Preface
Random Effect and Latent Variable Model Selection

In recent years, there has been a dramatic increase in the collection of multivariate and correlated data in a wide variety of fields. For example, it is now standard practice to routinely collect many response variables on each individual in a study. The different variables may correspond to repeated measurements over time, to a battery of surrogates for one or more latent traits, or to multiple types of outcomes having an unknown dependence structure. Hierarchical models that incorporate subject-specific parameters are one of the most widely-used tools for analyzing multivariate and correlated data. Such subject-specific parameters are commonly referred to as random effects, latent variables or frailties.

There are two modeling frameworks that have been particularly widely used as hierarchical generalizations of linear regression models. The first is the linear mixed effects model (Laird and Ware , 1982) and the second is the structural equation model (Bollen , 1989). Linear mixed effects (LME) models extend linear regression to incorporate two components, with the first corresponding to fixed effects describing the impact of predictors on the mean and the second to random effects characterizing the impact on the covariance. LMEs have also been increasingly used for function estimation. In implementing LME analyses, model selection problems are unavoidable. For example, there may be interest in comparing models with and without a predictor in the fixed and/or random effects component. In addition, there is typically uncertainty in the subset of predictors to be included in the model, with the number of candidate predictors large in many applications.

To address problems of this type, it is not appropriate to rely on classical methods developed for model selection and inferences in non-hierarchical regression models. For example, the widely used BIC criteria are not valid for random effects models, and likelihood ratio and score tests face difficulties, since the null hypothesis often falls on the boundary of the parameter space. The objective of the first part of this book is to provide an overview of a variety of promising strategies for addressing model selection problems in LMEs and related modeling frameworks.

In the chapter, "Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models," Ciprian Crainiceanu provides an applications-motivated overview of recent work on likelihood ratio and restricted likelihood ratio tests for

testing whether random effects have zero variance. The approaches he describes represent an important advance over the current standard practice in testing for zero variance components in hierarchical models. Such approaches include ignoring the boundary problem and assuming the likelihood ratio test statistic has a chi-square distribution under the null and relying on asymptotic results showing a mixture of chi-squares is more appropriate (Stram and Lee, 1994). Crainiceanu shows that asymptotic approximations may be unreliable in many applications, motivating use of finite sample approaches. He illustrates the ideas through several examples, including applications to nonlinear regression modeling.

Score tests provide a widely-used alternative to likelihood ratio tests, and in the chapter, "Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics," of this volume Daowen Zhang and Xihong Lin provide an excellent overview of the recent literature on score test-based approaches. In addition, Zhang and Lin consider a broader class of models, which includes GLMMs and generalized additive mixed models (GAMMs). GAMMs provide an extremely rich framework for semiparametric modeling of longitudinal data allowing flexible predictor effects through replacing linear terms in a generalized linear model with unknown non-linear functions, while also including random effects to account for within-subject dependence and heterogeneity.

The first part of the volume is completed with two companion chapters describing Bayesian approaches for variable selection in LMEs and GLMMs. The likelihood ratio and score test methods provide an approach for comparing two nested models with the smaller model having a random effect excluded. However, in many applications one is faced with a set of p candidate predictors, with uncertainty in which subsets should be included in the fixed and random effects components of the model. Clearly, the number of candidate models grows extremely rapidly with p, so that it often becomes impossible to fit each model in the list. One possibility is to use a likelihood ratio test within a stepwise selection procedure. However, the final model selected will depend on the order in which candidate predictors are added or deleted and it is difficult to adjust for uncertainty in subset selection in performing inferences and predictions. In non-hierarchical regression models, Bayesian variable selection implemented with stochastic search algorithms has been very widely used to address this problem. In the chapter, "Bayesian Model Uncertainty in Mixed Effects Models," Satkartar Kinney and I describe an approach for LMEs, while in the chapter, "Bayesian Variable Selection in Generalized Linear Mixed Models," Bo Cai and I describe an alternative for GLMMs.

The second part of the book switches gears to focus on structural equation models (SEMs), which have been very widely used in social science applications for assessing relationships among latent variables, such as poverty or violence, that can only be measured indirectly through multiple surrogates. SEMs provide a generalization of factor analysis, which allows for modeling of linear relationships among the latent factors through a linear structural relations (LISREL) model. SEMs are also quite useful outside of traditional application areas for sparse covariance structure modeling of high-dimensional multivariate data. However, one of the main issues in applying SEMs is how to deal with model uncertainty, which commonly arises

in deciding on the number of factors to include in each component and the relationships among these factors. In the chapter, "A Unified Approach to Two-Level Structural Equation Models and Linear Mixed Effects Models," Peter Bentler and Jiajuan Liang provide a bridge between the first and second parts of the volume in linking LMEs and SEMs, while also considering methods for model selection.

In the chapter, "Bayesian Model Comparison of Structural Equation Models," Sik-Yum Lee and Xin-Yuan Song provide a general Bayesian approach to comparison of SEMs. Typical Bayesian methods for comparing models rely on Bayes factors. However, Bayes factors have proved quite difficult to estimate accurately in SEMs. Lee and Song propose a useful and clever solution to this problem using path sampling. One well-known issue in model selection using Bayes factors is sensitivity to prior selection. This has motivated a rich literature on default priors. In the chapter, "Bayesian Model Selection in Factor Analytic Models" Joyee Ghosh and I build on the approach of Lee and Song, proposing a default prior, and an efficient approach for posterior computation relying on parameter expansion. In addition, an importance sampling algorithm is proposed as an alternative to path sampling.

In summary, this volume provides a practically-motivated overview of a variety of recently proposed approaches for model selection in random effects and latent variable models. The goal is to make these methods more accessible to practitioners, while also stimulating additional research in this important and under-studied area of statistics. There are a number of topics related to model selection in random effects and latent variable models that are in need of new research, with solutions having the potential for substantial applied impact. The first topic is the development of simple methods to calculate model selection criteria, which modify AIC and BIC to incorporate a penalty for model complexity that is appropriate for a hierarchical model. A second topic is the development of efficient methods for simultaneous model search and posterior computation in SEMs. Often, one has a high-dimensional set of SEMs that are plausible a priori and consistent with current scientific or sociologic theories. It is of substantial interest to identify high posterior probability models and to average across models in making predictions. However, typical tricks used in other model classes, such as zeroing out coefficients, do not work in general for SEMs, and efficient alternatives remain to be developed.

References

Bollen, K.A. (1989). *Structural Equation Models with Latent Variables*. New York: Wiley

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

David B. Dunson

http://www.springer.com/978-0-387-76720-8

Random Effect and Latent Variable Model Selection
Dunson, D. (Ed.)
2008, X, 170 p., Softcover
ISBN: 978-0-387-76720-8