imperial College, U.K.
j.wakefield@ma.ic.ac.uk

Part I

General Overview

1

Generalized Linear Models: A Bayesian View

Alan E. Gelfand
Malay Ghosh

ABSTRACT Generalized linear models (GLMs) offer a unifying class of models which are widely used in regression analysis. Though initially introduced into the community from a classical viewpoint, in the past decade the Bayesian literature employing these models has witnessed rapid growth. This is due in part to their attractiveness within familiar hierarchical modeling but as well to the wide availability of high speed computing to implement simulation based fitting of these models. The objective of this chapter is to provide a brief, somewhat selective, summary and overview of this recent literature. In particular, we focus upon the range of proposed GLMs, prior specification, propriety of the resultant posterior, semiparametric approaches and model determination, i.e., model adequacy and model choice.

1. Introduction

Generalized linear models (GLMs), originally introduced by Nelder and Wedderburn (1972), provide a unifying family of models that is widely used for regression analysis. These models are intended to describe non-normal responses. In particular, they avoid having to select a single transformation of the data to achieve the possibly conflicting objectives of normality, linearity and homogeneity of variance. Important examples include the binary and the count data. Over the years, GLMs have expanded much in scope and usage, and are currently applied to a very broad range of problems which include analysis of multicategory data, dynamic or state-space extensions of non-normal time series and longitudinal data, discrete time survival data, and non-Gaussian spatial processes.

By now, there are several excellent textbooks discussing inference for GLMs from a classical point of view (McCullagh and Nelder, 1989; Fahrmeir and Tutz, 1991, Lindsey, 1995). These books provide a rich collection of estimation and hypothesis testing procedures for various parameters of interest primarily from a frequentist point of view. Breslow and Clayton (1993) have extended these models further by introducing random effects in addition to the fixed effects. The resulting models, usually referred to as generalized linear mixed models (GLMM's) have further widened the scope of application of GLMs for data analysis. Software routines such as PROC MIXED in SAS have facilitated the computations involved in the classical implementation of GLMs.

Bayesian methods for analyzing GLMs are of more recent origin. A general hierarchical Bayesian (HB) approach for such analysis began with West (1985) and Albert (1988), although various special cases were considered earlier. Leonard and Novick (1986) considered Bayesian analysis of two-way contingency tables, while Albert (1985) considered simultaneous estimation of several Poisson means via a hierarchical log-linear model.

The present article aims to review the Bayesian perspective with regard to GLMs. This includes a range of hierarchical model GLM specifications, approaches for Bayesian model fitting and techniques for model checking and model choice. In Section 2, after a brief introduction to the model and the classical inference procedure, we proceed to the discussion of some of the Bayesian models that have appeared in the literature. We compare the various priors that have been proposed, and discuss also the methods required for their implementation. This includes the work of Albert (1988), Ibrahim and Laud (1991), Dellaportas and Smith (1993), Zeger and Karim (1991), Ghosh, Natarajan, Stroud and Carlin (1998) among others. Section 3 discusses the propriety of posteriors under the different priors. In particular, we discuss the results of Ibrahim and Laud (1991), Natarajan and McCulloch (1995), Hobert and Casella (1996), Gelfand and Sahu (1999), Ghosh, Natarajan, Stroud and Carlin (1998). Section 4 discusses semiparametric and nonparametric Bayesian procedures for GLMs. Section 5 briefly looks at overdispersed GLMs. Finally, Section 6 addresses the issues of model diagnostics and model selection.

2. GLMs and Bayesian Models

2.1 GLMs

Consider measurements (discrete or continuous) for n individuals. For the ith individual, the response variable is denoted by y_i, and the corresponding vector of covariates is denoted by $\boldsymbol{x}_i$. Responses may be continuous real variables, or counts or binary. Fahrmeir and Tutz (1991) contains many interesting examples of binary and count data. As an example of binary data, they consider infection from births by caesarean section. The response variable is the occurrence or non-occurrence of infection. They consider three dichotomous covariates: (a) planned or unplanned caesarean, (b) presence or absence of risk factors such as having diabetes or being overweight, and (c) use or nonuse of antibiotics as prophylaxis. An example of count data involves the effect of two agents of immuno-activating ability that may induce cell differentiation (Piegorsch, Weinberg and Margolin, 1988). As response variable, one considers the number of cells that exhibited markers after exposure. The covariates are the agents TNF (tumor necrosis factor) and IFN (interferon). It is of interest to know whether these agents stimulate cell differentiation independently or whether there is an interaction effect.

There are certain distributional and structural assumptions associated with GLMs. The key distributional assumption is that conditional on the θ_i, the y_i are independent with pdf's belonging to the one-parameter exponential family, that is,

$$f(y_i|\theta_i) = \exp[a^{-1}(\phi_i)\{y_i\theta_i - \psi(\theta_i)\} + c(y_i;\phi_i)], \tag{1}$$

where the θ_i are unknown, but the $a(\phi_i)(>0)$ are known. The usual structural assumption is that $\theta_i = h(\boldsymbol{x}_i^T\boldsymbol{b})$, where h is a strictly increasing sufficiently smooth function, $\boldsymbol{b}(p\times 1)$ is the vector of unknown regression coefficients, and the $\boldsymbol{x}_i(p\times 1)$

are known design vectors of dimension p. The parameters θ_i are usually referred to as the canonical parameters. Important special cases include the binomial distributions with success parameters $p_i = \exp(\theta_i)/[1+ \exp(\theta_i)]$, $a(\phi_i) = 1$, and the Poisson distributions with means $\lambda_i = \exp(\theta_i)$, $a(\phi_i) = 1$. The $N(\mu_i, \sigma_i^2)$ distributions are also covered by (1) with $\theta_i = \mu_i, a(\phi_i) = \sigma_i^2$. The gamma and the inverse Gaussian distributions are other important special cases of (1).

The classical estimation procedure for GLMs is maximum likelihood. For simplicity, we assume that the ϕ_i are known and that $\boldsymbol{X}^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ has rank p. The likelihood function is given by

$$L(\boldsymbol{b}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\{y_i h(\boldsymbol{x}_i^T \boldsymbol{b}) - \psi(h(\boldsymbol{x}_i^T \boldsymbol{b}))\}\right]. \tag{2}$$

The corresponding score vector is

$$\frac{d \log L(\boldsymbol{b})}{d\boldsymbol{b}} = \sum_{i=1}^{n} a_i^{-1}(\phi_i)\{y_i - \psi'(h(\boldsymbol{x}_i^T \boldsymbol{b}))\} h'(\boldsymbol{x}_i^T \boldsymbol{b}) \boldsymbol{x}_i, \tag{3}$$

and the Fisher information matrix is

$$\boldsymbol{I}(\boldsymbol{b}) = E\left[-\frac{d^2 \log L}{d\boldsymbol{b} d\boldsymbol{b}^T}\right] = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}(\boldsymbol{b}) \boldsymbol{\Delta}^2(\boldsymbol{b}) \boldsymbol{X}, \tag{4}$$

where $\boldsymbol{D} = \text{Diag}\,(a^{-1}(\phi_i), \cdots, a^{-1}(\phi_n))$, $\boldsymbol{V}(\boldsymbol{b}) = \text{Diag}\,(\psi''(h(\boldsymbol{x}_1^T \boldsymbol{b})), \cdots, \psi''(h(\boldsymbol{x}_n^T \boldsymbol{b}))$, and $\boldsymbol{\Delta}(\boldsymbol{b}) = \text{Diag}\,(h'(\boldsymbol{x}_1^T \boldsymbol{b}), \cdots, h'(\boldsymbol{x}_n^T \boldsymbol{b}))$.

The maximum likelihood estimators are obtained as iterative solutions of the likelihood equations $\frac{d \log L(\mathbf{b})}{d\mathbf{b}} = \mathbf{0}$. If the log-likelihood $\ell(\boldsymbol{b}) = \log L(\boldsymbol{b})$ is concave, then the MLE is unique when there exists at least one $\hat{\boldsymbol{b}}$ within the admissible parameter set where $\ell(\boldsymbol{b})$ attains the local or global maximum.

The asymptotic theory of the MLE works in this situation as well. Under mild regularity conditions, the MLE $\hat{\boldsymbol{b}}$ of $\boldsymbol{b}$ is asymptotically $N(\boldsymbol{b}, n^{-1}\boldsymbol{I}^{-1}(\boldsymbol{b}))$.

2.2 Bayesian Models

For a Bayesian model associated with the likelihood (2), we require a prior for $\mathbf{b}$. A commonly used choice is $N(\boldsymbol{b}_0, \boldsymbol{\Sigma})$, where $\boldsymbol{b}_0$ and $\boldsymbol{\Sigma}$ are known. This prior appears for example in Dellaportas and Smith (1993). Then, writing $\boldsymbol{y} = (y_1, \cdots, y_n)^T$, the posterior of $\boldsymbol{b}$ is given by

$$\pi(\boldsymbol{b}|\boldsymbol{y}) \propto \exp\left[\sum_{1}^{n} a^{-1}(\phi_i)\{y_i h(\boldsymbol{x}_i^T \boldsymbol{b}) - \psi(h(\mathbf{x}_i^T b))\} - \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{b} - \boldsymbol{b}_0)\right]. \tag{5}$$

The above posterior is not analytically tractable. In fact, there does not exist any closed form expression for the norming constant. Also, finding posterior means, variances etc. by numerical integration is not easy even for moderate p. The most convenient approach seems to be the Markov Chain Monte Carlo (MCMC) numerical integration techniques which require generating samples from the posterior. They can be implemented in general using the Metropolis-Hastings algorithm, but if the posterior is log-concave, then one can also use the adaptive rejection sampling approach of Gilks and Wild (1992).

With little or no prior information, an alternative is to use noninformative priors. This implies that the posterior distribution is essentially the likelihood that the Bayesian analysis will be close to a likelihood analysis, possibly attractive to frequentists. One of the commonly used noninformative priors due to Laplace (1812) is $\pi_L(\boldsymbol{b}) \propto 1$. However, the following example due to Laud and Ibrahim (1991), shows that such a prior can sometimes lead to an improper posterior.

Example 1. Suppose $n = 1$, $p = 1$, and

$$f(y_1|x_1) = (bx_1)^{-1} \exp[-y_1/(bx_1)]; \quad y > 0, x_1 > 0, \quad b > 0.$$

Then if $\pi_L(b) \propto 1$

$$\pi_L(b|y_1) \propto (bx_1)^{-1} \exp[-y_1/(bx_1)]. \tag{6}$$

Integration with respect to b over $(0, \infty)$ gives $\int_0^\infty \pi_L(b|y_1)db = \int_0^\infty z^{-1} \exp\left(-\frac{y_1}{z}\right) dz = +\infty$, so that the posterior is improper.

Laud and Ibrahim (1991) proposed Jeffreys' prior for this problem given by $\pi_J(\boldsymbol{b}) \propto |\boldsymbol{I}(\boldsymbol{b})|^{1/2}$ with $\boldsymbol{I}(\boldsymbol{b})$ given in (4). They provided sufficient conditions under which the resulting posterior $\pi_J(\boldsymbol{b}|\boldsymbol{y})$ is proper.

The likelihood function given in (2) generalizes the normal fixed effects model to the one-parameter exponential family. Breslow and Clayton (1993) have extended this further to include mixed effects models. For a fixed effects model, under the link function $g \equiv h^{-1}$, we have $g(\theta_i) = \boldsymbol{x}_i^T \boldsymbol{b}$. In contrast, for the random effects model, one incorporates the random effects as well, and writes $g(\theta_i) = \boldsymbol{x}_i^T \boldsymbol{b} + \boldsymbol{z}_i^T \boldsymbol{u}_i$, where the $\boldsymbol{z}_i$ are also known and the $\boldsymbol{u}_i$ are i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$. Breslow and Clayton (1993) advocated penalized quasi-likelihood estimates (PQL) for estimating $\boldsymbol{b}$.

We note an example of generalized linear mixed effects model (GLMM), considered in Crowder (1978), and also in Breslow and Clayton (1993). This concerns data on the proportion of seeds that germinated on each of 21 plates arranged according to a 2×2 factorial layout by seed variety and type of root extract. It turns out in this example that the within-group variation exceeds that predicted by the binomial sampling theory. The heterogeneity due to plate-to-plate variability is accounted for by Crowder (1978) and Breslow and Clayton (1993) by means of a GLMM that employs the canonical link setting

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{b} + u_i,$$

$i = 1, \cdots, 21$, where $\boldsymbol{b}$ represents the fixed effects associated with seed and the extract, and the u_i, assumed to be iid $N(0, \sigma_u^2)$, represent random effects associated the plates.

The Bayesian procedure, as before assigns a $N(\boldsymbol{b}_0, \boldsymbol{\Sigma}_b)$ distribution to $\boldsymbol{b}$. More generally, a hierarchical Bayesian model is considered whereby one assigns distributions to $\boldsymbol{\Sigma}_b$ and σ_u. One option is to use an inverse Wishart distribution for $\boldsymbol{\Sigma}_b$ and inverse gamma distribution for σ_u^2. Such distributions can possibly be improper, but care must be exercised in order that the resulting posterior is proper.

Specifically, let $\boldsymbol{\Sigma}_b$ and σ_u^2 have independent priors

$$\pi(\boldsymbol{\Sigma}_b) \propto \exp[-\frac{1}{2} tr(\boldsymbol{\Psi} \boldsymbol{\Sigma}_b^{-1})] |\boldsymbol{\Sigma}_b|^{-\frac{1}{2}\nu}, \tag{7}$$

and

$$\pi(\sigma_u^2) \propto \exp(-\frac{a}{2\sigma_u^2})(\sigma_u^2)^{-\frac{g}{2}-1}. \tag{8}$$

We shall write symbolically $\boldsymbol{\Sigma}_b \sim IW(\boldsymbol{\Psi}, \nu)$ and $\sigma_u^2 \sim IG(\frac{a}{2}, \frac{g}{2})$.

The joint posterior of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$, $\boldsymbol{b}$, $\boldsymbol{\Sigma}_b$ and σ_u^2 is now given by

$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\Sigma}_b, \sigma_u^2|\boldsymbol{y}) \quad &\propto \quad \exp[\sum_{i=1}^{n}\{a^{-1}(\phi)(y_i\theta_i - \psi(\theta_i))\}] \\ &\times \quad (\sigma_u^2)^{-\frac{n+q}{2}-1}\exp[-\frac{\sum_{i=1}^{n}(\theta_i - \boldsymbol{x}_i^T\boldsymbol{b})^2 + a}{2\sigma_u^2}] \\ &\times \quad |\boldsymbol{\Sigma}_b|^{-\frac{p+\nu}{2}}\exp[-\frac{1}{2}tr\{(\boldsymbol{b}-\boldsymbol{b}_0)(\boldsymbol{b}-\boldsymbol{b}_0)^T + \boldsymbol{\Psi}\}\boldsymbol{\Sigma}_b^{-1}]. \quad (9)\end{aligned}$$

This posterior is analytically intractable, and one needs numerical integration for posterior analysis. Gibbs sampling (Gelfand and Smith, 1990; Gelfand, Hill, Racine-Poon and Smith, 1990) has proved to be very useful for implementation of the Bayesian model fitting. This requires sampling from the full conditionals

$$[\boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\Sigma}_b, \sigma_u^2, \boldsymbol{y}] \sim N((\sigma_u^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_b^{-1})^{-1}(\sigma_u^{-2}\boldsymbol{X}^T\boldsymbol{\theta} + \boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}_0), (\sigma_u^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_b^{-1})^{-1}]; \quad (10)$$

$$[\boldsymbol{\Sigma}_b|\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2, \boldsymbol{y}] \sim IW(\boldsymbol{\Psi} + (\boldsymbol{b}-\boldsymbol{b}_0)(\boldsymbol{b}-\boldsymbol{b}_0)^T, \nu + p); \quad (11)$$

$$[\sigma_u^2|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\Sigma}_b, \boldsymbol{y}] \sim IG(\frac{\sum_{i=1}^{n}(\theta_i - \boldsymbol{x}_i^T\boldsymbol{b})^2 + a}{2}, \frac{n+g}{2}); \quad (12)$$

$$\pi(\theta_i|\theta_j(j \neq i, \boldsymbol{b}, \boldsymbol{\Sigma}_b, \sigma_u^2, \boldsymbol{y}) \propto \exp[a^{-1}(\phi)(y_i\theta_i - \psi(\theta_i)) - \frac{(\theta_i - \boldsymbol{x}_i^T\boldsymbol{b})^2}{2\sigma_u^2}], \quad (13)$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$. It is easy to generate samples from the normal, inverse Wishart and the inverse gamma distributions. The only nonstandard conditionals are the $\pi(\theta_i|.)$, $i = 1, \cdots, n$, which are known only up to multiplicative constants. One can use the Metropolis-Hastings algorithm to generate samples. Alternately, one can use the adaptive rejection sampling of Gilks and Wild (1992) since these posteriors are log-concave.

The above model should be contrasted to that of Albert (1988). Albert (1988) begins with the likelihood given in (1), but does not model the θ_i as $\theta_i = h(\boldsymbol{x}_i^T\boldsymbol{b})$. Instead, he considers independent conjugate priors for the θ_i at the first stage, namely

$$\Pi(\theta_i|m_i, \lambda) = \exp[\lambda(m_i\theta_i - \psi(\theta_i)) + k(m_i; \lambda)] \quad (14)$$

An easy calculation shows that $E[\psi'(\theta_i)] = m_i$. He models the prior means m_i as $m_i = h(\boldsymbol{x}_i^T\boldsymbol{b})$. Thus Albert moves the GLM to the second stage specification, that is, for the θ_i's rather than the customary first stage specification for the y_i's. The remaining prior parameter λ is a precision parameter that reflects the strength of one's prior beliefs about the means m_i. As λ approaches infinity, the prior distribution of the $\psi'(\theta_i)$ becomes increasingly concentrated about the mean m_i, and the Bayesian model approaches the first stage specification in (2).

To complete the prior specification at the second stage, a distribution needs to be assigned to $\boldsymbol{b}$ and λ. West (1985) assigns a normal distribution to $\boldsymbol{b}$, and a chi-squared distribution to λ. Albert (1988) assigns instead the prior $\Pi(\boldsymbol{b}, \lambda) \propto (1+\lambda)^{-2}$, that is a priori $\boldsymbol{b}$ and λ are independent with $\boldsymbol{b} \sim$ uniform (R^p), and $\Pi(\lambda) \propto (1+\lambda)^{-2}$, a heavy tailed prior with infinite first moment.

With Albert's model, it is possible to calculate

$$E[\psi'(\theta_i)|\boldsymbol{b}, \lambda, \boldsymbol{y}] = [a^{-1}(\phi_i)y_i + \lambda m_i]/[a^{-1}(\phi_i) + \lambda].$$

This implies

$$E[\psi'(\theta_i)|\boldsymbol{y}_i] = y_i - E\left[\frac{\lambda}{a^{-1}(\phi_i) + \lambda}(y_i - m_i)|\boldsymbol{y}\right]. \quad (15)$$

Again, explicit evaluation of the conditional expectation given in the right hand side of (16) is difficult. Albert uses three approximations: (a) Laplace's method, (b) Quasi-likelihood approach, and (c) Brooks's method. A detailed application is given for the binomial-logit hierarchical model.

Albert's model differs from the one given in (5), (7) and (8) in that he connects the regression variables not directly with the parameters of interest, but indirectly with the prior means. The approximations that he considers are attractive alternatives to direct numerical integration, especially when p is large. On the other hand, these approximations do not appear to be simpler than using MCMC integration methods.

A slight variant of the model given in (1), (5), (7) and (8) is due to Zeger and Karim (1991). They consider a stratified sampling situation where y_{ij} is the response of the jth unit in the ith stratum. (In a longitudinal data or repeated measurements situation, j denotes the jth measurement on the ith subject). The corresponding vector of auxiliary characteristics or covariates is denoted by $\boldsymbol{x}_{ij}(p \times 1)$. The one-parameter exponential family model is now given by

$$f(y_{ij}|\theta_{ij}) = \exp[a^{-1}(\phi_{ij})\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\} + c(y_{ij};\phi_{ij})], \tag{16}$$

where the link function $h(.)$ yields $h(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{b} + \boldsymbol{z}_{ij}^T\mathbf{u}_i, j = 1, \cdots, n_i; i = 1, \cdots, m$. The $\boldsymbol{u}_i$ are iid $N(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

At the final stage of the hierarchical model, one assigns mutually independent priors to $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_u$ with $\boldsymbol{b} \sim$ uniform (R^p) and $\boldsymbol{\Sigma}_u \sim IW(1\boldsymbol{\Psi}, \nu)$. This posterior is also analytically intractable, and Gibbs sampling can be used to generate samples from the necessary conditionals.

The Zeger-Karim formulation does not include possible error in misspecifying the model. Ghosh, Natarajan, Stroud and Carlin (1998) consider slightly more general modeling of $g(\theta_{ij})$, namely $h(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{b} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_i + e_{ij}$. The errors e_{ij} account for model misspecification. A special case of the latter model will be discussed in the next section for studying the propriety of posteriors.

3. Propriety of Posteriors

If only proper priors are used, then one necessarily gets proper posteriors. However, as mentioned earlier, Bayesian analysis often relies on diffuse and flat priors which are mostly improper. In such instances, it is imperative to verify the propriety of posteriors. Otherwise, descriptive measures such as moments, quantiles etc. of the posteriors do not carry any meaning. Checking the propriety of posteriors is all the more important when the Bayesian procedure is implemented via MCMC technique as it may so happen that all the full conditionals are proper distributions, and yet the posterior is improper. (See Casella and George (1992) for an elementary example).

As shown in the previous section, Laplace's prior does not necessarily lead to a proper posterior. The same comment applies to Jeffreys' prior. Ibrahim and Laud (1991) have investigated conditions under which Jeffreys' prior leads to proper posteriors for GLMs. We present their main result below.

THEOREM 1. Consider the likelihood function given in (2), and Jeffreys' prior $\Pi_J(\boldsymbol{b}) \propto |\boldsymbol{I}(\boldsymbol{b})|^{1/2}$, where $\boldsymbol{I}(\boldsymbol{b})$ is given in (4). Assume that $r(\boldsymbol{X}) = p$ and the likelihood function is bounded. Then, a sufficient condition that the posterior dis-

tribution $\Pi(\boldsymbol{b}|\boldsymbol{y})$ is proper is that the integrals

$$\int \exp[(y_i\theta - \psi(\theta))/a(\phi_i)][\psi''(\theta)]^{1/2}d\theta \tag{17}$$

are finite for all $i = 1, \cdots, n$.

Ibrahim and Laud (1988) also show that a necessary and sufficient condition for Π_J to be proper is that $\int_\Theta [\psi''(\theta)]^{1/2}d\theta$ is finite where Θ denotes the parameter space. Dey, Gelfand and Peng (1997) extend these results to the case of overdispersed GLMs (see Section 5). Hobert and Casella (1996) have provided conditions ensuring propriety of posteriors for normal models. Natarajan and McCulloch (1995) have considered a version of a hierarchical model for binary data which ensures the propriety of posteriors.

Ghosh et al. (1998) have considered the model given in (14), but $h(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{b} + u_i + e_{ij} (j = 1, \cdots, n_i; i = 1, \cdots, m)$. The u_i and the e_{ij} are mutually independent with the u_i iid $N(0, \sigma_u^2)$ and the e_{ij} are iid $N(0, \sigma^2)$. Also, $\boldsymbol{b}, \sigma_u^2$ and σ^2 are mutually independent with $\boldsymbol{b} \sim$ uniform $(R^p), \sigma_u^2 \sim IG\left(\frac{1}{2}a, \frac{1}{2}g\right)$ and $\sigma^2 \sim IG\left(\frac{1}{2}c, \frac{1}{2}d\right)$.

Ghosh et al. (1998) provide sufficient conditions for the propriety of posteriors under this model. A slightly more general version of their theorem is proved in Ghosh and Natarajan (1998) which we present below.

THEOREM 2: Assume that $f(y_{ij}|\theta_{ij})$ is bounded for all $j = 1, \cdots, n_i; i = 1, \cdots, m$. Let $S = \{(i, j) : \int f(y_{ij}|\theta_{ij})d\theta_{ij} < \infty\}$, and $s =$ cardinality of S. Assume that $s \geq 1$. Then the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2, \sigma^2|\boldsymbol{y})$ is proper if $a > 0, c > 0, m + g > 0$ and $s + d > p$.

In a more recent article, Gelfand and Sahu (1999) have linked the issue of propriety of posteriors with Bayesian identifiability. Suppose the Bayesian model is denoted by the likelihood $L(\boldsymbol{\theta}; y)$ and the prior $\pi(\boldsymbol{\theta})$. Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Following Dawid (1979), if $\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{y}) = \pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$, we say that $\boldsymbol{\theta}_2$ is not identifiable. This means that if observing data $\boldsymbol{y}$ does not increase our prior knowledge about $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$, then $\boldsymbol{\theta}_2$ is not identified by the data. Noting that

$$\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{y}) \propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_1), \tag{18}$$

$\boldsymbol{\theta}_2$ is nonidentifiable if and only if $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})$ is free of $\boldsymbol{\theta}_2$, that is $L(\boldsymbol{\theta}; \boldsymbol{\theta}_2; \boldsymbol{y}) = L(\boldsymbol{\theta}_1, \boldsymbol{y})$. Hence, Dawid's formal definition of non-identifiability is equivalent to lack of identifiability of the likelihood. We may also observe that if $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})$ is free of $\boldsymbol{\theta}_2$, then the posterior

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y}) \propto L(\boldsymbol{\theta}_1; \boldsymbol{y})\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1|\boldsymbol{y})\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) \tag{19}$$

is proper if and only if both $\pi(\boldsymbol{\theta}_1|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ are proper.

To see how (19) works for GLMs suppose that rank $(\boldsymbol{X}) = r < p$. Then it is possible to make a one-to-one transformation from $\boldsymbol{b}$ to $(\boldsymbol{\delta}, \boldsymbol{\rho})$ such that (2) has the alternate representation

$$L(\boldsymbol{\delta}, \boldsymbol{\rho}) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\{\boldsymbol{u}_i h(\boldsymbol{x}_{0i}^T\boldsymbol{\delta}) - \psi(h(\boldsymbol{x}_{0i}^T\boldsymbol{\delta}))\}\right], \tag{20}$$

where $\boldsymbol{X}_0^T = (\boldsymbol{x}_{01}, \cdots, \boldsymbol{x}_{on})$ is a $r \times n$ matrix of rank r. Thus the likelihood does not depend on $\boldsymbol{\rho}$. Since propriety of $\pi(\boldsymbol{b}|\boldsymbol{y})$ is equivalent to the propriety of $\pi(\boldsymbol{\delta}, \boldsymbol{\rho}|\boldsymbol{y})$, it follows from (19) and (20) that $\pi(\boldsymbol{b}|\boldsymbol{y})$ is proper if and only both $\pi(\boldsymbol{\delta}|\boldsymbol{y})$ and $\pi(\boldsymbol{\rho}|\boldsymbol{\delta})$ are proper. If the latter holds, all we need to verify is that $\pi(\boldsymbol{\delta}|\boldsymbol{y})$ is proper.

Gelfand and Sahu (1999) have shown that with the canonical link $\theta_i = \boldsymbol{x}_i^T \boldsymbol{b}$, if $\text{rank}(\boldsymbol{X}) = r < p$, a sufficient condition for a proper $\pi(\boldsymbol{b}|\boldsymbol{y})$ is that the likelihood is bounded, and that at least r of the y_i's belong to the interiors of their respective domains.

4. Semiparametric GLMs

Sections 2 and 3 contain discussion of fully parametric Bayesian GLMs. In this section, we enrich this class of models by wandering nonparametrically near (in a suitable sense) this class. As a result, parts of the modeling are captured parametrically, in particular, the linear regression structure on some monotonically increasing transformed scale of the canonical parameters. Other aspects such as the link function or the distribution of the random effects are specified nonparametrically. This is an example of what has now become known as semiparametric regression modeling.

One of the difficulties with Bayesian modeling in the nonparametric case is that, unlike its parametric counterpart where the dimension of the parameter space is finite, nonparametric modeling requires an "infinite dimensional" parameter. Thus, the Bayesian approach, in assuming all unknowns are random, requires an infinite dimensional stochastic specification. However, significant advancement of research in this area during the past twenty five years or so has provided tractable ways to make such specifications. Further, more recent advances in Bayesian computation enable the fitting of models incorporating these specifications and even extensions of these specifications.

The various probabilistic specifications which yield Bayesian nonparametric modeling include discrete mixtures, Dirichlet processes, mixtures of Dirichlet processes, Polya tree distributions, Gamma processes, extended Gamma processes and Beta processes. Gelfand (1998) contains a review of all these. For brevity, we will discuss only the discrete mixtures and mixtures of Dirichlet processes in the context of conditionally independent hierarchical GLMs.

We may note that the basic object which we are attempting to model is an unknown function, say $g(\cdot)$. The parametric approach writes g as $g(\cdot, \theta)$, $\theta \epsilon \Theta$ and then places a prior distribution over $\theta \epsilon \Theta$. The nonparametric approach assumes only that $g \in \mathcal{G}$, where $\mathcal{G}$ is some class of functions. A Bayesian approach requires assigning a prior over the elements of $\mathcal{G}$. In what follows, we illustrate this with examples where the elements of $\mathcal{G}$ are monotone functions.

We begin our discussion with mixture models. First we notice that modeling a strictly monotone function g is equivalent to modeling a distribution function. For instance, if the range of g is R^1, then $T(g(\cdot))$ with $T(z) = k_1 \exp(k_2 z)/[1 + k_1 \exp(k_2 z)]$ with $k_1 > 0, k_2 > 0$ is a df. Similarly, if the range of g is R^+, then $T(g(\cdot))$ with $T(z) = k_1 z^{k_2}(1 + k_1 z^{k_2})^{-1}$ with $k_1 > 0, k_2 > 0$ is a df.

The mixture model approach models an unknown df using a dense class of mixtures of standard distributions. For instance, Diaconis and Ylvisaker (1985) observe that discrete mixture of Beta densities provide a dense class of models for densities on $[0, 1]$.

As a special case, consider modeling the link function h in a generalized linear model. Mallick and Gelfand (1994) suggested modeling $h(\theta)$ by

$$T(h(\theta)) = \sum_{\ell=1}^{r} w_\ell IB(T(h_0(\theta)); \ c_\ell, d_\ell), \tag{21}$$

where r is the number of mixands, $w_\ell \geq 0, \sum_{\ell=1}^{r} w_\ell = 1$ are the mixing weights, $IB(u,c,d) = \int_0^u \{x^{c-1}(1-x)^{d-1}/B(c,d)\}dx$, and h_0 is a centering function for h. Then (21) provides a generic member of the dense class. Inversion of (21), provides a generic h. Since h is determined by specification of $r, \boldsymbol{w}^{(r)} = (w_1, \cdots, w_r)$, $\boldsymbol{c}^{(r)} = (c_1, \cdots, c_r)$ and $\boldsymbol{d}^{(r)} = (d_1, \cdots, d_r)$, introducing a distribution on g requires specification of a distribution of the form $f(r)f(\boldsymbol{w}^{(r)}, \boldsymbol{c}^{(r)}\boldsymbol{d}^{(r)}|r)$. Mallick and Gelfand (1994) suggest-fixing $c^{(r)}$ and $d^{(r)}$ to provide a set of Beta densities which "blanket" $[0,1]$, for example $c_\ell = \ell, d_\ell = r+1-\ell, \ell = 0, 1, \cdots, r$, but treat $\boldsymbol{w}^{(r)}$ as random, given r. As shown by these authors, if $\boldsymbol{w}^{(r)} \sim$ Dirichlet $(\alpha \mathbf{1}_r)$, where $\mathbf{1}_r = (1, \cdots 1)^T$, then $E[T(h(\theta))] \approx T(h_0(\theta))$, that is h is roughly centered about h_0. This class of models is easy for inferential purposes. Since finding the posterior of h is equivalent to finding the posterior of $\boldsymbol{w}^{(r)}$, all one needs is to generate samples from the latter.

Next we turn to Dirichlet processes. Since the appearance of the classic paper of Ferguson (1973), such processes have been used quite extensively for Bayesian nonparametric inference. A probability measure G on $\mathcal{G}$ is said to follow a Dirichlet process with parameter αG_0, symbolically written as $G \sim DP(\alpha G_0)$, if for any measurable partition $B_1, \cdots, B_m$ of $\mathcal{G}$, $(G(B_1), \cdots, G(B_m)) \sim$ Dirichlet $(\alpha G_0(B_1), \cdots, \alpha G_0(B_m))$. Here, G_0 is a specified probability measures, and α is the "precision" parameter. This name for α is justified since $V[G_0(B_\ell)] = G_0(B_\ell)(1-G_0(B_\ell))/(\alpha+1)$ which decreases in α for every ℓ.

Computationally, it is most convenient to work with a family of Dirichlet mixture distributions. Let $\{f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \epsilon \Theta \subset R^P\}$ be a parametric family of densities with respect to some dominating measure μ. Consider the family of probability distributions $\mathcal{F} = \{F_G : G \epsilon \mathcal{G}\}$ with densities

$$f(y|G) = \int f(y|\boldsymbol{\theta}) dG(\boldsymbol{\theta}). \tag{22}$$

Here $G(\boldsymbol{\theta})$ is viewed as the conditional distribution of $\boldsymbol{\theta}$ given G. It is assumed that $G \sim DP(\alpha G_0)$, whence $f(y|G)$ arises by mixing with respect to a distribution having a Dirichlet process.

Mukhopadhyay and Gelfand (1997) provide a general discussion of Bayesian inference based on Dirichlet process mixed models. Mixtures of Dirichlet processes were first introduced in Antoniak (1974), and have been considered subsequently in Lo (1984), Brunner and Lo (1989, 1994), (1991), Escobar (1994), Escobar and West (1995), MacEachern and Müller (1994), Gelfand and Mukhopadhyay (1995), and Newton, Czado and Chappell (1996) among others.

To illustrate the implementation of the semiparametric Bayesian procedure in the context of GLMs, we begin with conditionally independent observations y_i with pdf given in (1). In the next stage, we model $h(\theta_i) = \boldsymbol{x}_i^T \boldsymbol{b} + u_i$, where the u_i are iid from G with $G \sim DP(\alpha G_0)$. Integrating over G, the u_i have joint pdf $f(u_1, \cdots, u_n|G_0, \alpha)$. They are no longer independent, but this joint distribution can be written explicitly.

To complete the prior specification, one needs to specify G_0, and also specify a distribution for $\boldsymbol{b}$ and α. Typically, one assigns a $N(\boldsymbol{b}_0, \boldsymbol{\Sigma})$ distribution for $\boldsymbol{b}$ and, following Escobar and West (1995), a gamma distribution for α. (Empirical experience suggests setting $\alpha = 1$ may be preferable to adding a hyperprior.) A simple choice for G_0 is a normal distribution. However, the Bayesian model fitting can be implemented, at least in principle, for an arbitrary G_0 and α.

In order to implement the model fitting, one begins with the joint posterior of $\boldsymbol{b}$

and $\boldsymbol{u} = (u_1, \cdots, u_n)^T$ as

$$\begin{aligned} & \pi(\boldsymbol{b}, \boldsymbol{u}, \alpha | G_0, \boldsymbol{y}) \\ \propto\ & \exp\left[\sum_{i=1}^n \left\{ y_i h^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{b} + u_i\right) - \psi\left(h^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{b} + u_i\right)\right)\right\}\right] \\ & \times \exp\left[-\tfrac{1}{2}\left(\boldsymbol{b} - \boldsymbol{b}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{b} - \boldsymbol{b}_0\right)\right] f\left(u_1, \cdots, u_n | G_0, \alpha\right) \pi\left(\alpha\right). \end{aligned} \tag{23}$$

The MCMC implementation of the Bayesian procedure requires generating samples from the full conditionals

$$\begin{aligned} \text{(a)} \qquad \pi\left(\boldsymbol{b} | \boldsymbol{u}, \alpha, G_0, \boldsymbol{y}\right) \ & \propto \ \exp\left[\sum_{i=1}^n \left\{ y_i h^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{b} + u_i\right) - \psi\left(h^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{b} + u_i\right)\right)\right\}\right] \\ & \times \exp\left[-\tfrac{1}{2}\left(\boldsymbol{b} - \boldsymbol{b}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{b} - \boldsymbol{b}_0\right)\right]; \end{aligned} \tag{24}$$

(b) $\pi\left(u_i | \boldsymbol{b}, u_j\left(j \neq i\right), \alpha, G_0, \boldsymbol{y}\right)$, a mixed density placing point masses proportional to $f(y_i | \boldsymbol{b}, u_j)$ at each $u_j (j \neq i)$ and continuous mass proportional to $\alpha f(y_i | \boldsymbol{b}, u_i)$ $f(u_i | G_0)$;

(c) $\pi(\alpha | \boldsymbol{b}, \boldsymbol{u}) \propto f(u_1, \cdots, u_n | G_0, \alpha)\pi(\alpha)$. If $\pi(\alpha)$ is gamma, Escobar and West (1995) have discussed how to generate samples from $\Pi(\alpha | \boldsymbol{b}, \boldsymbol{u})$ by introducing an additional parameter, say, γ. The reader is referred to their paper for details.

5. Overdispersed Generalized Linear Models

Being based on the one-parameter exponential family of distributions, GLMs assume a known functional relationship between the mean and the variance. This makes these models unsuitable for certain applications, especially those where the samples are too heterogeneous to be explained by such a simple functional relationship. In such instances, one is naturally led to a wider class of models.

A popular approach for creating a larger class has been through mixture models. For instance, the one parameter exponential family defining the GLM is mixed with a two parameter exponential family for the canonical parameter θ (or equivalently the mean parameter μ) resulting in a two parameter marginal mixture family for the data. The resulting overdispersed family of mixture models no longer belongs to the exponential family (e.g. beta-binomial, gamma-Poisson). More importantly, since the likelihood depends on the sample size, while the mixture distribution does not, the relative overdispersion of the resulting mixture family to the original exponential family tends to infinity as the sample size increases. An implication of such models is that taking additional observations within a population does not increase knowledge regarding heterogeneity across populations.

A second class of models, usually referred to as exponential dispersion models (EDM), arises when $a(\phi) = \phi$ so that ϕ behaves like a scale parameter. Jorgensen (1987) provides an extensive treatment of such models. The resulting two parameter family of distributions no longer belongs to the exponential family.

An alternative approach due to Efron (1986) models overdispersion through so-called "double-exponential" families. Such families are derived as a saddle point approximation to the density of an average of n^* random variables from a one parameter exponential family with large n^*. The parameter n^* written as $n\rho$ for actual sample size n introduces ρ as a second parameter in the model along with canonical parameter θ. Ganio and Schafer (1992) have shown that EDM's can be

embedded within Efron's double exponential family, and the associated asymptotic inference applies. These asymptotics result in overdispersion relative to the original exponential family which tends to a constant as $n \to \infty$, unlike the mixture case.

A more general class of models was introduced by Gelfand and Dalal (1990). For a given one-parameter exponential family, they introduced a two-parameter exponential family where one parameter is the overdispersion parameter in addition to the canonical parameter. This model includes Efron's model as a special case, and also includes a family discussed in Lindsay (1986).

Dey, Gelfand and Peng (1997) adopted a Bayesian approach for fitting these models using Jeffreys' prior. Following the technique of Laud and Ibrahim (1991), they also proved the propriety of posteriors under such priors under certain conditions. Some of their results are presented below.

We begin with conditionally independent random variables y_i $(i = 1, \cdots, n)$ such that

$$f(y_i|\theta_i, \tau_i) = b(y_i) \exp[\theta_i y_i + \tau_i T(y_i) - \rho(\theta_i, \tau_i)]. \tag{25}$$

In the above f is a density with respect to some σ-finite measure μ. Assuming that (25) is integrable with respect to y_i, if $T(y_i)$ is convex, then for distributions with common means, $V(y_i)$ increases in τ_i. Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$, and define $\theta_i = h_1(x_i^T \boldsymbol{b})$ and $\tau_i = h_2(\boldsymbol{z}_i^T \boldsymbol{\alpha})$, where h_1 and h_2 are strictly increasing. The resulting likelihood is

$$L(\boldsymbol{b}, \boldsymbol{\alpha}; \boldsymbol{y}) = \exp[\sum_{i=1}^{n} \{\theta_i y_i + \tau_i T(y_i) - \rho(\theta_i, \tau_i)\}]. \tag{26}$$

We shall use the notation $\rho^{(r,s)} = \partial \rho^{r+s} / (\partial \theta^r \partial \tau^s)$. Then straightforward calculations yield

$$E\left(-\frac{\partial^2 \log L}{\partial b_j \partial b_k}\right) = \sum_{i=1}^{n} \rho^{(2,0)}(\theta_i, \tau_i) x_{ij} x_{ik} [g'(\boldsymbol{x}_i^T \boldsymbol{b})]^2; \tag{27}$$

$$E\left(-\frac{\partial^2 \log L}{\partial \alpha_j \partial \alpha_k}\right) = \sum_{i=1}^{n} \rho^{(0,2)}(\theta_i, \tau_i) z_{ij} z_{ik} [h'(\boldsymbol{z}_i^T \boldsymbol{\alpha})]^2; \tag{28}$$

$$E\left(-\frac{\partial^2 \log L}{\partial b_j \partial \alpha_k}\right) = \sum_{i=1}^{n} \rho^{(1,1)}(\theta_i, \tau_i) x_{ij} z_{ik} g'(\boldsymbol{x}_i^T \boldsymbol{b}) h'(\boldsymbol{z}_i^T \boldsymbol{\alpha}). \tag{29}$$

Writing $\boldsymbol{X}^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, $\boldsymbol{Z}^T = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$, $\boldsymbol{M}_\theta$ as the $n \times n$ diagonal matrix with $(\boldsymbol{M}_\theta)_{ii} = \rho^{(2,0)}(\theta_i, \tau_i)(g'(\boldsymbol{x}_i^T \boldsymbol{b}))^2$, $\boldsymbol{M}_\tau$ a diagonal matrix with $(\boldsymbol{M}_\tau)_{ii} = \rho^{(0,2)}(\theta_i, \tau_i)(h'(\boldsymbol{z}_i^T \boldsymbol{\alpha}))^2$, and $\boldsymbol{M}_{\theta,\tau}$ a diagonal matrix with $(\boldsymbol{M}_{\theta,\tau})_{ii} = \rho^{(1,1)}(\theta_i, \tau_i) g'(\boldsymbol{x}_i^T \boldsymbol{b}) h'(\boldsymbol{z}_i^T \boldsymbol{\alpha})$, one gets the Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{b}, \boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{X}^T \boldsymbol{M}_\theta \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{M}_{\theta,\tau} \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{M}_{\theta,\tau} \boldsymbol{X} & \boldsymbol{Z}^T \boldsymbol{M}_\tau \boldsymbol{Z} \end{pmatrix}. \tag{30}$$

Jeffreys' prior is then given by $|\boldsymbol{I}(\boldsymbol{b}, \boldsymbol{\alpha})|^{1/2}$.

Suppose we assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of full column rank, and that $L(\boldsymbol{b}, \boldsymbol{\alpha}; \boldsymbol{y})$ is bounded above. Then, the posterior of $(\boldsymbol{b}, \boldsymbol{\alpha})$ is proper if, for each $y_i (i = 1, \cdots, n)$,

$$\int_T \int_\Theta \exp[\theta y_i + \tau T(y_i) - \rho(\theta, \tau)] (\rho^{(2,0)}(\theta, \tau) \rho^{(0,2)}(\theta, \tau)^{1/2} d\theta d\tau < \infty. \tag{31}$$

This result is proved in Dey et al. (1997).

Recognizing the limitations of the one parameter exponential family, for example, an implicit mean-variance relationship and unimodality, Mukhopadhyay and

Gelfand (1997) introduced a Dirichlet process mixed GLM (DPMGLM). These models provide a more flexible first stage specification, while retaining linear structure on a transformed scale.

6. Model Determination Approaches

With the availability of a wide range of GLMs to consider in analyzing a dataset, the problem of model determination becomes critical. Model determination comprises model checking - is the model adequate? - and model selection - among a set of adequate models, which one is best?

First we consider model adequacy which has received much less attention in the literature than model choice. In providing the probabilistic components of a hierarchical model, we rarely believe that any of the distributions is correct. Those specifications further removed from the data are often intentionally made less precise, not because we *believe* them to be correct but in order to permit the data to drive the inference. However, what is *true* is apart from model checking. If we undertake model criticism we must examine the adequacy of what is specified and we must assume proper priors (or else the observed data could not have arisen under the model). High dimensional models, e.g., those having more parameters than data points, as well as very vaguely specified hierarchical models will be difficult to criticize.

A formal Bayesian model adequacy criterion (as in Box, 1980) proposes that the marginal density of the data be evaluated at the observations. Large values support the model, small values do not. Assessment of the magnitude of this value could be facilitated by standardizing, using the maximum value or an average value of this density (Berger, 1985). However, a high dimensional density ordinate will be difficult to estimate well and hopeless to calibrate. In addition, with hierarchical models, failures, such as outliers, mean structure errors, dispersion misspecifications and inappropriate exchangeabilities, can occur at each hierarchical stage. The formal procedure does not provide feedback regarding the adequacy of the stagewise specifications.

Chaloner and Brant (1988), Chaloner (1994) and Weiss (1995), focusing on outlier detection suggest posterior-prior comparison. Their strategy is to identify random variables whose distribution, a priori, is a standard one. In particular, they choose functions of so-called realized residuals. Given the data, the posterior distribution of each such function is obtained. If it differs considerably from its associated prior, using tail area comparison, a lack of model fit is claimed. For a realized residual itself, an outlying observation is asserted. If the entire model specification is correct, such comparisons will be successful on average but will fail to recognize the *variability* in the posterior.

A second approach, referred to as model expansion or elaboration, captures model failures by specifying a more complex model using mixtures. Though most often used to detect outliers, recently, Albert and Chib (1997) use this approach for other model failures, in particular, exchangeability in the direction of partial exchangeability. Regardless, the model of interest becomes nested within the expanded or full model, so model choice procedures replace model checking to criticize the adequacy of the reduced one. Recent work of Müller and Parmigiani (1995) and Carota, Parmigiani and Polson (1993) combines elaboration with posterior-prior comparison using the Kullback-Leibler distance between these two distributions for the elab-

oration parameter. This approach requires, for each sort of failure, a non-unique specification of an expanded model.

A third approach is taken up in Gelman, Meng and Stern (1995) who propose a posterior predictive strategy. These authors define a discrepancy measure as a function of data and parameters, treating both as unknown in one case, inserting the observed data in the other. They then compare the resulting posterior distributions given the observed data. Gelman, et al. dismiss prior predictive checking arguing that the prior predictive distribution treats the prior as a true "population distribution" whereas the posterior predictive distribution treats the prior as an outmoded first guess. However, model checking must examine the acceptability of the model fitted to the data. Model parameters must be generated from the prior prescribed under the model. Gelman, et al. can be criticized for using the data twice. The observed data, through the posterior, suggests values of the parameter which are likely under the model. Then, to assess adequacy, the observed data is checked against data generated using such parameter values, apparently making it difficult to criticize the model.

A fourth approach is developed in recent work of Hodges (1998). Limited to the case where all levels are Gaussian, he reexpresses linear hierarchical models as standard linear models with simple covariance structure. He then suggests the use of familiar linear models diagnostic tools, e.g., residual plots, added variable plots, transformations, collinearity checks, case influence, etc. Ad-hoc method is needed in tailoring some of these tools to the hierarchical structure.

Finally, Dey, Gelfand, Swartz and Vlachos (1998) suggest an approach which is entirely simulation based, requiring only the model specification and that, for a given data set, one be able to simulate draws from the posterior under the model. By replicating a posterior of interest using data replicates obtained under the model, the extent of variability in such a posterior can be seen. Then, the posterior obtained under the observed data can be compared with this medley of posterior replicates to ascertain whether the former is in agreement with them and accordingly, whether it is plausible that the observed data come from the proposed model. Such comparison can be implemented using a Monte Carlo test. Many such tests can be run, each focusing on a potential model failure.

Turning to model choice, for a collection of models $m = 1, 2, \cdots, M$, the formal Bayesian approach assumes that one is "true" but which is the true one is unknown. Assigning prior probabilities p_m that model m is true, the posterior probability of model m, is $Pr(m|\boldsymbol{y}) \propto f(\boldsymbol{y}|m)p_m$ where $f(\boldsymbol{y} \mid m)$ is the marginal or prior predictive density of $\boldsymbol{y}$ under model m. Hence, if $\boldsymbol{y}_{obs}$ denotes the realized data, the model which maximizes $f(\boldsymbol{y}_{obs}|m)p_m$ is selected. If $p_m = M^{-1}$ for all m, we choose the model with the largest $f(\boldsymbol{y}_{obs}|m)$, suggesting the use of this quantity as a general screening criterion. When models are compared in pairs, the Bayes factor emerges, $B = f(\boldsymbol{y}_{obs}|m_1)/f(\boldsymbol{y}_{obs}|m_2)$ for say models m_1 and m_2. B is viewed as a weight of evidence; $B > 1$ supports model m_1, $B < 1$ supports model m_2.

Bayes factors have a wide advocacy in the Bayesian community; see Kass and Raftery (1995) for a review. However, they lack interpretation in the case of improper priors which are frequently used in complex hierarchical specifications and they are difficult to compute for such models with large datasets (though there is much recent discussion, see, e.g., Raftery, 1995). The use of Schwarz's (1978) Bayesian information criterion (BIC) as an approximation to the Bayes factor requires the specification of model dimension. Unfortunately in the context of GLMs involving mixed effects, the dimension of the model is unclear. Moreover, the asymptotics associated with such approximation are invalid when the number of model

parameters grows with sample size, as in random effects settings where the number of individuals grows large.

Recently, attractive alternatives have appeared. Gelfand and Ghosh (1998), noting that posterior prediction is often a primary use for a model, suggest a formal utility maximization approach for model selection. In particular, their approach amounts to obtaining a minimized expected posterior predictive loss for a given model and then selecting the model which provides the overall minimum.

For a version of log scoring (or deviance) loss, the minimization for a given model can be done explicitly yielding an expression which can be interpreted as a penalized deviance criterion. The criterion is comprised of a piece which is a Bayesian deviance measure and a piece which is interpreted as a penalty for model complexity. The penalty function arises without specifying model dimension or asymptotic justification.

Under the model in (1) with $a(\phi_i) = \phi/w_i, w_i$ known, the criterion becomes, for model m,

$$\begin{aligned} D_k(m) &= 2\sum_{i=1}^{n} w_i(t_i^{(m)} - t(\mu_i^{(m)}))+ \\ &\quad +2(k+1)\sum_{i=1}^{n} w_i \left\{ \frac{t(\mu_i^{(m)})+kt(y_{i,obs})}{k+1} - t\left(\frac{\mu_i^{(m)}+ky_{i,obs}}{k+1}\right)\right\} \end{aligned} \tag{32}$$

In (32), $t(y) = y\theta(y) - \psi(\theta(y))$ where $\theta(\mu) = \psi'^{-1}(\mu), \mu_i^{(m)} = E(y_i|\boldsymbol{y}_{obs}, m)$ and $t_i^{(m)} =$
$E(t(y_i)|\boldsymbol{y}_{obs}, m)$, and k is a weight which typically does not affect the ordering of the models and so may be set to 1 for convenience. Since $t(y)$ is convex, Jensen's inequality ensures that each term in the right side of (32) is nonnegative. Gelfand and Ghosh clarify that the first term can be interpreted as a penalty function and the second as a goodness-of-fit term. The choice (1) determines $t(y)$, hence (32). For instance, in the Possion case $t(y) = y\log y - y$, in the binomal case, $t(y) = \log\left(\frac{y}{n}\right) + \frac{n-y}{n}\log\left(\frac{n-y}{n}\right)$. Usual continuity corrections are imposed to ensure that $t(y)$ can be calculated for any $y_{i,obs}$ and that $t_i^{(m)}$ exists.

Another, somewhat similar criterion has been discussed in Spiegelhalter, Best and Carlin (1998). Motivated by the work of Dempster (1974), their suggestion is to obtain the posterior distribution of the log likelihood at the observed data for each model and then compare these across models. In particular, for (1) they define the "Bayesian deviance" $D(\boldsymbol{\theta})$ to be $\sum_{i=1}^{n} D(\theta_i)$ where

$$D(\theta_i) = -2\log f(y_i|\theta_i) + 2\log f(y_i|\theta(y_i)). \tag{33}$$

Defining $\overline{D} = E(D(\boldsymbol{\theta})|y)$ and $p_D = \overline{D} - D(E(\boldsymbol{\theta}|\boldsymbol{y}_{obs}))$, Spiegelhalter, et al. propose the criterion

$$DIC = \overline{D} + p_D \tag{34}$$

where DIC denotes Deviance Information Criteria. They argue that $\overline{D}$, the posterior expected deviance, summarizes model fit while p_D, interpreted as the effective number of parameters, measures the complexity of the model. They show that DIC generalizes the familiar Akaike Information Criteria (AIC) (Akaike, 1973).

Both the Gelfand and Ghosh criteria and the DIC are readily computed from posterior samples. The BUGS software (Spiegelhalter, et al. 1996) provides a convenient package for fitting most Bayesian GLMs, and thus for providing posterior samples.

Finally, informal Bayesian model selection in the case of nested GLMs can be effected by obtaining the posterior distribution of the discrepancy parameter between

the full and reduced models as in Albert and Chib (1997). Exploratory approaches using cross validation ideas, applicable to small or even moderate sized datasets are discussed in Gelfand, Dey and Chang (1992) and Gelfand (1995).

Acknowledgement

The work of the first author was supported in part by NSF grant DMS 96-25383. The work of the second author was supported in part by NSF grant SBR 98-10968.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of International Symposium on Information Theory*, ed. B.N.Petrov and F. Czaki, Budapest, Academia Kiado, 267-281.

Albert, J.H. (1985). Simultaneous estimation of Poisson means under exchangeable and independence models. *Journal of Statistical Computation and Simulation*, **23**, 1-14.

Albert, J.H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, **83**, 1037-1044.

Albert, J.H. and S. Chib (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, **92**, 916-925.

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, **2**, 1152-1174.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Box, G.E.P. (1980). Sampling and Bayes's inference in scientific modeling (with discussion). *Journal of the Royal Statistical Society, Series A*, **143**, 383-430.

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

Brunner, L.J. and Lo, A.Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics*, **17**, 1550-1566.

Brunner, L.J. and Lo, A.Y. (1994). Nonparametric Bayes methods for directional data. *The Canadian Journal of Statistics*, **22**, 401-412.

Carota, C., G. Parmigiani and N.G. Polson (1997). Diagnostic measures for model criticism. *Journal of the American Statistical Association*, **92**, 753-762.

Casella, G and George, E.I. (1994). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

Chaloner, K. (1994). Residual analysis and outliers in Bayesian hierarchical models. In: *Aspects of Uncertainty*, eds. A.F.M. Smith and P.R. Freeman. Chichester, U.K., John Wiley, 153-161.

Chaloner, K. and R. Brant (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651-659.

Crowder, M. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics*, **27**, 34-37.

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, series B*, **41**, 1-31.

Dellaportas, P. and Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443-459.

Dempster, A. (1974). The direct use of likelihood for significance testing. In: *Proceedings of Conference on Foundational Questions In Statistical Inference.* Eds: O. Barndorff-Nielsen, P. Blaesild and G. Schou, p. 335-352, Department of Theoretical Statistics: University of Aarhus.

Dey, D.K., Gelfand, A.E. and Peng, F. (1997). Overdispersed generalized linear models. *Journal of Statistical Planning and Inference,* **64**, 93-107.

Dey, D.K., Gelfand, A.E., Swartz, T. and Vlachos, P.K. (1998). Simulation based model checking for hierarchical models. *Test*, **7**, 325-346.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 709-721.

Escobar, M.D. (1994). Estimating normal means with Dirichlet process priors. *Journal of the Statistical Planning and Inference*, **43**, 97-106.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577-588.

Fahrmeir, L and Tutz, G. (1994). *Multivariate Statistical Modelling based on Generalized Linear Models.* Springer-Verlag, New York.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209-230.

Ganio, L.M. and Schafer, D.W. (1992). Diagnostics of overdispersion. *Journal of the American Statistical Association*, **87**, 795-804.

Gelfand, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, eds. W. Gilks et al., London, Chapman and Hall, 145-161.

Gelfand, A.E. (1998). Approaches for semiparametric Bayesian regression. In: *Asymptotics, Nonparametrics and Time Series.* Ed.: S. Ghosh, Marcel Dekker, Inc, New York. (to appear).

Gelfand, A.E. and Dalal S. (1990). A note on overdispersed exponential families. *Biometrika*, **77**, 55-64.

Gelfand, A.E., D.K. Dey and H. Chang (1992). Model determination using predictive distributions with implementations via sampling-based methods. In: *Bayesian Statistics 4*, eds. J.M. Bernardo et al., Oxford, U.K., Oxford University Press, 147-167.

Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1-11.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972-985.

Gelfand, A.E. and Sahu, S.K. (1999). On the propriety of posteriors and Bayesian identifiability in generalized linear models. *Journal of the American Statistical Association* (to appear).

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., X-L. Meng and H.S. Stern (1995). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733-807.

Ghosh, M., and Natarajan, K. (1998). Small area estimation : a Bayesian perspective. *Multivariate, Design and Sampling*, Ed., S. Ghosh, Marcel Dekker, New York (to appear).

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, **93**, 273-282.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337-348.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors in Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**, 1461-1473.

Hodges, J. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society, Series B*, **60** (to appear),

Ibrahim, J.G. and Laud, P.W. (1991). On Bayesian analysis of general linear models using Jeffreys's prior. *Journal of the American Statistical Association*, **86**, 981-986.

Jorgensen, B. (1987). Exponential dispersion models. (with discussion). *Journal of the Royal Statistical Society, Series B*,**50**, 150-?.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Leonard, T. and Novick, M.R. (1986). Bayesian full rank marginalization for two-way contingency tables. *Journal of Educational Statistics*, **11**, 33-56.

Lindsay, B. (1986). Exponential family mixture models. *The Annals of Statistics*, 14, 124-137.

Lindsey, J.K. (1995). *Modelling Frequency and Count Data.* London, Clarendon Press.

Lo, A.Y. (1984). On a class of Bayesian nonparametric density estimation:I, density estimates. *The Annals of Statistics*, 12, 351-357.

MacEachern, S.N. and Müller, P. (1994). Estimating mixture of Dirichlet process models. Technical Report, Institute of Statistics and Decision Sciences, Duke University.

Mallick, B.K. and Gelfand, A.E. (1994). Generalized linear models with unknown link functions. *Biometrika*, 81, 237-245.

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

Meng, X-L. (1994). Posterior predictive p-values. *Annals of Statistics*, **22**, 1142-1160.

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, **92**, 633-639.

Müller, P. and G. Parmigiani (1995). Numerical evaluation of information theoretic measures. Inc: *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner.* Eds: Berry, D.A., Chaloner, K.M., Geweke, J.F., John Wiley, New York, 397-406.

Natarajan, R. and McCulloch, C.E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, 82, 639-643.

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

Newton, M.A., Czado, C. and Chappell, R. (1996). Semiparametric Bayesian inference for binary regression. *Journal of the American Statistical Association*, **91**, 142-153.

Piergorsch, W.W., Weinberg, C.R. and Margolin, B.H. (1988). Exploring simple independent action in multifactor table of proportions. *Biometrics*, **44**, 595-603.

Raftery, A.E. (1995). Hypothesis testing and model selection. In: *Markov Chain Monte Carlo in Practice.* Eds. W. Gilks, et al. London, Chapman and Hall, 163-187.

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

Spiegelhalter, D.J., Best, N.G. and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. Preprint.

Weiss, R.E. (1995). Residuals and outliers in repeated measures random effects models. Tech. Rpt., Dept. of Biostatistics, UCLA.

West, M. (1985). Generalized linear models: scale parameters, outlier accomodation and prior distributions. In *Bayesian Statistics*,**2**, Oxford, Oxford University Press, 531-557.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79-86.

2

Random Effects in Generalized Linear Mixed Models (GLMMs)

Dongchu Sun
Paul L. Speckman
Robert K. Tsutakawa

ABSTRACT In this chapter, we examine the use of special forms of correlated random effects in the generalized linear mixed model (GLMM) setting. A special feature of our GLMM is the inclusion of random residual effects to account for lack of fit due to extra variation, outliers and other unexplained sources of variation. For random effects, we consider, in particular, the correlation structure and improper priors associated with the autoregressive (AR) model of Ord (1975) and the conditional autoregressive (CAR) model of Besag (1974). We give conditions for the propriety of the posterior distribution of the GLMM when the fixed effects have a constant improper prior and the random effects have a possibly improper conditional autoregressive prior. Several examples of exponential families as well as computational details for Markov chain Monte Carlo simulation are also presented.

1. Introduction

Traditional treatment of random effects in mixed linear and nonlinear models generally assumes that these effects are independent following some standard distributions such as normal or gamma. However, with the advent of Markov chain Monte Carlo (MCMC) methods and, in particular, the Gibbs sampler (cf. Gelfand and Smith, 1990), such restrictions are no longer necessary, and a much broader class of models, including those with correlated random effects, can be used in practice. (See Clayton (1996) for a general review of this recent development.)

In this chapter we consider generalized mixed linear models with random effects having the autoregressive and conditionally autoregressive properties commonly encountered in temporal and spatial covariates where one expects similarities among closely situated observations. Examples from disease mapping will be used to motivate these models.

The computational simplicity of MCMC methods enables one to extend the commonly used generalized linear mixed model (GLMM) to one that appends random residual effects to the linear term to account for lack of fit. These extra terms allow for the minor perturbations and occasional outliers commonly encountered in practice. However, the remarkable ease of application of the Gibbs sampler does not come without a price. There is potential nonconvergence and other annoying

problems when using the algorithm, especially in situations where noninformative prior distributions are employed.

In Section 2 we formally define the GLMM with residual effects. Two examples are given. One has the normal distribution and the other the gamma distribution, with the choice depending on the nature of the observed data. For example when the data are Poisson, it is more natural to use the conjugate gamma distribution, although the normal may be just as appropriate and simple to use.

In Section 3 we discuss several forms of correlated random effects including the AR process of Ord (1974) and the CAR process of Besag (1974), which are useful in describing spatial correlations. We examine the joint distributions associated with these processes to get a better understanding of the underlying association implied by these models. Of particular interest are distributions that are improper and could create problems when used in the GLMMs.

In Section 4, we consider the incorporation of these spatial random variables into the GLMM setting and emphasize the special role of the link function in a Bayesian hierarchical framework. In the case where the residual effects are normally distributed, the fixed effects have a constant prior and random effects may have an improper prior, we give sufficient conditions for the existence of a proper posterior distribution of all parameters including the fixed and random effects and variance components.

In Section 5, we summarize the computational details including the full conditional distributions required for the implementation of the Gibbs Sampler.

2. The Model

Let $Y_1, \ldots, Y_N$ be the independent random observations, where Y_i has the probability density

$$f_i(y_i|\eta_i, \phi) = \exp[A_i(\phi)^{-1}\{y_i\eta_i - B_i(\eta_i)\} + C_i(y_i; \phi)]. \tag{1}$$

The function $A_i(\phi)$ is commonly of the form $A_i(\phi) = \phi w_i^{-1}$, where the w_i are prespecified weights. It is often assumed that the scale parameter ϕ is known. Consider, for example, the case when the population size in area i is m_i with unknown mortality rate p_i, and Y_i is Poisson distributed with mean $m_i p_i$. This is a special case of (1) with $\phi = 1, A_i(\phi) = 1, \eta_i = \log(m_i p_i), B_i(\eta_i) = \exp(\eta_i)$, and $C_i(y_i; \phi) = -\log(y_i!)$. When Y_i has a binomial distribution with parameters m_i and p_i, $\phi = 1, A_i(\phi) = 1, \eta_i = \log\{p_i/(1-p_i)\}, B_i(\eta_i) = m_i \log\{1 + \exp(\eta_i)\}$, and $C_i(y_i; \phi) = \log[m_i!/\{y_i!(m_i - y_i)!\}]$.

Generalized Linear Models. We wish to model the variability in η_i to account for various fixed covariates. The natural parameters η_i are modeled as

$$h_i(\eta_i) = \boldsymbol{x}_{1i}^t \boldsymbol{\theta}, \tag{2}$$

where the h_i are known monotone functions, $\boldsymbol{X}_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n})^t$ is an $N \times p$ design matrix and $\boldsymbol{\theta}$ is the vector of fixed effects. Such a model is commonly referred to as a generalized linear model (GLM) with canonical parameter η_i, scale parameter ϕ, and link function h_i (cf. McCullagh and Nelder, 1989). (Note that usually there is a single link function $h_i \equiv h$.)

Generalized Linear Mixed Models. We now extend the model to include random effects as follows. Let

$$h_i(\eta_i) = \boldsymbol{x}_{1i}^t \boldsymbol{\theta} + \boldsymbol{x}_{2i}^t \boldsymbol{Z}, \tag{3}$$

where h_i is a known monotone function, $\boldsymbol{X}_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n})^t$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n})^t$ are $N \times p$ and $N \times k$ design matrices, the $p \times 1$ vector $\boldsymbol{\theta}$ represents fixed effects, and $\boldsymbol{Z}$ is a $k \times 1$ vector of random effects. Models given by (1) and (3) are often called generalized linear mixed models (GLMMs) and have been widely used in many problems such as disease mapping e.g., Breslow and Clayton (1993).

We can further extend the model to add additional residual effects by taking

$$h_i(\eta_i) = \boldsymbol{x}_{1i}^t \boldsymbol{\theta} + \boldsymbol{x}_{2i}^t \boldsymbol{Z} + e_i. \tag{4}$$

Here $\boldsymbol{e} = (e_1, \ldots, e_N)^t$ are residual effects satisfying some restriction such as $\mathbb{E}(e_i) = 0$ or $\mathbb{E}\exp(e_i) = 1$. In addition, $\boldsymbol{Z}$ and $\boldsymbol{e}$ are assumed mutually independent. We include random residual effects e_i to account for the lack of fit of (3) due to extra variation, outliers, and other unexplained sources of variation. Note that the random effect e_i is quite different from $\boldsymbol{Z}$ in the sense that $\boldsymbol{Z}$ often accounts for some special pattern such as random geographical effects and spatial correlation. In addition, the number of components of $\boldsymbol{Z}$ is often much smaller than N, the number of residual effects e_i. By a suitable choice of the design matrix, (4) may be encompassed under (3), but we do not do this in order to emphasize the separate roles of $\boldsymbol{Z}$ and $\boldsymbol{e}$. We will call the model given by (1) and (4) a GLMM as well.

There are many possible choices for the link functions h_i in models (2)–(4). For example, in the mortality setting cited earlier, Y_i has the Poisson distribution with mean $m_i p_i$ and $\eta_i = \log(m_i p_i)$. One possibility is to take $h_i(\eta_i) = \eta_i - \log(m_i) = \log(p_i)$, and a loglinear regression model may be applied. Alternatively, Y_i can be modeled with a binomial distribution. Then the logit link is canonical, and $\text{logit}(p_i) = \log\{p_i/(1-p_i)\} = \eta_i - \log(m_i - e^{\eta_i}) = h_i(\eta_i)$, resulting in logistic regression.

The random effects term $\boldsymbol{Z}$ in (3)–(4) is typically assumed to have a multivariate normal distribution. We will discuss in detail the choice of the distribution of $\boldsymbol{Z}$ in the next section.

Distribution of Residual Effects. We will assume that the residual effects e_i or some monotone functions of e_i have distributions belonging to an exponential family (1), with known common canonical parameter η but unknown scale parameter ϕ. For illustration, we will consider the following two classes of distributions for residual effects.

- *Normal Residual Effects.* Residual effects e_i are independent and identically normal with mean 0 and variance δ_0.
- *Gamma Residual Effects.* The $\exp(e_i)$ are iid gamma(R, R). Here a random variable W has the gamma(α, β) distribution if W has p.d.f.

$$f(w) = \alpha^\beta \{\Gamma(\alpha)\}^{-1} w^{\alpha-1} \exp(-\beta w).$$

Special cases of these models have appeared previously. Clayton and Kaldor (1987) and Waller *et al.* (1997) use a Poisson-normal model (Poisson for Y_i and normal for $\boldsymbol{Z}$) but without the residual term $\boldsymbol{e}$. This a special case of (3). Ghosh *et al.* (1998) use $\boldsymbol{e}$ in the binomial-normal model and treat spatial effects by taking $\boldsymbol{X}_2\boldsymbol{Z} = \boldsymbol{U}$, with $\boldsymbol{U}$ having a distribution defined by the conditional auto regressive CAR(1) model of Besag (1974). This is a special case of (4). In Sun, Tsutakawa, Kim and He (1997) and Sun, Tsutakawa and He (1998), $\boldsymbol{Z}$ consists of block-wise independent random effects, where each block contains random effects and the e_i are independent random variables with mean 0 and a common variance. West and

Aguilar (1997) give another interesting example analysing hospital quality monitors with an extra residual term in (1.4).

Special cases of Poisson-gamma models are found in Clayton and Kaldor (1987) and Tsutakawa (1988). Specifically, in Tsutakawa (1988), $\boldsymbol{Z}$ contains independent random effects, and the $\exp(e_i)$ are independent gamma variables with mean 1 and a common variance.

The general model (1) and (2) can be used for both continuous and discrete data. A discrete example of (4), which motivated much of this work, is studied in Sun, Tsutakawa, Kim and He (1998), where a spatio-temporal model for cancer mortality data is proposed. For a given gender, let Y_{ijk} denote the frequency of deaths from some specific cause in the ith region and jth age group during the kth time period, $i = 1, \ldots, I$; $j = 1, \ldots, J$; $k = 1, \ldots, K$. Conditionally on the fixed and random parameters, assume the Y_{ijk} are independent and Poisson with means $m_{ijk}p_{ijk}$, where m_{ijk} is the size of the ijkth target population. The model of Sun *et al.* takes the form

$$\log(p_{ijk}) = \theta_j + Z_i + (\mu_j + W_{ij})(t_k - \bar{t}) + e_{ijk},$$

where θ_j is the effect of the jth age group, Z_i is the effect of the ith region, t_k is the midpoint of the kth time period, and $\bar{t} = \sum_{k=1}^{K} t_k/K$. The rate of change over time is represented by $(\mu_j + W_{ij})$ for the jth age group in the ith region. Both θ_j and μ_j are treated as fixed effects, while Z_i and W_{ij} are random. The residual effects e_{ijk} are also random. A detailed description of the distributions of the random effects and prior distributions are given in Sun *et al.* (1998), where disease mapping and interpretation of numerical results for male lung cancer in the state of Missouri can be found.

3. Random Effects

3.1 Independent Random Effects

Historically, it was common to assume independent random effects for linear mixed models, i.e., $Z_1, \ldots, Z_N$ are independently and identically $\mathrm{N}(0, \delta_1)$ distributed. (See Harville (1977).) Typical examples include one-way ANOVA and two-way ANOVA models with random effects. Hobert and Casella (1996) gave necessary and sufficient conditions for the propriety of the posterior distribution for a class of noninformative priors for variances components assuming independence of random effects.

3.2 Correlated Random Effects

There are many important situations where the random effects should be modeled as correlated. Correlated models are especially appropriate for spatial effects. A number of related methods are commonly used.

Direct specification of correlation matrix. If the random effects are linearly ordered, as for example with longitudinal data, it may be convenient to specify a correlation structure directly. For example, to model correlation decreasing with distance, $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^t$ can be taken to have the $\mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma} = (\sigma_{ij})$ is the $k \times k$ matrix with elements

$$\sigma_{ij} = \tau\rho^{|i-j|}, \tag{5}$$