UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Bayesian and frequentist cross-validation methods for explanatory item response models

Permalink
https://escholarship.org/uc/item/97q125mp

Author
Furr, Daniel Coulter

Publication Date
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Bayesian and frequentist cross-validation methods for explanatory item response models

by

Daniel C. Furr

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sophia Rabe-Hesketh, Chair
Professor Alan Hubbard
Assistant Professor Zachary Pardos

Summer 2017

Bayesian and frequentist cross-validation methods for explanatory item response models

Copyright 2017
by
Daniel C. Furr

Abstract

Bayesian and frequentist cross-validation methods for explanatory item response models

by

Daniel C. Furr

Doctor of Philosophy in Education

University of California, Berkeley

Professor Sophia Rabe-Hesketh, Chair

The chapters of this dissertation are intended to be three independent, publishable papers, but they nevertheless share the theme of predictive inferences for explanatory item models. Chapter 1 describes the differences between the Bayesian and frequentist statistical frameworks in the context of explanatory item response models. The particular model of focus, the "doubly explanatory model", is a model for dichotomous item responses that includes covariates for person ability and covariates for item difficulty. It includes many Rasch-family models as special cases. Differences in how the model is understood and specified within the two frameworks are discussed. The various predictive inferences available from the model are defined for the two frameworks.

Chapter 2 is situated in the frequentist framework and focuses on approaches for explaining or predicting the difficulties of items. Within the frequentist framework, the linear logistic test model (LLTM) is likely to be used for this purpose, which in essence regresses item difficulty on covariates for characteristics of the items. However, this regression does not include an error term, and so the model is in general misspecified. Meanwhile, adding an error term to the LLTM makes maximum likelihood estimation infeasible. To address this problem, a two-stage modeling strategy (LLTM-E2S) is proposed: in the first stage Rasch model maximum likelihood estimates for item difficulties and standard errors are obtained, and in the second stage a random effects meta-analysis regression of the Rasch difficulties on covariates is performed that incorporates the uncertainty in the item difficulty estimates. In addition, holdout validation, cross-validation, and Akaike information criteria (AIC) are discussed as means of comparing models that have different sets of item predictors. I argue that AIC used with the LLTM estimates the expected deviance of the fitted model when applied to new observations from the *same* sample of items and persons, which is unsuitable for assessing the ability of the model to predict item difficulties. On the other hand, AIC applied to the LLTM-E2S provides the expected deviance related to new observations arising from *new* items, which is what is needed. A simulation study compares parameter recovery and model comparison results for the two modeling strategies.

Chapter 3 takes a Bayesian outlook and focuses on models that explain or predict person abilities. I argue that the usual application of Bayesian forms of information criteria to these models yields the wrong inference. Specifically, when using likelihoods that are conditional on person ability, information criteria estimate the expected fit of the model to new data arising from the *same* persons. What are needed are likelihoods that are marginal over the distribution for ability, which may be used with information criteria to estimate the expected fit to new data from a *new* sample of persons. The widely applicable information criterion (WAIC), Pareto-smoothed importance sampling approximation to leave-one-out cross-validation, and deviance information criterion (DIC) are discussed in the context of these conditional and marginal likelihoods. An adaptive quadrature scheme for use within Markov chain Monte Carlo estimation is proposed to obtain the marginal likelihoods. Also, the moving block bootstrap is investigated as a means to estimate the Monte Carlo error for Bayesian information criteria estimates. A simulation study using a linear random intercept model is conducted to assess the accuracy of the adaptive quadrature scheme and the bootstrap estimates of Monte Carlo error. These methods are then applied to an real item response dataset, demonstrating the practical difference between conditional and marginal forms of information criteria.

Contents

Contents **i**

List of Figures **iii**

List of Tables **iv**

Preface **v**

1 A comparison of the frequentist and Bayesian frameworks in relation to explanatory item response models **1**
1.1 Introduction 1
1.2 A doubly explanatory item response model 2
1.3 Estimated and predicted quantities 8
1.4 Discussion 13

2 Frequentist approaches to cross-validation for item-explanatory models **14**
2.1 Introduction 14
2.2 Models 16
2.3 Model selection strategies 19
2.4 Simulation 23
2.5 Discussion 30

3 Bayesian approaches to cross-validation for person-explanatory models **37**
3.1 Introduction 37
3.2 A simple hierarchical Bayesian model 38
3.3 Information criteria for hierarchical Bayesian models 39
3.4 Adaptive Gaussian quadrature for marginal likelihoods 45
3.5 Circular block bootstrap for estimating Monte Carlo error 46
3.6 Simulation study of adaptive quadrature and circular block bootstrap 47
3.7 Applied example 52
3.8 Discussion 55

Bibliography **57**

A Software details **62**
A.1 Software details for chapter 2 62
A.2 Software details for chapter 3 62

List of Figures

1.1 The doubly explanatory model 3
1.2 Predictive distributions of various forms for responses under the doubly explanatory model 12

2.1 Bias (mean of parameter estimates minus generating values) with 95% confidence intervals for Model 2 26
2.2 Q-Q plots for the observed $\frac{\hat{\beta}_6}{\text{se}(\hat{\beta}_5)}$ for Model 3 across simulation iterations versus standard normal quantiles 27
2.3 Percentages of times each model was chosen for simulation replications in which the proportion of explained item variance is varied 32
2.4 Percentages of times each model was chosen for simulation replications in which the number of items is varied 33
2.5 Estimated penalties implied by holdout validation with new items, AIC, and LOCO-CV 34
2.6 Mean for the root mean squared error of prediction for each model across simulation conditions 35
2.7 The mean for the mean root mean squared error of prediction for the selected model 36

3.1 Differences in marginal information criteria between calculations using adaptive quadrature and the multivariate normal density function 50
3.2 Circular block bootstrap standard error estimates for WAIC by block size for the simulation 52
3.3 Estimated effective number of parameters for the five latent regression Rasch models 54
3.4 Information criteria values for the five latent regression Rasch models 55

List of Tables

1.1 Specification of several special cases of the doubly explanatory model 7

2.1 Generating values for parameters across the simulation conditions 24

3.1 Conditional and marginal DIC, WAIC, and PSIS-LOO for the simulated datasets using the multivariate normal density 49
3.2 Means and standard deviations for the "brute force" WAIC results 51
3.3 Counts of problematic observations for WAIC and PSIS-LOO by model 55

Preface

This dissertation focuses on predictive inferences for item response models that account for factors associated with person ability and item difficulty, known as explanatory item response models. Of particular interest is the use of information criteria for the comparison of competing models, such as models that include different sets of item- or person-related covariates. While the context of explanatory item response models is niche, the insights made in this work also apply more broadly, having implications for model comparison for clustered or cross-classified data in general.

The chapters of this dissertation are intended to be three independent, publishable papers, but they are interrelated nonetheless. Chapter 1 introduces explanatory item response models, comparing model specification in the Bayesian and frequentist statistical frameworks. The various predictive inferences and how they vary depending on framework are explicated. In this way, the first chapter serves as a conceptual basis for the others, which seek to solve specific problems.

Chapter 2 is situated in the frequentist framework and focuses on models for the prediction of item difficulty. I argue that the usual model for item difficulty, the linear logistic test model, is in general misspecified, yielding biased parameter estimates and inaccurate standard errors. Moreover, using the Akaike information criterion with this model provides misleading results. I propose a two-stage estimation strategy that yields better parameter estimates and may be paired with AIC or leave-one-out cross-validation.

Chapter 3 takes a Bayesian outlook and focuses on models for the prediction of person abilities. I argue that the usual application of Bayesian forms of information criteria to these models yields the wrong inference. Specifically, when using likelihoods that are conditional on person ability, information criteria estimate the expected fit of the model to new data arising from the same persons. I propose an adaptive quadrature scheme for use within Markov chain Monte Carlo simulation to obtain likelihoods that are marginal over the ability distribution, which may be used with information criteria to estimate the expected fit of the model to a new sample of persons.

Chapter 1

A comparison of the frequentist and Bayesian frameworks in relation to explanatory item response models

1.1 Introduction

Item response data are cross-classified; that is, any given response to an item is nested both within a person and within an item. In developing a model for such data, the effects on response probabilities of either or both of persons and items may be regarded as arising from a distribution, and the mean of these distributions may be a function of characteristics of persons or items. In this way, an item response model may be described as explanatory if it provides estimates of the effects of person and/or item characteristics, in contrast to purely descriptive models that do not include these sorts of effects.

Explanatory Rasch-family item response models are the focus of this chapter. In the frequentist framework, some explanatory Rasch-family models cannot be estimated using the usual marginal maximum likelihood estimation. In particular, marginal maximum likelihood estimation is infeasible if both the person and item effects are modeled as arising from distributions. However, when effects of persons but not items are assumed to arise from a distribution, a variety of standard software packages are available to fit such models with relative ease. There is no such barrier in the Bayesian framework, as models for cross-classified data may be estimated using Markov chain Monte Carlo (MCMC) simulation. The existing software for MCMC tends to be highly flexible but more cumbersome to use.

In this chapter, frequentist and Bayesian approaches to explanatory item response modeling are compared. Special attention is paid to the predictive inferences that are available under the two frameworks. Despite the particular context of item response models, the discussion applies to models for clustered data more generally.

1.2 A doubly explanatory item response model

General formulation

A useful model for dichotomous item response data is the Rasch model (Rasch, 1960):

$$\Pr(y_{ip}|\theta_p, \delta_i) = \frac{\exp(\theta_p - \delta_i)^{y_{ip}}}{1 + \exp(\theta_p - \delta_i)}, \tag{1.1}$$

where $y_{ip} = 1$ if person p ($p = 1, \ldots, P$) responded to item i ($i = 1, \ldots, I$) correctly and $y_{ip} = 0$ otherwise, θ_p is the ability of person p, and δ_i is the difficulty of item i. The individual instances of θ_p and δ_i may be collected into vectors θ and δ, respectively. This is a "descriptive" item response model (Wilson & De Boeck, 2004); it fully accounts for abilities and difficulties, assuming the appropriateness of the model, but does not offer insight into the factors associated with abilities and difficulties.

The model in Equation 1.1 may be expanded to a "person explanatory" model by decomposing θ_p as

$$\theta_p = w_p'\gamma + \zeta_p, \tag{1.2}$$

where w_p is a row from a design matrix W for person-related covariates, γ is a vector of regression parameters, and ζ_p is the residual person ability. The above may be interpreted as a latent regression of ability on covariates w_p. θ_p may be referred to as the composite ability, $w_p'\gamma$ the structured part of ability, and ζ_p the residual part.

Similarly, decomposing δ_i as

$$\delta_i = x_i'\beta + \epsilon_i \tag{1.3}$$

results in an "item explanatory" model, in which x_i is a row from a design matrix X for item-related covariates, β is a vector of regression parameters, and ϵ_i is the residual item difficulty. The above is then a latent regression of item difficulty on covariates x_i. In parallel with the preceding terminology for ability, δ_i may be referred to as the composite difficulty, $x_i'\beta$ the structured part of difficulty, and ϵ_i the residual part.

Equations 1.1, 1.2, and 1.3 together form a "doubly explanatory" item response model, which incorporates covariates associated with both the persons and items. Note that the model may still serve descriptive purpose as the composite abilities and difficulties remain a part of the model. The final step in formulating the model is to specify distributions for the residuals. In this chapter, normal distributions are assumed,

$$\zeta_p \sim N(0, \sigma^2) \tag{1.4}$$

and

$$\epsilon_i \sim N(0, \tau^2), \tag{1.5}$$

though other choices could be considered. The person and item "sides" of the model are specified in directly parallel ways, and much of the discussion that follows will make use of this point.

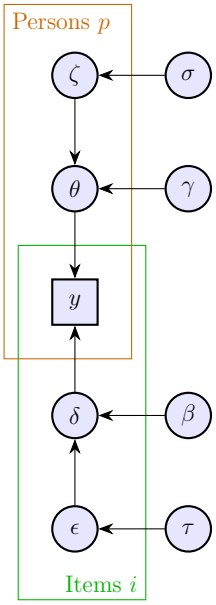


Figure 1.1: The doubly explanatory model presented as a directed graphical model. Circles represent parameters or latent variables, and squares represent data. Person covariates w_p and item covariates x_i are omitted. The boxed regions indicate whether the parameters vary over persons, items or neither.

The model is also presented as a directed graphical model (for example, Dawid, Cowell, Lauritzen, & Spiegelhalter, 1999; Jordan, 2004) in Figure 1.1. In the diagram, unknowns are represented by circles and data are represented by squares. The boxed regions indicate whether the parameters vary over persons, items or neither, and naturally the item responses y vary over both. The direction of the arrows indicates dependence. For example, θ depends on γ and ζ directly, while it depends indirectly on σ. This sort of diagram is associated with Bayesian modeling, where all the unknowns may be considered parameters. From a frequentist perspective, the unknowns are a mixture of parameters and latent variables, which will be discussed in greater depth later.

Hierarchical Bayes modeling approach

In Bayesian methodology, the posterior distribution for the parameters is factorized by way of Bayes theorem:

$$p(\omega|y, W, X) \propto p(\omega)p(y|\omega, W, X), \tag{1.6}$$

which indicates that the posterior distribution is proportional to the product of the prior distribution and the likelihood. In the above, ω is the set of model parameters. In this chapter, the following terminology for different types of parameters is used.

1. *Basic parameters* are the foundational parameters. They are plugged into the likelihood directly or affect it indirectly through intermediate parameters. Priors are specified for them.

    a) *Exchangeable basic parameters* have hierarchical prior distributions. They are exchangeable draws from a distribution, the characteristics of which are determined by hyperparameters. Residuals ζ_p and ϵ_i are examples.

    b) *Non-exchangeable basic parameters* have non-hierarchical priors. They are not thought of as exchangeable or drawn from a distribution, except in the loose sense that there is some prior distribution. Regression coefficients γ and β are examples.

2. *Intermediate parameters* are composites built from basic parameters. They may be included in Bayesian modeling to streamline specifying a model or they may be quantities of genuine interest. They may be plugged into the likelihood in place of basic parameters. They do not have explicit prior distributions, but instead their priors follow from the priors for the basic parameters. Parameters θ_p and δ_i are examples

3. *Hyperparameters* are parameters for the distributions for exchangeable basic parameters. They do have prior distributions themselves.

Of the above terms, only "hyperparameters" is in general usage, while the remaining types of parameters are not typically distinguished from one another.

The doubly descriptive model has a likelihood (Equation 1.1) based on intermediate parameters θ and δ, which are in turn built from basic parameters γ, ζ, β, and ϵ (Equations 1.2 and 1.3). γ and β are non-exchangeable basic parameters, while ζ and ϵ are exchangeable basic parameters whose priors depend on hyperparameters σ and τ, respectively. The prior distribution in Equation 1.6 may be rewritten as

$$p(\omega) = p(\gamma)p(\sigma)\left[\prod_{p=1}^{P} p(\zeta_p|\sigma)\right]p(\beta)p(\tau)\left[\prod_{i=1}^{I} p(\epsilon_i|\tau)\right] \tag{1.7}$$

if independent priors are specified, which is the usual case. No prior is included for θ and δ as they are wholly determined from the basic parameters. The likelihood part of Equation 1.6 may be rewritten in terms of basic parameters as

$$p(y|\omega, W, X) = \prod_{p=1}^{P}\prod_{i=1}^{I} \Pr(y_{ip}|w_p, x_i, \zeta_p, \gamma, \epsilon_i, \beta), \tag{1.8}$$

or in terms of intermediate parameters as

$$p(y|\omega, W, X) = p(y|\theta, \delta) = \prod_{p=1}^{P}\prod_{i=1}^{I} \Pr(y_{ip}|\theta_p, \delta_i). \tag{1.9}$$

Given that both the prior and the likelihood may be specified ignoring the intermediate parameters θ and δ, it is clear that they are redundant. For many applications, however, they have useful interpretations, and for that reason estimation of their posteriors may be desired. Further, the posterior distributions of θ and δ can be estimated easily from the posterior draws of the basic parameters.

The posterior for a single parameter, marginal in regards to all other parameters, may be obtained by integrating the full joint posterior over all other parameters. Let $D = \{y, W, X\}$, represent the full data. Then,

$$p(\sigma|D) = \iiiint\!\!\int p(\zeta, \gamma, \sigma, \epsilon, \beta, \tau|D)\ d\zeta d\gamma d\epsilon d\beta d\tau \tag{1.10}$$

is the posterior for the standard deviation of the ability residuals. The mean and standard deviation of the marginal posterior for a parameter may be taken to represent a point estimate and standard error. Further, the joint posterior of a subset of parameters, $p(\beta, \zeta_p, \gamma, \epsilon_i|D)$ for example, likewise may be obtained by integrating out the other parameters. Despite the high-dimensional integral involved, these quantities are readily available from Monte Carlo simulation by simply ignoring the draws for the parameters to be integrated out, and so no special effort is required to obtain them.

The model could equivalently be specified using hierarchical centering (Gelfand, Sahu, & Carlin, 1995) by replacing the preceding prior with

$$p(\omega) = p(\gamma)p(\sigma)\left[\prod_{p=1}^{P} p(\theta_p|w_p, \gamma, \sigma)\right] p(\beta)p(\tau)\left[\prod_{i=1}^{I} p(\delta_i|x_i, \beta, \tau)\right] \tag{1.11}$$

where $p(\theta_p|w_p, \gamma, \sigma) = \mathrm{N}(w_p\gamma, \sigma^2)$ and $p(\delta_i|x_i, \beta, \tau) = \mathrm{N}(x_i\beta, \tau^2)$, respectively. The likelihood is still specified as in Equation 1.9. In this formulation, residuals ζ and ϵ are omitted altogether, and θ and δ are treated as exchangeable (conditional on covariates) basic parameters rather than as intermediate parameters. Depending on the data and on the algorithm used, this formulation may improve the efficiency of the MCMC simulation. Because this chapter includes discussion of inferences related to the residuals, the "decentered" formulation described before is preferred.

Frequentist modeling approach

In the frequentist approach, only the non-exchangeable basic parameters (γ and β) and the hyperparameters (σ and τ) are treated as parameters to be estimated. In this framework, marginal maximum likelihood may be used to estimate the model, which involves marginalizing the residuals out of the likelihood. The marginal likelihood is

$$p(y|\gamma, \beta, \sigma, \tau) = \int \cdots \iint \cdots \int \prod_{p=1}^{P}\prod_{i=1}^{I} \left[\Pr(y_{ip}|\zeta_p, \gamma, \epsilon_i, \beta)p(\zeta_p|\sigma)p(\epsilon_i|\tau)\right]\ d\zeta_1 \ldots d\zeta_P d\epsilon_1 \ldots d\epsilon_I, \tag{1.12}$$

in which the probability of a response is marginal over the distributions for person and item residuals. Point estimates $\hat{\gamma}$, $\hat{\sigma}$, $\hat{\beta}$, and $\hat{\tau}$ are obtained by maximizing this likelihood

Within this framework, the exchangeable parameters ζ and ϵ are called latent variables or random effects because parameters cannot have distributions. Rather than obtain direct estimates for random effects, marginal maximum likelihood estimation obtains estimates for the parameters of their distributions only, in this case, $\hat{\sigma}$ and $\hat{\tau}$. The non-exchangeable basic parameters γ and β are sometimes referred to as "fixed-effects."

A model of this kind may be formulated in the generalized linear mixed model framework. The response variable, conditional on covariates and so-called random effects, is specified as arising from a Bernoulli distribution:

$$y_{ip}|w_p, x_i, \zeta_p, \epsilon_i \sim \text{Bernoulli}(\pi_{ip}). \tag{1.13}$$

Then the model may be written in terms of an inverse link function

$$\pi_{ip} = \Pr(y_{ip} = 1|w_p, x_i, \zeta_p, \epsilon_i) = \text{logit}^{-1}[\eta_{ip}] \tag{1.14}$$

and a linear predictor

$$\eta_{ip} = (w_p'\gamma + \zeta_p) - (x_i'\beta + \epsilon_i). \tag{1.15}$$

Because the random-effects ζ_p and ϵ_i are not nested, the model may be described as a crossed-random effects model. Such a model is difficult to estimate efficiently via marginal maximum likelihood because the integrals in Equation 1.12 do not factorize as they do with nested random effects. The result is an $I \times P$ dimensional integral, though Rasbash and Goldstein (1994) describe a means of reducing this to an $I + 1$ dimensional integral.

Special cases

Many dichotomous item response models are special cases of the doubly explanatory model that arise from restrictions placed on the composite abilities and difficulties. For example, the Rasch model (Rasch, 1960) as fit by marginal maximum likelihood estimation (Bock & Aitkin, 1981) can be written as

$$\Pr(y_{ip}|\theta_p, \delta_i) = \frac{\exp(\theta_p - \delta_i)^{y_{ip}}}{1 + \exp(\theta_p - \delta_i)} \tag{1.16}$$

$$\theta_p = \zeta_p \tag{1.17}$$

$$\delta_i = x_i'\beta, \tag{1.18}$$

where X is an $I \times I$ identity matrix (I_I) and β is a vector of length I, such that $\delta_i = \beta_i$. In other words, δ_i is set equal to the (unstructured) structural part of item difficulty, while θ_p is set equal to the ability residuals.

Model	θ_p	δ_i	Notes
MML Rasch	ζ_p	$x_i'\beta$	$X = I_I$
JML Rasch	$w_p'\gamma$	$x_i'\beta$	$W = I_{P-1}$, $X = I_I$
Random item Rasch	ζ_p	ϵ_i	
Latent regression	$w_p'\gamma + \zeta_p$	$x_i'\beta$	$X = I_I$
Linear logistic test	ζ_p	$x_i'\beta$	
Linear logistic test with error	ζ_p	$x_i'\beta + \epsilon_i$	
Doubly explanatory	$w_p'\gamma + \zeta_p$	$x_i'\beta + \epsilon_i$	

Table 1.1: Specification of several special cases of the doubly explanatory model.

In the Bayesian approach, the posterior for this Rasch model variant is given by

$$p(\theta, \sigma, \delta|y) \propto \left[p(\delta)p(\sigma)\prod_{p=1}^{P} p(\theta_p|\sigma)\right]\left[\prod_{p=1}^{P}\prod_{i=1}^{I} \Pr(y_{ip}|\theta_p, \delta_i)\right], \tag{1.19}$$

in which the left hand bracketed quantity is the prior and and the right hand quantity is the likelihood. The marginal likelihood for the frequentist approach is

$$p(y|\sigma, \delta) = \prod_{p=1}^{P}\int \prod_{i=1}^{I} \Pr(y_{ip}|\theta_p, \delta_i)p(\theta_p|\sigma)\ d\theta_p. \tag{1.20}$$

The single-dimensional integration is simpler than the $I \times P$ dimensional integral in Equation 1.12 and may be approximated using adaptive quadrature (Rabe-Hesketh, Skrondal, & Pickles, 2002).

Other special cases arise from different choices of restrictions placed on the composite abilities and difficulties, and these are summarized in Table 1.1. As mentioned earlier, the Rasch model as fit by joint maximum likelihood estimation (for example, Embretson & Reise, 2000) includes only the structured parts of ability and difficulty with identity matrices for W and X (one difficulty or ability parameter must be constrained for identifiability). In contrast, the random item Rasch model (for example, De Boeck, 2008) has only the residual parts for both sides (a model intercept must be added). The latent regression item response model (Mislevy, 1985; Adams, Wilson, & Wu, 1997) includes both parts of the composite ability and the structured part of item difficulty, where X is an identity matrix. The linear logistic test model (LLTM) (Fischer, 1973), has the residual part for ability and the structured part for difficulty. Its extension, the linear logistic test model with error (LLTM-E) (for example, Mislevy, 1988; Janssen, Schepers, & Peres, 2004), adds an item difficulty residual.

1.3 Estimated and predicted quantities

Several quantities from the model may be of interest, whether they are estimated directly or obtained after estimation. At the macro-level, γ represents the effects of the person covariates, and $W\gamma$ together with σ describes the conditional distribution for person abilities. Likewise, β represents the effects of the item covariates, and $X\beta$ together with τ describes the conditional distribution for item difficulties. Depending on the choice of either a frequentist and Bayesian framework, the maximum likelihood estimates $\hat{\gamma}$ and $\hat{\beta}$ or posterior distributions $p(\gamma|D)$ and $p(\beta|D)$ will be obtained for these parameters.

For some applications, such as measurement "per se", the specific persons and items will be of interest. This is the case when, for example, measurements are needed for person abilities and a Wright map (Wilson, 2004) is used in interpreting them in relation to the item difficulties. In this case, attention will be placed on θ and δ, though ζ and ϵ may be of interest in the identification of outliers. These are within-sample quantities; that is, the estimation sample contains a person p who is associated with ζ_p and θ_p and also an item i that is associated with ϵ_i and δ_i.

There may be a (real or hypothetical) person p' not represented in the estimation data. This out-of-sample person has a covariate vector $w_{p'}$ and is associated with parameters $\tilde{\zeta}_{p'}$ and $\tilde{\theta}_{p'}$, none of which play a role in fitting the model. Likewise, an out-of-sample item i' associated with $x_{i'}$, $\tilde{\epsilon}_{i'}$, and $\tilde{\delta}_{i'}$ may be envisioned. Inferences for these out-of-sample quantities may be obtained from the fitted model.

Inferences for the within-sample quantities θ_p, δ_i, ζ_p, and ϵ_i are called predictions in the frequentist framework because they are random variables (and not parameters) that are not directly estimated from the model. The same inferences are estimates in a Bayesian setting where ζ_p and ϵ_i are drawn from the posterior and θ_p δ_i are functions of parameters drawn from the posterior. Inferences for the out-of-sample quantities $\tilde{\theta}_{p'}$, $\tilde{\delta}_{i'}$, $\tilde{\zeta}_{p'}$, and $\tilde{\epsilon}_{i'}$ are considered predictions in either case.

Lastly, inferences may be made regarding new responses, which are always considered predictions. A new response may be conceived as arising from a within-sample person to a within-sample item, indicated by $\tilde{y}_{ip}$. This is, in other words, simply a model-predicted response for an existing observation. Several possibilities exist for out-of-sample responses: $\tilde{y}_{i'p}$ represents a new response from a within-sample person to an out-of-sample item, $\tilde{y}_{ip'}$ represents a new response from an out-of-sample person to a within-sample item, and $\tilde{y}_{i'p'}$ represents a new response when both the associated item and person are out-of-sample.

Inferences for residuals

Starting with the Bayesian perspective, the posterior for residual ζ_p is

$$p(\zeta_p|D) = \iiiiiint p(\zeta, \gamma, \sigma, \epsilon, \beta, \tau|D)\ d\zeta_{-p} d\gamma d\sigma d\epsilon d\beta d\tau, \tag{1.21}$$

where ζ_{-p} is the vector ζ omitting ζ_p. This is simply the full posterior integrating out all other parameters and its distribution be approximated in MCMC simulation simply by ζ_p^s, where $s = 1 \dots S$ indexes the draws from the simulation. The distribution for the residual of a new person, $\tilde{\zeta}_{p'}$, is

$$p(\tilde{\zeta}_{p'}|D) = \int p(\tilde{\zeta}_{p'}|\sigma)p(\sigma|D)\ d\sigma, \tag{1.22}$$

which is referred to as a mixed predictive distribution (Gelman, Meng, & Stern, 1996). It may be approximated by taking random draws for $\tilde{\zeta}_{p'}^s$ from its prior, $p(\tilde{\zeta}_{p'}|\sigma^s)$. On the item side, the parallel quantities are

$$p(\epsilon_i|D) = \iiiiiint p(\zeta, \gamma, \sigma, \epsilon, \beta, \tau|D)\ d\zeta d\gamma d\sigma d\epsilon_{-i} d\beta d\tau, \tag{1.23}$$

and

$$p(\tilde{\epsilon}_{i'}|D) = \int p(\tilde{\epsilon}_{i'}|\tau)p(\tau|D)\ d\tau. \tag{1.24}$$

Marshall and Spiegelhalter (2007) have recommended using mixed predictive distributions like $p(\tilde{\zeta}_{p'}|D)$ and $p(\tilde{\epsilon}_{i'}|D)$ to detect outlying residuals.

From the frequentist perspective, "empirical Bayes" predictions for residuals may be obtained post-estimation. The empirical Bayes mean prediction for ζ_p is

$$\hat{\zeta}_p^{\text{EB}} = \int \zeta_p\ p(\zeta_p|D, \hat{\gamma}, \hat{\sigma}\hat{\beta}, \hat{\tau})\ d\zeta_p, \tag{1.25}$$

where $p(\zeta_p|D, \hat{\gamma}, \hat{\sigma}, \hat{\beta}, \hat{\tau})$ is the conditional posterior

$$p(\zeta_p|D, \hat{\gamma}, \hat{\sigma}, \hat{\beta}, \hat{\tau}) \propto p(\zeta_p|\hat{\sigma})p(y_p|w_p, X, \hat{\gamma}, \zeta_p, \hat{\beta}, \hat{\tau}). \tag{1.26}$$

The rightmost quantity in the above is the likelihood conditional on ζ_p but marginal in regard to ϵ:

$$p(y_p|w_p, X, \hat{\gamma}, \zeta_p, \hat{\beta}, \hat{\tau}) = \int p(\epsilon|\hat{\tau})p(y_p|w_p, X, \hat{\gamma}, \zeta_p, \hat{\beta}, \epsilon)\ d\epsilon. \tag{1.27}$$

The above form for the empirical Bayes prediction is more complicated than usual owing to the need to integrate out the ϵ vector, which arises from the model being for cross-classified data. Instead of the empirical Bayes mean prediction, the modal prediction may be obtained by finding the mode of the conditional posterior. Of course, the empirical Bayes prediction for either the mean or mode of ϵ_i may be written in a way parallel to that for ζ_p. The main difference between the frequentist empirical Bayes approach and actual Bayesian approach is the propagation of uncertainty; while the frequentist approach treats the model parameters as known when obtaining the prediction, the Bayesian approach incorporates the residuals as a part of the full posterior. Lastly, frequentists may take $p(\tilde{\zeta}_p|\hat{\sigma})$ and $p(\tilde{\epsilon}_i|\hat{\tau})$ as representing the distributions for new instances of the residuals, and as both have a mean of zero, zero may be assigned as the point predictions for residuals for new persons or new items.

Inferences for composites

Returning to the Bayesian perspective, the posterior for composites like $p(\theta_p|D)$ are easily approximated from the posterior draws of MCMC simulation:

$$\theta_p^s = w_p'\gamma^s + \zeta_p^s. \tag{1.28}$$

The posterior for a new composite ability, $p(\tilde{\theta}_{p'}|D)$, is approximated by the empirical distribution of

$$\tilde{\theta}_{p'}^s = w_{p'}'\gamma^s + \tilde{\zeta}_{p'}^s \tag{1.29}$$

where $w_{p'}$ is the covariate vector for the new person and $\tilde{\zeta}_{p'}^s$ is as given above. In a parallel way, $p(\delta_i|D)$ and $p(\tilde{\delta}_{i'}|D)$ may be approximated by the distributions of

$$\delta_i^s = x_i'\beta^s + \epsilon_i^s. \tag{1.30}$$

and

$$\tilde{\delta}_{i'}^s = x_{i'}'\beta^s + \tilde{\epsilon}_{i'}^s, \tag{1.31}$$

respectively.

In the frequentist perspective, a prediction for an in-sample composite ability is a combination of the regression prediction and the empirical Bayes estimate for the residual:

$$\hat{\theta}_p = w_p'\hat{\gamma} + \hat{\zeta}_p^{\mathrm{EB}}. \tag{1.32}$$

For an out-of-sample composite ability, the residual part of the prediction may be set to zero (the mean of residuals):

$$\tilde{\theta}_{p'} = w_{p'}'\hat{\gamma}^s. \tag{1.33}$$

The equivalent quantities on the item side are

$$\hat{\delta}_i = x_i'\hat{\beta} + \hat{\epsilon}_i^{\mathrm{EB}} \tag{1.34}$$

and

$$\tilde{\delta}_{i'} = x_{i'}'\hat{\beta}^s. \tag{1.35}$$

As with the predictions for residuals, each of these are point estimates and do not involve the propagation of uncertainty realized in Bayesian modeling.

Inferences for responses

Returning again to the Bayesian framework, the posterior predictive distribution (Rubin, 1984) for new a response $\tilde{y}_{ip}$ from a within-sample person and item is

$$p(\tilde{y}_{ip}|D) = \iint \Pr(\tilde{y}_{ip}|\theta_p, \delta_i)p(\theta_p, \delta_i|D)\ d\theta_p d\delta_i \tag{1.36}$$

$$= \iiiint \Pr(\tilde{y}_{ip}|w_p, x_i, \gamma, \zeta_p, \beta, \epsilon_i)p(\gamma, \zeta_p, \beta, \epsilon_i|D)\ d\gamma d\zeta_p d\beta d\epsilon_i. \tag{1.37}$$

The predictive distribution for a new response arising from an out-of-sample person and out-of-sample item is

$$p(\tilde{y}_{i'p'}|D) = \iint \Pr(\tilde{y}_{i'p'}|\tilde{\theta}_{p'}, \tilde{\delta}_{i'}) p(\tilde{\theta}_{p'}, \tilde{\delta}_{i'}|D)\ d\tilde{\theta}_{p'} d\tilde{\delta}_{i'} \tag{1.38}$$

$$= \iiiint \Pr(\tilde{y}_{i'p'}|w_{p'}, x_{i'}, \gamma, \tilde{\zeta}_{p'}, \beta, \tilde{\epsilon}_{i'}) p(\gamma, \tilde{\zeta}_{p'}, \beta, \tilde{\epsilon}_{i'}|D)\ d\gamma d\tilde{\zeta}_{p'} d\beta d\tilde{\epsilon}_{i'}, \tag{1.39}$$

where $p(\tilde{\theta}_{p'}, \tilde{\delta}_{i'}|D)$ is the joint mixed predictive distribution for $\tilde{\theta}_{p'}$ and $\tilde{\delta}_{i'}$, and $p(\gamma, \tilde{\zeta}_{p'}, \beta, \tilde{\epsilon}_{i'}|D)$ includes the mixed predictive distributions for $\tilde{\zeta}_{p'}$ and $\tilde{\epsilon}_{i'}$, as described previously. In MCMC simulation, $p(\tilde{y}^s_{ip}|D)$ may be obtained as a random draw from Bernoulli($\text{logit}^{-1}(\theta^s_p - \delta^s_i)$), and likewise $p(\tilde{y}^s_{i'p'}|D)$ may be obtained as random draw from Bernoulli($\text{logit}^{-1}(\tilde{\theta}^s_{p'} - \tilde{\delta}^s_{i'})$).

Figure 1.2 shows four ways of making inferences for new responses. On the left side of each panel is a graphical representation of the model, similar to the one shown earlier, though the boxed regions indicating which parameters vary over persons and which vary over items are omitted for simplicity. On the right side of each is a shaded region for the out-of-sample predictions. Figure 1.2a shows that the posterior distribution for $\tilde{y}_{ip}$, a new response from an in-sample person and in-sample item, arises directly from an existing θ_p and δ_i pair. In this way, it is clear that $\tilde{y}_{ip}$ is closely related to the observed y_{ip}. Figure 1.2b shows that the posterior for $\tilde{y}_{i'p'}$ arises from the mixed predictive distributions for $\tilde{\theta}_{p'}$ and $\tilde{\delta}_{i'}$. Further, it depicts how the various predictive distributions are influenced by the posteriors for the modeled parameters. Lastly, predictive distributions for responses from a new person to an in-sample item, $p(\tilde{y}_{ip'}|D)$, as well as responses from an in-sample person to a new item, $p(\tilde{y}_{i'p}|D)$, may be obtained by mixing and matching posterior and mixed predictive distributions as needed, as shown in Figures 1.2c and 1.2d.

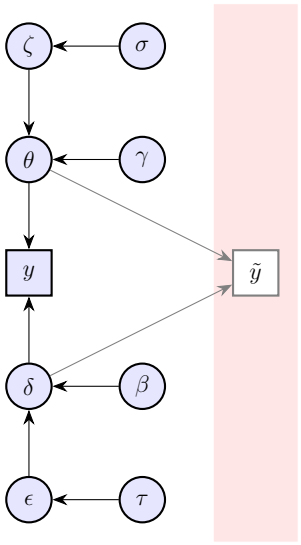
In the frequentist perspective, the predicted probabilities for a correct response are based on point estimates of model parameters, but are otherwise similar to the Bayesian predictions. For a new response from a within-sample person-item pair, the predicted probability of a correct response is

$$\Pr(\tilde{y}_{ip} = 1|w_p, x_i, \hat{\gamma}, \hat{\sigma}, \hat{\beta}, \hat{\tau}) = \iint \Pr(\tilde{y}_{ip} = 1|w_p, x_i, \hat{\gamma}, \zeta_p, \hat{\beta}, \epsilon_i) p(\zeta_p, \epsilon_i|D, \hat{\gamma}, \hat{\sigma}, \hat{\beta}, \hat{\tau})\ d\zeta_p d\epsilon_i. \tag{1.40}$$

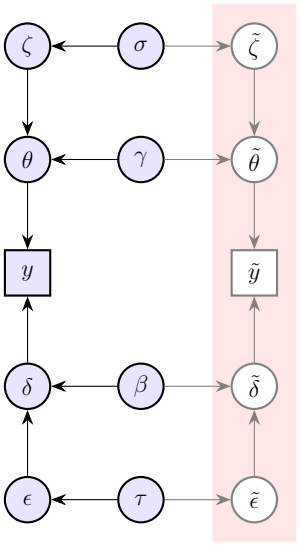
Like empirical Bayes predictions, it uses the conditional posterior for ζ_p and ϵ_i. This corresponds to the cluster-averaged expectation for generalized linear mixed models described by Skrondal and Rabe-Hesketh (2009), except that the prediction given here marginalizes over the posterior for two sets of residuals rather than just one. The predicted probability for a new response from a new person to a new item is

$$\Pr(\tilde{y}_{i'p'} = 1|w_{p'}, x_{i'}, \hat{\gamma}, \hat{\beta}, \hat{\tau}, \hat{\sigma}) = \iint \Pr(\tilde{y}_{i'p'} = 1|w_{p'}, x_{i'}, \hat{\gamma}, \hat{\beta}, \zeta_{p'}, \epsilon_{i'}) p(\zeta_{p'}, \epsilon_{i'}|\hat{\tau}, \hat{\sigma})\ d\zeta_{p'} d\epsilon_{i'}, \tag{1.41}$$

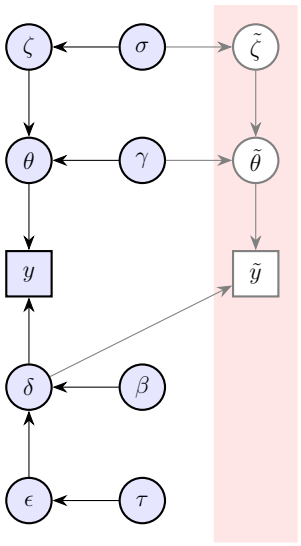
which uses the prior for ζ_p and ϵ_i. This corresponds to what Skrondal and Rabe-Hesketh (2009) refer to as the population-averaged expectation, again with the exception that two



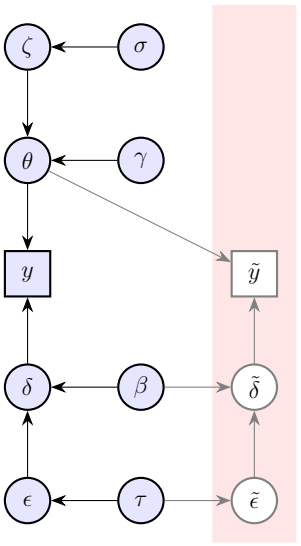
(a) Within-sample persons and items ($\tilde{y}_{ip}$)



(b) Out-of-sample persons and items ($\tilde{y}_{i'p'}$)



(c) Out-of-sample persons and within-sample items ($\tilde{y}_{ip'}$)



(d) Within-sample persons and out-of-sample items ($\tilde{y}_{i'p}$)

Figure 1.2: Predictive distributions of various forms for responses under the doubly explanatory model. Circles represent parameters and squares represent data. The shaded region indicates predictive quantities that are not involved in the estimation. Covariates W and X are omitted.

sets of residuals are involved here. Lastly, predictions for new responses of the form $\tilde{y}_{i'p}$ and $\tilde{y}_{ip'}$ are obtained by mixing the use of posterior and prior distributions for the residuals.

Inferences for special cases

If a special case of the full model is fitted, such as any described in Section 1.2, some predictive inferences may not be available. For example, with the Rasch model (either the marginal or joint maximum likelihood formulations) the predictive distribution for $\tilde{\delta}_{i'}$ is unavailable because for the Rasch model X is a series of indicator variables for the existing items and ϵ and τ are omitted. By extension, predictive distributions for $\tilde{y}_{i'p}$ and $\tilde{y}_{i'p'}$ also cannot be obtained for the Rasch model.

1.4 Discussion

For models that are not hierarchical, frequentist analysis will often be equivalent to a Bayesian analysis using uniform priors. In a more complicated model, like the doubly explanatory model, the results are still expected to be very similar if the priors for Bayesian analysis are uniform or diffuse. Nonetheless, some advantages have been identified in the Bayesian approach.

First, in Bayesian modeling it is natural to make inferences about basic parameters, intermediate parameters, and hyperparameters simultaneously, while frequentist analysis does not directly estimate exchangeable basic parameters or intermediate parameters. Paradoxically, in frequentist item response modeling, the actual measurement of persons, that is obtaining a prediction for θ_p, must occur in a second, post-estimation step when marginal maximum likelihood estimation is used.

Second, Bayesian analysis propagates uncertainty regarding parameters while frequentist analysis does not. For example, frequentist analysis may obtain an empirical Bayes prediction $\hat{\zeta}_p^{\mathrm{EB}}$ that will depend on point estimate $\hat{\sigma}$ and other parameters, and standard errors for $\hat{\zeta}_p^{\mathrm{EB}}$ will be unduly small as $\hat{\sigma}$ is treated as known. In contrast, the Bayesian posterior for ζ_p will be marginal over the posterior for σ (and all other parameters) and so will more accurately represent the uncertainty regarding ζ_p. The difference is more pronounced with an intermediate parameter like θ_p, as the true Bayesian posterior for it will also reflect the uncertainty regarding γ in addition to ζ_p.

Chapter 2

Frequentist approaches to cross-validation for item-explanatory models

2.1 Introduction

Several models have been developed that account for the factors associated with item difficulty: the linear logistic test model (LLTM; Fischer, 1973), the linear logistic test model with error (LLTM-E Janssen et al., 2004), the 2PL-constrained model (Embretson, 1999), the additive multilevel item structure model (Cho, De Boeck, Embretson, & Rabe-Hesketh, 2014), and others. Such models are described as item explanatory models by Wilson and De Boeck (2004) and lend themselves to the prediction of item difficulties for new items. Predictions regarding new items are useful in automatic item generation, especially in regards to adaptive testing when the goal is to generate an optimally informative item during administration (see for example, Embretson, 1999). Such a prediction is sometimes referred to as "precalibration" (see for example, Gierl & Haladyna, 2013). Even when item generation is not automatic, the ability to anticipate the difficulty of new items may be useful in the development of new test forms.

The best set of predictors for item difficulty may not be known a priori, and in this case a model selection strategy may be employed to select a preferred model for the prediction of new item difficulties from a set of candidate models. A model selection strategy requires a choice of a score function to evaluate the prediction error, and the deviance (-2 times the log-likelihood) is a natural choice. Holdout validation, cross-validation, or information criteria may be used to select a preferred model on the basis of the score function. In holdout validation, a model is estimated on one dataset and then evaluated in a second dataset using the score function. Cross-validation is similar but involves splitting the data multiple times and aggregating the results across splits. The Akaike Information Criterion (AIC; Akaike, 1974) is asymptotically equivalent to cross-validation (Stone, 1977) but requires only a single

fit of the model. In holdout validation, the estimated prediction error is conditional on the fitted model, whereas cross-validation and AIC approximate the expected prediction error, with the expectation taken over possible training datasets (Hastie, Tibshirani, & Friedman, 2009, chapter 7).

Predictive utility is a good basis for model selection in general even if the goal is not actually prediction per se. In particular, if it is believed that none of the candidate models are true, then the model that best predicts new data may be justified as the best available approximation to the unknown true model. For example, a researcher may be interested in identifying the factors associated with item difficulty in order to develop or evaluate the theory pertaining to the latent construct if interest. In this way, the purpose of modeling item difficulty may be explanation rather than prediction, but predictive utility may still form the basis for selecting a preferred model. It must be noted that a true model, even if it were available, may not be the most predictive model when it is estimated with finite data, owing to potentially high variance in the parameter estimates. However, this possibility need not be troubling because it is more realistic to consider models as being approximations to a complex reality than as being true data generating mechanisms. Further, it is realistic and appropriate that additional information, in the form of increased amounts of data, should affect judgments about which model best approximates the unknown true model.

Obtaining predictions for new data requires consideration of how new observations would arise. For the case of independent observations, new observations come about in a straightforward way. For the case of clustered observations, new observations may come from within the existing clusters or instead from within new clusters. Item response data are more complicated still as the observations are cross-classified, or in other words, responses are clustered simultaneously within persons and items. As a result, new item responses could arise from any combination of new or same persons and new or same items. In most applications, predictions are made for the response variable, but with clustered data prediction may also be made for the clusters themselves. For item response data, this could mean predictions for item difficulties or for person abilities, though this chapter focuses on the prediction of new item difficulties.

Generalized linear mixed models are commonly used in the analysis of clustered data. In this framework, models are built from "fixed" and "random" effects. Fixed effects are much like standard regression coefficients and are directly estimated. Fixed effects are usually constant across clusters, while random effects are cluster-specific. Random effects are not estimated directly, but instead estimation focuses on the parameters of their assumed joint distribution. In this way, the clusters are treated as sampled from a distribution, implying that a new data collection would entail a new set of clusters. An alternative modeling strategy is to treat cluster-specific effects as fixed effects, in which case the clusters are fixed rather than random, implying that new observations would arise from the existing set of clusters.

Rasch family item response models, including the LLTM but not the LLTM-E, are readily specified as generalized linear mixed models (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). Items are customarily modeled as fixed effects, for example as item-specific parameters

for the Rasch model or as a weighted sum of covariates for the LLTM, and persons are modeled as random effects, perhaps with fixed effects for the mean structure of the person ability distribution (for example, Zwinderman, 1991; Adams et al., 1997). In this way, common item response models imply that persons are random, and so would be different in new data, and that the items would be the same in new data. It follows that such a modeling strategy would be well-suited to the selection of a preferred model for predicting person abilities but not for predicting item difficulties, which presents a paradox for the LLTM. In particular, information criteria may perform poorly for the LLTM given that they are based on a likelihood that treats the items as fixed, while holdout validation or cross-validation may perform more reasonably so long as they are based on new items.

The LLTM-E treats both the persons and items as random, which is reflected in the likelihood for it. For this reason, information criteria used with the LLTM-E should exhibit correct behavior, unlike for the LLTM. However, given that the model is simultaneously marginal in regard to persons, and the persons and items are crossed, it is infeasible to estimate using marginal maximum likelihood. I propose a two-stage estimation method for the LLTM-E, called the LLTM-E2S, that appropriately treats the items as random.

In this chapter, simulation study results for several model selection strategies are reported for the LLTM and LLTM-E2E. It is expected that holdout validation using a new set of items will yield better item predictions than holdout validation repeating the same set of items. For the LLTM, model selection results using AIC are expected to resemble results from holdout validation with the same items, given that the likelihood for the LLTM treats items as fixed. For the LLTM-E2S, model selection results using AIC are expected to resemble results from holdout validation with new items, given that its likelihood treats items as random.

2.2 Models

The linear logistic test model with error

The data generating model in the simulation study is the LLTM-E:

$$\Pr(y_{ij} = 1|x_i, \theta_j) = \text{logit}^{-1}\left[\theta_j - (x_i'\beta + \epsilon_i)\right] \tag{2.1}$$

$$\theta_j \sim \text{N}(0, \sigma^2) \tag{2.2}$$

$$\epsilon_i \sim \text{N}(0, \tau^2), \tag{2.3}$$

where $y_{ij} = 1$ if person j ($j = 1, \cdots, J$) responded to item i ($i = 1, \cdots, I$) correctly and $y_{ij} = 0$ otherwise. Latent ability is denoted by θ_j, which follows a normal distribution. The quantity $x_i'\beta + \epsilon_i$ is a latent regression of item difficulty in which x_i is a vector of item covariates, β is a vector of regression coefficients, and ϵ_i is a residual. The residual is necessary because it is unrealistic that the item covariates would perfectly account for item difficulty. Further, x_i is a row from a matrix of item covariates X, which will in general include a column with all elements equal to one for the model intercept. The model may

be understood as a generalization of the Rasch model (Rasch, 1960) that decomposes the Rasch difficulty parameters into a structural part ($x_i'\beta$) and residual part (ϵ_i). Omitting the residual part from the LLTM-E yields the standard LLTM.

Fitting the model using marginal maximum likelihood estimation is infeasible, given the need to integrate over the vectors θ and ϵ simultaneously when calculating the marginal likelihood. Maximizing the likelihood is equivalent to minimizing the deviance, which is -2 times the log likelihood. The deviance for the LLTM-E is

$$\text{dev}(y|\hat{\omega}_m(y)) = -2\log \int \cdots \int \left[\prod_{i=1}^{I} \prod_{j=1}^{J} \Pr(y_{ij}|\hat{\beta}, \epsilon_i, \theta_j) \phi(\epsilon_i; 0, \hat{\tau}^2) \phi(\theta_j; 0, \hat{\sigma}^2) \right] \, \text{d}\epsilon \text{d}\theta, \quad (2.4)$$

where ϕ is the normal density function. The integral does not factorize and at best may be reduced from $I + J$ to $I + 1$ dimensional integrals (Goldstein, 1987; Rasbash & Goldstein, 1994). In the above equation, $\hat{\omega}_m(y)$ is shorthand for all estimated parameters ($\hat{\beta}$, $\hat{\sigma}$, and $\hat{\tau}$) for model m, which are estimated from data $\{x, y\}$, and the hats on parameters denote marginal maximum likelihood estimates. In the notation $\hat{\omega}_m(y)$, x is omitted for convenience but would be appropriate to include for completeness. Also, for the moment it may seem redundant to indicate that the parameter estimates arise from y in the notation $\hat{\omega}_m(y)$, but this notation will become useful later.

The linear logistic test model

A common model for studying the effects of item covariates, the LLTM (Fischer, 1973), omits the item residual ϵ_i:

$$\Pr(y_{ip} = 1|x_i, \theta_j) = \text{logit}^{-1}\left[\theta_j - x_i'\beta\right] \quad (2.5)$$

$$\theta_j \sim \text{N}(0, \sigma^2). \quad (2.6)$$

Otherwise, the model is the same as the LLTM-E. The likelihood for the LLTM is marginal over persons but not items. Expressing the likelihood in terms of deviance,

$$\text{dev}(y|\hat{\omega}_m(y)) = -2\sum_{j=1}^{J} \log \int \left[\prod_{i=1}^{I} \Pr(y_{ij}|\hat{\beta}, \theta_j) \right] \phi(\theta_j; 0, \hat{\sigma}^2) \, \text{d}\theta_j, \quad (2.7)$$

where $\hat{\omega}_m(y)$ again represents all estimated parameters, this time only $\hat{\beta}$ and $\hat{\sigma}$. Only a one-dimensional integration is involved, and the LLTM is readily fit using marginal maximum likelihood estimation (Bock & Aitkin, 1981). While no closed-form solution exists for the integration due to the logit link function (Equation 2.5), it is easily approximated by adaptive quadrature (Pinheiro & Bates, 1995; Rabe-Hesketh, Skrondal, & Pickles, 2005). As mentioned earlier, the LLTM may be expressed as a generalized linear mixed model and is readily fit in standard software in addition to more specialized software for item response theory models.

Fischer (1997, p. 232) recommended testing the goodness of fit of the LLTM by conducting a likelihood ratio test comparing the LLTM to the Rasch model, suggesting that the LLTM was to be interpreted only if it is not rejected. However, as he admitted, the LLTM will generally be rejected, leaving the researcher with two options: they may either refrain from studying the sources of item difficulty or interpret the LLTM anyway. The danger in the second option, which he did not identify, is that standard errors for the item predictors will be inappropriately small, often substantially so, when the predictions $x_i'\beta$ fail to replicate the "complete" item difficulties $x_i'\beta + \epsilon_i$. This problem directly parallels the situation in multilevel modeling in which a non-hierarchical model is fit to clustered data, and then the omission of cluster-level residuals leads to an overstatement of the amount of information available to estimate coefficients of cluster-level covariates.

Two-stage estimation of the linear logistic test model with error

To avoid the high-dimensional integral in Equation 2.4, I propose a two-stage estimation of the LLTM-E, which I will refer to as the LLTM-E2S. In the first stage, the Rasch model is fit to the data. The Rasch model is

$$\Pr(y_{ij} = 1|\delta_i, \theta_j) = \text{logit}^{-1}\left[\theta_j - \delta_i\right] \tag{2.8}$$

$$\theta_j \sim \text{N}(0, \sigma^2), \tag{2.9}$$

where δ_i is an item-specific difficulty parameter. Point estimates $\hat{\delta}_i$ and standard errors for δ_i are obtained by marginal maximum likelihood estimation, minimizing a deviance similar to that in Equation 2.7. These results are compiled into a constructed dataset of I observations that includes the difficulty estimates, standard errors, and predictors for each item.

In the second stage, the $\hat{\delta}_i$ are regressed on the item covariates using the constructed data set. This second stage model is

$$\hat{\delta}_i = x_i'\beta + u_i + \epsilon_i \tag{2.10}$$

$$u_i \sim \text{N}(0, \widehat{\text{var}}(\hat{\delta}_i)) \tag{2.11}$$

$$\epsilon_i \sim \text{N}(0, \tau^2), \tag{2.12}$$

where u_i is a residual related to uncertainty in the estimated $\hat{\delta}_i$, and ϵ_i is the usual residual in linear regression. The residual u_i has known variance $\widehat{\text{var}}(\hat{\delta}_i)$, which is the square of the estimated standard error for $\hat{\delta}_i$ obtained in the first stage. The variance for ϵ_i, τ^2, is a model parameter to be estimated. This is a random-effects meta-regression model (see for example Raudenbush & Bryk, 1985). For the LLTM-E2S, let $\hat{\omega}_m(y)$ represent the set of parameter estimates from the second stage model. Then the deviance to be minimized in the second stage is

$$\text{dev}(y|\hat{\omega}_m(y)) = \sum_{i=1}^{I} -2\log\phi(\hat{\delta}_i; x_i'\hat{\beta}, \widehat{\text{var}}(\hat{\delta}_i) + \hat{\tau}^2), \tag{2.13}$$

where $\hat{\delta}_i$ are estimates carried over from the first step, and $\hat{\beta}$ and $\hat{\tau}$ are estimates obtained in the second step. This deviance is suitable only for the selection of an item difficulty model. Two-stage estimation has been used elsewhere. For example, Borjas and Sueyoshi (1994) estimate a probit model with dummy variables for group effects and then regress the estimated group effects on group-level covariates. In this way, their two-stage estimation is similar to the LLTM-E2S except that it is for non-cross-classified hierarchical models.

2.3 Model selection strategies

Holdout validation

In holdout validation, a large dataset is split into three parts: the *training*, *validation*, and *evaluation* subsets. (The evaluation subset may also be referred as the test subset.) In this process, parameter estimates are obtained for a model by first fitting it to the training subset, and then the fitted model is used to evaluate the score function in the validation subset. These steps are repeated for each candidate model In this chapter the deviance is used as the score function, and so models are both estimated and evaluated using the deviance. The model with the lowest deviance in the validation subset is selected as the best model and then evaluated a second time in the evaluation subset. The use of a validation subset addresses the bias that would arise from both fitting and evaluating the model on the training subset. The use of an evaluation subset addresses the bias that would arise from selecting and evaluating a model using the validation subset alone.

This chapter extends the usual holdout validation scheme by considering what elements differ or persist between the three data subsets. For the case of selecting a model that best predicts item difficulties, the relevant detail is whether the subsets include the same or different items. Let y^{t} be the responses from the training subset. A validation subset may include the *same* items as the training subset, and the responses from such a validation subset will be denoted y^{s}. Alternatively, a validation subset might include a *new* set of items, and the responses associated with that training subset will be denoted y^{n}. Also, let y^{e} be the responses from the evaluation subset, which in this chapter will always contain a set of items distinct from those in the other subsets. Each of y^{t}, y^{s}, y^{n}, and y^{e} are assumed to arise from separate, random samples of persons. Further, each of y^{t}, y^{s}, y^{n}, and y^{e} is associated with item covariates x^{t}, x^{s}, x^{n}, and x^{e} respectively, though in general this chapter will omit the item covariates from notation.

A candidate model is selected based on $\mathrm{dev}(y^{\mathrm{n}}|\hat{\omega}_m(y^{\mathrm{t}}))$ or $\mathrm{dev}(y^{\mathrm{s}}|\hat{\omega}_m(y^{\mathrm{t}}))$, depending on the form of the validation subset. Let m^* represent the model selected from this process. The deviance for it in the evaluation subset is $\mathrm{dev}(y^{\mathrm{e}}|\hat{\omega}_{m^*}(y^{\mathrm{t}}))$, where m^* may differ depending on which type of validation subset was used. In this chapter, the evaluation subset always contains new items; that is, items that are different from those featured in the training and validation subsets.

The estimated prediction error (deviance) in holdout validation is conditional on the

particular training data used. This is clear in the notation $\text{dev}(y^{\text{e}}|\hat{\omega}_{m^*}(y^{\text{t}}))$, in which the deviance of the chosen model in the evaluation subset is conditional on parameter estimates obtained from the training subset. This fact distinguishes holdout validation from the cross-validation methods discussed in the next section.

In summary, two approaches to holdout validation for item prediction are considered in this chapter: one in which the validation subset features the same items as the training subset and one in which the validation subset features new items. Whether the LLTM or LLTM-E2S is used, holdout validation with the same items is expected to perform poorly because the training and validation subsets will be similar, particularly in regards to the items. It may be expected to choose overly complex models too often, as idiosyncrasies related to the items, specifically the realizations of ϵ_i, will be repeated in the two subsets. Holdout validation with new items is expected to choose the model with the correct set of item predictors, or when the amount of information in the data is low, a simpler model. The LLTM-E2S may be more successful in this regard than the LLTM, given that its second stage deviance is more targeted toward item prediction than the deviance for the LLTM.

Cross-validation and AIC

If the available data are not abundant enough to support holdout validation, single dataset methods for model selection may be considered instead. In k-fold cross-validation, the data are split into K (approximately) equally sized partitions, most often $K = 5$ or 10. A model is fit to all data *not* in fold k, and then the fitted model is evaluated using the score function on the data in fold k. This process is performed for every fold, and the resulting deviances are summed over the folds. Hastie et al. (2009) provide a thorough description of k-fold cross-validation. Asymptotically, the model selected by cross-validation performs as well as the candidate model that minimizes the loss function with respect to the true probability distribution, and this is sometimes referred to as the oracle property (van der Laan & Dudoit, 2003).

When the observations are clustered, k-fold cross-validation may or may not keep the clusters intact, depending on the whether the desired inference requires new clusters. For item response data, in which item and person clusters are crossed, either the person clusters or the item clusters may be kept intact. For example, if the goal is to compare models that explain the difficulty of $I = 20$ items differently and $K = 5$ folds are used, then each fold should contain all responses for four ($\frac{I}{K}$) items. In this case, the items clusters are held intact, as the responses to a given item are not split up across folds. However, the person clusters are broken up across folds, as each fold will have only four responses per person. On the other hand, if the purpose is to compare models with different sets of person covariates, folds should instead keep the person clusters intact.

For models for item prediction, a possibility is to assign each item to its own fold, which

may be referred to as leave-one-cluster-out cross-validation (LOCO-CV). Then

$$\text{LOCO-CV}_m = \sum_{i=1}^{I} \text{dev}(y_i|\hat{\omega}_m(y_{-i})) \tag{2.14}$$

where y_i indicates all responses for item i, and y_{-i} indicates all responses not associated with item i. The LLTM-E2S is particularly suited to performing LOCO-CV, given that the first stage (fitting the Rasch model) need only be carried out once. Then the second stage is performed I times per candidate model, leaving one item out and then obtaining a prediction for the left out item difficulty. These predictions are substituted for $x_i'\hat{\beta}$ in Equation 2.13 to obtain LOCO-CV$_m$. LOCO-CV could be performed for the LLTM, but it is time-consuming when compared to the LOCO-CV with the LLTM-E2S. For this reason, the simulations that follow only use the LLTM-E2S when performing LOCO-CV.

The Akaike Information Criterion (AIC; Akaike, 1974) requires only a single fit of the model and has the form

$$\text{AIC} = \text{dev}(y|\hat{\omega}_m(y)) + 2q_m, \tag{2.15}$$

where q_m is the count of parameters in model m. AIC is an approximation related to the Kullback-Leibler distance, which is a measure of the information lost when a model is used to approximate the true data generating distribution. Calculating the Kullback-Leibler distance would require knowing the true data generating distribution, but it is possible to approximate the expected *relative* distance

$$-2 \text{ ERD}_m = \text{E}_{y^{\text{v}}}\text{E}_{y^{\text{t}}}[\text{dev}(y^{\text{v}}|\hat{\omega}_m(y^{\text{t}}))] \approx \text{AIC}, \tag{2.16}$$

where L is the log-likelihood rather than deviance, is tractable. In the above equation, y^{t} and y^{v} are (hypothetical) independent datasets and the expectations are taken over the true data generating distribution for y^{t} and y^{v}. The difference between the Kullback-Leibler distance and expected relative distance is an unknown constant that is a function only of the true data generating distribution. This constant will be the same for all candidate models because it does not depend on the models. AIC is an approximation to the expected relative distance multiplied by negative two, putting it on the scale of deviance. For models without random effects, AIC is asymptotically equivalent to "leave-one-observation-out" cross-validation (Stone, 1977), and for models with random effects it is asymptotically equivalent to LOCO-CV (Fang, 2011), at least for linear mixed effects models.

Kuha (2004) describes two requirements for AIC to be a good approximation of the expected relative distance. First, the sample size is assumed to be large. Corrections for small samples exist but must be derived for every model type. Second, the candidate models are assumed to be true. This is a result of the derivation of AIC; the AIC penalty (two times the number of parameters) is a property of the true distribution. For untrue models, the penalty is biased but has zero variance. "Other, less biased, estimates for the same quantity exist, but their variances must also be larger. Thus, the constant estimate used in [AIC], besides being trivial to calculate, is likely to have a lower mean squared error than

alternatives in many models in which its assumptions are at least roughly satisfied" (Kuha, 2004, p. 208).

Vaida and Blanchard (2005) demonstrate that, for linear mixed effects models, "marginal" AIC (as in Equation 2.15) assumes that (hypothetical) new datasets would entail a different set of clusters than the original data. Also in the context of linear mixed effects models, Greven and Kneib (2010) show that marginal AIC is not asymptotically unbiased, favoring models with fewer random effects, but suggest this may not be a problem in choosing between models that merely have differing fixed effects. In addition, Vaida and Blanchard (2005) develop a conditional AIC for inferences pertaining to new datasets that would have the same, fixed set of clusters, and this work has been extended by others (Liang, Wu, & Zou, 2008, 3; Greven & Kneib, 2010; Yu & Yau, 2012; Yu, Zhang, & Yau, 2013; Saefken, Kneib, van Waveren, & Greven, 2014). However, conditional AIC is not suitable for this application.

For linear regression models, a corrected form of AIC is available which adjusts for the finite sample by modifying the penalty (Sugiura, 1978; Hurvich & Tsai, 1989). I apply this corrected AIC to the LLTM-E2S:

$$\text{AIC}_c = \text{dev}(y|\hat{\omega}_m(y)) + 2(q_m + 1)\frac{I}{I - q_m - 2}. \tag{2.17}$$

The appropriateness of the corrected AIC here is uncertain, as the second stage LLTM-E2S model includes the estimated variances for $\hat{\delta}$ and as such is not a simple linear regression model. In fact, Kuha (2004, p. 208) notes that a disadvantage of corrected AIC is that the particular adjustments will differ across types of models. Nonetheless, the corrected AIC is expected to be more accurate than standard AIC and so will be used in the simulation study. Corrected AIC has been used with meta-regression in other contexts (see for example, Knowles, Nakagawa, & Sheldon, 2009; Jones, Nakagawa, & Sheldon, 2009; Chen, Ibrahim, Shah, Lin, & Yao, 2012). Also, Vaida and Blanchard (2005) provide a similar correction for linear mixed effects models, but this correction is not used with the LLTM because the correction may not apply to logistic models and also because it would have almost no impact, as the sample size under the LLTM is large ($I \times P$).

As the deviance for the LLTM is marginal only over persons and not over items, or in other words treats persons but not items as random, AIC is expected to perform similarly to holdout validation with the same items for the LLTM. By extension, it is unlikely to be effective for selecting models for item prediction. For the LLTM-E2S, AIC_c is expected to perform similarly to holdout validation with new items, given that the LLTM-E deviance is marginal in regards to items. Results for LOCO-CV with the LLTM-E2S are expected to more closely match those of holdout validation with new items than AIC_c, given that it does not rely on assumptions regarding the appropriate penalty.

BIC and likelihood ratio testing

For completeness, two other model selection strategies are considered despite the fact that they are not motivated by prediction. First, the Bayesian Information Criterion (BIC;

Schwarz, 1978) is

$$\text{BIC}_m = \text{dev}(y|\hat{\omega}_m(y)) + q_m \log N, \tag{2.18}$$

where N is the count of observations. For the LLTM approach $N = I \times P$, while for the two stage approach $N = I$. The penalty for BIC is based on an approximation to the Bayes factor for an assumed multivariate normal prior distribution with means equal to the parameter estimates and a covariance matrix that is as informative as one observation (Kuha, 2004, p. 196). The model with the lowest BIC_m is preferred.

Second, the likelihood ratio test is based on hypothesis testing and is suitable only for comparing nested models. Let Δ_{dev} be the difference in deviance between two models, and let Δ_q be the difference in the number of parameters. Then the asymptotic null distribution for Δ_{dev} is $\chi^2(\Delta_q)$, and most often a p-value less than .05 is deemed statistically significant. If the likelihood ratio test provides a statistically significant result, the simpler model is rejected in favor of the more complex one. When multiple comparisons are needed, the comparisons may be made in ordered pairs. For example, the simplest model may be compared against the second simplest, and if the likelihood ratio test rejects the simplest model, then the second simplest is then compared against the third simplest, and so on. As such, the researcher selects the simplest unrejected model in the end.

2.4 Simulation

Simulation study design

In each replication of the simulation, the LLTM-E is used to generate a training subset, a same items validation subset, a new items validation subset, and an evaluation subset. The two forms of holdout validation (using the same items versus new items) are performed for competing models using both the LLTM and LLTM-E2S. In this way the simulation is a 2×2 (holdout validation type by modeling strategy) design. The simulation replications track which model is selected and the score function values, $\text{dev}(y^{\text{e}}|\hat{\omega}_m(y^{\text{t}}))$, for the competing models. In addition, model selection is also preformed using AIC, BIC, LOCO-CV, and likelihood ratio testing using only the training subset.

The fixed part of the data generating model for the item difficulties is

$$x_i'\beta \equiv \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{2i} x_{3i}, \tag{2.19}$$

which includes constant $x_{1i} = 1$, covariates x_{2i}, x_{3i}, x_{4i}, and one interaction $x_{2i}x_{3i}$. The predictors $x_{2i} \cdots x_{4i}$ are independent draws from a standard normal distribution. A key feature of the generated datasets (and data of this type more generally) is the extent to which the item covariates account for the item difficulties. Let υ^2 represent the variance of the structural part of item difficulty ($x_i'\beta$). Then

$$R^2 = \frac{\upsilon^2}{\upsilon^2 + \tau^2} \tag{2.20}$$

Branch	R^2	I	P	β_1	$\beta_2 \cdots \beta_5$	τ	υ	σ
1	0.3	32	500	0.00	0.41	1.25	0.82	1.50
	0.6	32	500	0.00	0.58	0.95	1.16	1.50
	0.9	32	500	0.00	0.71	0.47	1.42	1.50
2	0.6	16	500	0.00	0.58	0.95	1.16	1.50
	0.6	32	500	0.00	0.58	0.95	1.16	1.50
	0.6	64	500	0.00	0.58	0.95	1.16	1.50

Table 2.1: Generating values for parameters across the simulation conditions. In the first simulation branch, the proportion of explained item variance (R^2) is varied while the number of items (I) is fixed. In the second branch, I is varied while R^2 is fixed. The condition in which $R^2 = .6$ and $I = 32$ is duplicated in the two branches.

represents the proportion of item variance accounted for by the item predictors. For the simulation, let the total item variance be $\tau^2 + \upsilon^2 = 1.25$, and let each of $\beta_2 \cdots \beta_5$ equal the same value. Then for a given value of R^2, the variance of the structural part of item difficulty is $\upsilon = \sqrt{1.25R^2}$. Given the above, $\tau = \sqrt{1.25 - \upsilon^2}$ and $\beta_2 \cdots \beta_5 = \frac{\upsilon}{2}$. The remaining parameters are the intercept, which is set to $\beta_1 = 0$, and the person variance, which is set to $\sigma^2 = 1.25$. In short, the simulation is designed such that the proportion of item variance accounted for by the predictors, R^2, may be varied while maintaining the same the same total item variance.

In the first of two simulation branches, values for R^2 are manipulated, $R^2 \in \{.3, .6, .9\}$, while the number items $I = 32$ is held constant. These generating values are provided in the first part of Table 2.1. In the second simulation branch, the number of items is varied, $I \in \{16, 32, 64\}$, while R^2 is held at .6, and these generating values are depicted in the second part of Table 2.1. The condition in which $R^2 = .6$ and $I = 32$ is duplicated in the two branches, and so there are really five conditions rather than six. In all conditions in both branches, the number of persons is $P = 500$. A total of five hundred replications are carried out for each condition.

Three competing models are subjected to the various model selection strategies. Model 1 includes the main effects only:

$$x_i'\beta \equiv \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}. \tag{2.21}$$

Model 2 matches the data generating model in terms of $x_i'\beta$. Model 3 includes an extra interaction:

$$x_i'\beta \equiv \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{2i} x_{3i} + \beta_6 x_{2i} x_{4i}. \tag{2.22}$$

Using the LLTM-E2s with Model 2 matches the data generating model, while using the LLTM with Model 2 does not because the LLTM does not include the item residual.

Parameter recovery and standard error estimates

Parameter recovery is investigated for both the LLTM and LLTM-E2S. Results for the LLTM-E2S are of interest as confirmation that the method works, while results for the LLTM are of interest because the misspecification of the LLTM may lead to biased parameter estimates. This bias is assessed by the mean difference between the estimated and generating parameters across 500 simulation replications. For this purpose, I focus on estimation results $\hat{\omega}(y^{\mathrm{t}})$ for Model 2 because the fixed part matches that of the generating model. Figure 2.1 presents estimates of bias (the mean of differences) with 95% confidence intervals ($\pm 1.96 \frac{\mathrm{sd}}{\sqrt{500}}$).

For the LLTM, there is evidence of downward bias in the coefficient estimates ($\hat{\beta}_2 \cdots \hat{\beta}_5$) in all of the simulation conditions, and in relation to the magnitude of these coefficients ($\beta_2 \cdots \beta_5 = .58$), the bias is often substantial. However, no such problem is seen for the estimated intercept ($\hat{\beta}_1$). Because of the absence of item residuals, the LLTM is like a population-average model in regards to the items, which is known to exhibit attenuated coefficients for the logistic case (Ritz & Spiegelman, 2004). The estimate of the person standard deviation ($\hat{\sigma}$) also exhibits a downward bias that depends on R^2 (or by extension, τ). For the LLTM-E2S, there is no systematic evidence for bias in $\hat{\beta}$, though there is a downward bias in $\hat{\tau}$ that is mitigated in the high information conditions. The bias in $\hat{\tau}$ may be alleviated by using restricted maximum likelihood estimation or more recent estimators (see for example, Viechtbauer, 2005), but for simplicity and speed, maximum likelihood estimation is used in the simulation.

The LLTM and LLTM-E2S provide very different standard error estimates. To illustrate, I focus on β_6 for Model 3. Because $\beta_6 = 0$ in data generation, $\frac{\hat{\beta}_6}{\mathrm{se}(\hat{\beta}_6)}$ should follow a standard normal distribution across simulation iterations if the standard error estimates are correct. Figure 2.2 presents Q-Q plots of the observed $\frac{\hat{\beta}_6}{\mathrm{se}(\hat{\beta}_6)}$ against the quantiles of the standard normal distribution. In all conditions, values for $\frac{\hat{\beta}_6}{\mathrm{se}(\hat{\beta}_6)}$ for the LLTM deviate greatly from the expected results from a standard normal distribution and indicate that the estimated standard errors are too small. In contrast, the LLTM-E2S shows no such problem, conforming to the appropriate distribution.

Clearly the LLTM yields inappropriately small standard errors, which is a result of the omission of item residuals. For example, the mean standard error for $\hat{\beta}_6$ was 0.03 for the LLTM but 0.20 for the LLTM-E2S in the simulation condition in which $R^2 = .6$ and $I = 32$. As mentioned earlier, Fischer (1997, p. 232) recommended conducting a likelihood ratio test comparing the LLTM and Rasch model as a goodness of fit test. Failure to reject the LLTM would imply that τ is about zero, and then it may be that the LLTM would yield approximately correct standard errors. However, this is an improbable scenario as it requires perfect predictors for item difficulty.

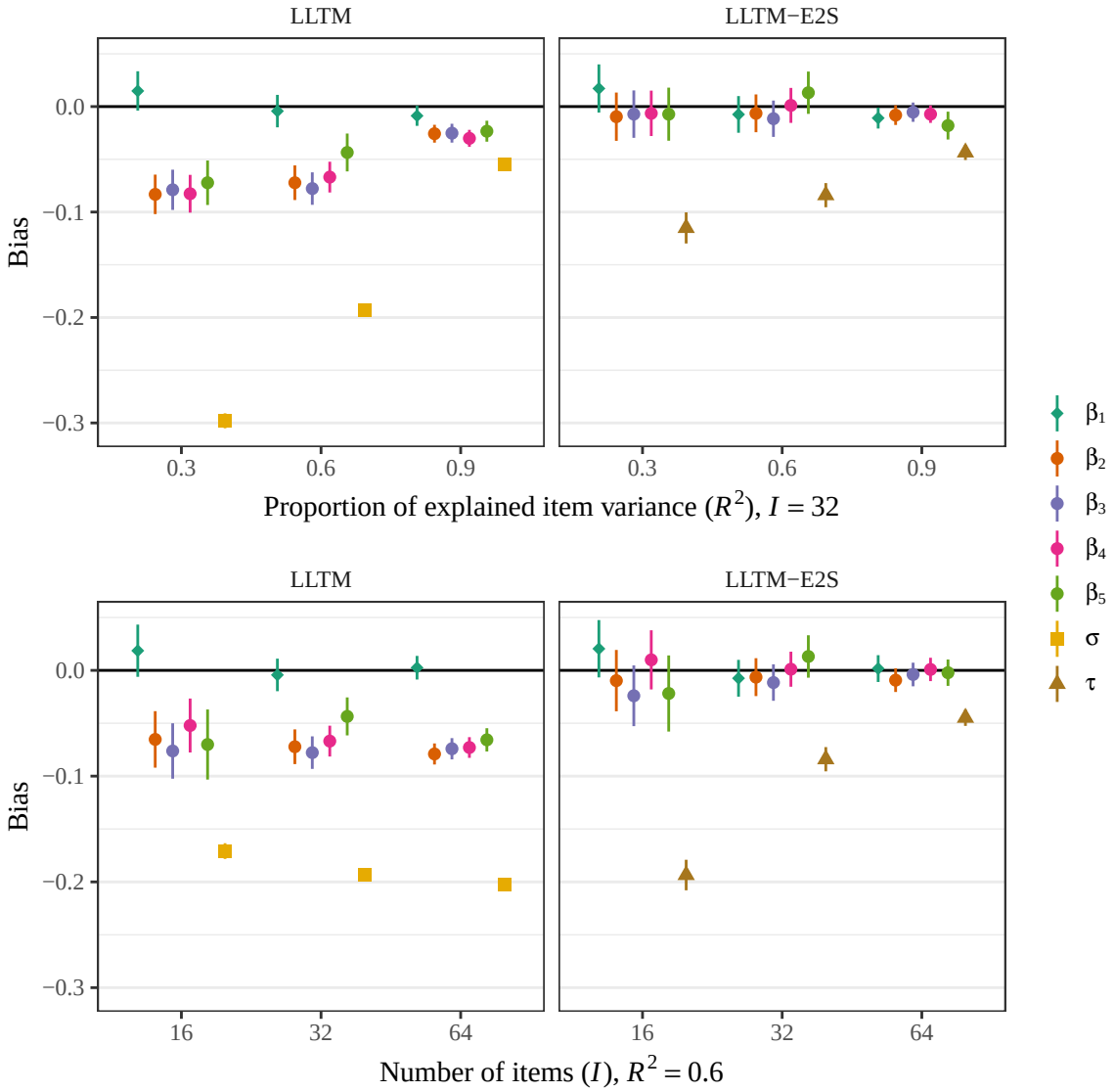


Figure 2.1: Bias (mean of parameter estimates minus generating values) with 95% confidence intervals for Model 2. Results are for 500 simulation replications per condition. The LLTM does not include estimation of τ, and the LLTM-E2S does not include estimation of σ in the second stage.

Model selection

In describing model selection results, it is useful to consider the relative amount of information about the regression coefficients contained in the simulated datasets. Larger values for R^2 are associated with item covariates that are strong predictors, so $R^2 = .9$ is a "high

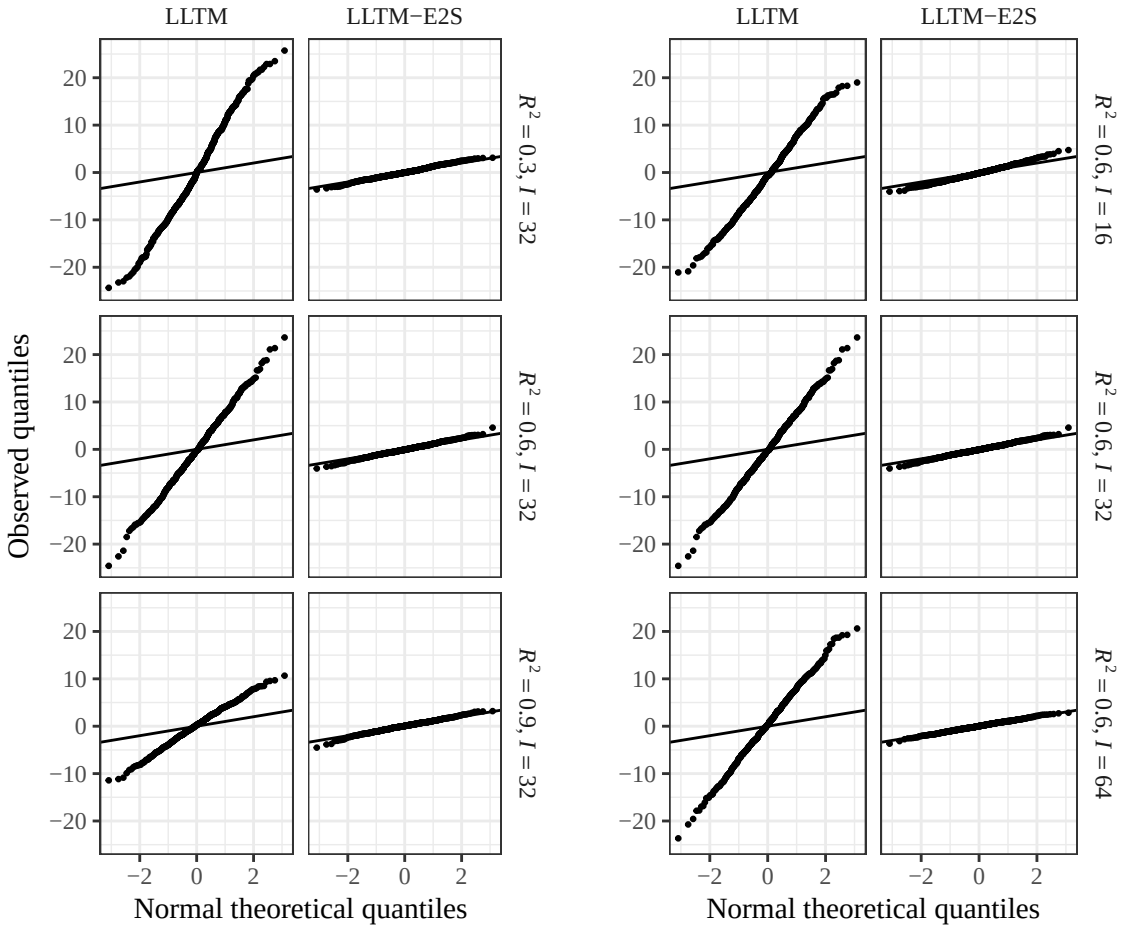


Figure 2.2: Q-Q plots for the observed $\frac{\hat{\beta}_6}{\text{se}(\hat{\beta}_5)}$ for Model 3 across simulation iterations versus standard normal quantiles. On the left are results for simulation replications in which τ varies, and on the right are results for replications in which the number of items vary. Because $\beta_6 = 0$ in data generation, $\frac{\hat{\beta}_6}{\text{se}(\hat{\beta}_6)}$ should follow a standard normal distribution across simulation iterations if the standard errors are correct. The lines have an intercept of zero and slope of one, indicating where the points should lay if the standard error estimates are correct.

information" condition. Also, when the number of items is large, more precise estimates of β are possible, so $I = 64$ is also a high information condition. Conversely, $R^2 = .3$ and $I = 16$ are "low information" conditions. It is expected that Model 2 should provide the best predictions given that it matches the generating model, though there is no guarantee of this with finite data. In high information conditions, it is expected that Model 2 will tend to be selected using holdout validation, LOCO-CV, or AIC because enough information should

be available to obtain good parameter estimates. In low information conditions, the simpler Model 1 may be selected more often, as estimates for its fewer parameters may be more stable compared to those in larger models. This is a matter of bias-variance trade off (Hastie et al., 2009, p. 223), and the expected difference in selection between high and low information conditions is correct behavior as the goal is to identify the most predictive model rather than a true model. Lastly, there is no condition in which Model 3, the most complex model, should tend to be favored, though random variation in the generated datasets are expected to lead to it being selected some small number of times.

Figure 2.3 provides the model selection results for the simulation conditions in which R^2 is varied, and Figure 2.4 shows the same for conditions in which the number of items is varied. Holdout validation using new items performs similarly between the LLTM (left side of both figures) and the LLTM-E2S (right side). In addition, this method selects Model 2 the majority of times in all conditions, though Model 1 is selected with increasing frequency in low information conditions. Holdout validation using the same items also behaves similarly between the LLTM and LLTM-E2S, but for this approach Model 3 is chosen the large majority of times in all simulation conditions. This result is problematic but not surprising; idiosyncrasies that arise from a particular set of item residuals in the training subset are repeated again in the validation subset, and so the two subsets are very similar. In this way the overly complex Model 3 is provided an opportunity to capitalize on chance. Clearly, holdout validation for item prediction must feature datasets with different sets of items in order to be effective.

Focusing now on the LLTM, AIC performs similarly to holdout validation with the same items in terms of model selection (Figures 2.3 and 2.4 again) and quite differently from holdout validation with new items. This supports the argument that using AIC with the deviance from the LLTM corresponds to an inference involving the same set of items. The likelihood ratio test is more conservative, that is, tends to prefer simpler models, when compared to AIC. Assuming an alpha level of .05, a model would have to have a deviance 3.8 lower than a competing model with one fewer parameter in order to reject the simpler model, compared to a difference in deviance of 2 for AIC. BIC is more conservative still with a penalty ranging from 8.99 to 10.37 per parameter, depending on the number of items. The relative conservatism of the likelihood ratio test and BIC are noticeable in the two figures. However, AIC, BIC, and the likelihood ratio test for the LLTM all bear resemblance to holdout validation with the same items and too frequently select the overly complex Model 3.

In contrast, AIC_c paired with the LLTM-E2S performs similarly to holdout validation with new items in terms of model selection (Figures 2.3 and 2.4 again) and is generally more apt to select Model 2 than holdout validation with new items. Results for LOCO-CV with the LLTM-E2S are similar to those for AIC, as expected. For these data and models, AIC_c applies penalties ranging from 2.66 to 5.22 per parameter; larger penalties occur with greater numbers of parameters and with smaller numbers of items. The likelihood ratio test implies a penalty of 3.8 again (assuming the models differ by one parameter) and BIC implies penalties ranging from 2.77 to 4.16 per parameter, depending on the number of items. The penalties

suggested by AIC_c may be higher or lower than for BIC or the likelihood ratio test. Unlike for the LLTM, AIC_c, BIC, and the likelihood ratio test appear more like holdout validation with new items when paired with the LLTM-E2S, and the three perform as expected in model selection.

Implied penalties associated with holdout validation

Standard AIC features a penalty to $\text{dev}(y^{\text{t}}|\hat{\omega}_m(y^{\text{t}}))$ that is a function of the number of model parameters. Let this be $d_{\text{AIC}} = 2q$. In this context, d_{AIC} is an asymptotic approximation for

$$d_{\text{HV}} = \text{E}_{y^{\text{e}}}\text{E}_{y^{\text{t}}}\left[\text{dev}\left(y^{\text{e}}|\hat{\omega}_m(y^{\text{t}})\right) - \text{dev}\left(y^{\text{e}}|\hat{\omega}_m(y^{\text{e}})\right)\right], \tag{2.23}$$

which is the expected difference between the holdout validation deviance and the deviance obtained from both fitting and evaluating the model using the evaluation subset. In effect, d_{HV} is the correct but unknown penalty. An empirical estimate for it may be obtained from the simulation as

$$\hat{d}_{\text{HV}} = \frac{1}{R}\sum_{r=1}^{R}\left[\text{dev}\left(y^{\text{e},r}|\hat{\omega}_m(y^{\text{t},r})\right) - \text{dev}\left(y^{\text{e},r}|\hat{\omega}_m(y^{\text{e},r})\right)\right], \tag{2.24}$$

where r indexes the R simulation replications. In a large sample, d_{AIC} should approximate d_{HV} well. Also, let d_{AIC_c} be the penalty associated with AIC_c, as in Equation 2.17. This penalty should approximate d_{HV} well in smaller samples. Last, the penalty implied by LOCO-CV may be estimated as

$$\hat{d}_{\text{CV}} = \frac{1}{R}\sum_{r=1}^{R}\left[\left[\sum_{i=1}^{I}\text{dev}\left(y_i^{\text{t},r}|\hat{\omega}_m(y_{-i}^{\text{t},r})\right)\right] - \text{dev}\left(y^{\text{t},r}|\hat{\omega}_m(y^{\text{t},r})\right)\right] \tag{2.25}$$

which should also approximate d_{HV} well when the sample is large.

Figure 2.5 displays the different estimated and calculated penalties across simulation conditions. For the LLTM, the $\hat{d}_{\text{HV}}$ is much greater than d_{AIC} across all simulation conditions. Clearly the penalty implied by AIC is incorrect for the LLTM, given that that d_{AIC} and $\hat{d}_{\text{HV}}$ bear no resemblance. Interestingly, $\hat{d}_{\text{HV}}$ for the LLTM appears to depend on R^2, and if R^2 were near one (making the LLTM the correct model), it may be that $\hat{d}_{\text{HV}}$ would approximately equal d_{AIC} for the LLTM. It is clear that $\hat{d}_{\text{HV}}$ depends on R^2, but AIC_c for the LLTM-E2S fails to capture this phenomenon. When $R^2 = .9$ or $I = 16$, AIC_c differs substantially from $\hat{d}_{\text{HV}}$, but otherwise AIC_c appears to reasonably approximate $\hat{d}_{\text{HV}}$. On the other hand, $\hat{d}_{\text{CV}}$ is quite close to $\hat{d}_{\text{HV}}$, though less so when $I = 16$.

Comparison of predictive performance

The root mean squared error of prediction for model m is

$$\text{RMSEP}_m = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left[x_i^{\text{e}\prime}\hat{\beta}_m(x^{\text{t}}, y^{\text{t}}) - \delta_i^{\text{e}}\right]^2}, \tag{2.26}$$

where $x_i^{e\prime}$ is a vector of item predictors associated with the evaluation data, $\hat{\beta}_m(x^t, y^t)$ is a vector of coefficients for model m estimated on the training subset, and δ_i^e is a known item difficulty (fixed plus residual parts) associated with the evaluation subset. Here the item predictors x are brought into the notation to emphasize that the coefficients are trained with x^t (and y^t), but predictions for δ_i^e are made with x^e. The RMSEP_m is based on the difference between predicted and actual item difficulties for new data. In a best case scenario, $x_i^{e\prime}\hat{\beta}_m(x^t, y^t)$ may fully account for the fixed part of δ_i^e, but it cannot account for the residual part. In this way, the residual standard deviation (τ) is the best value that may be expected for RMSEP_m. As RMSEP_m relies on known δ_i^e, it is only available in a simulation context, though an alternative could be based on estimates $\hat{\delta}_i^e$ from the Rasch model.

Figure 2.6 presents the mean RMSEP_m for each model, and τ is indicated by the dashed lines. In general, Models 2 and 3 both produce predictions that are close to τ, while Model 1 performs more poorly. The exceptions are the low information conditions, in which case Models 1 and 3 perform similarly. Further, let RMSEP_{m^*} be the root mean squared error of prediction for the model chosen by a given selection strategy. Then Figure 2.7 presents the mean of RMSEP_{m^*} across simulation replications for several selection strategies. The selection criteria and and whether the LLTM or LLTM-E2s are used makes little difference in RMSEP_{m^*}. Part of the reason for the similarity is that the differences between competing models in regards to RMSEP_m is small.

2.5 Discussion

Some recommendations may be made based on the simulation study. First, the LLTM should not be used because it yields biased parameter estimates and incorrect standard error estimates when the true model has an error term for item difficulty. Further, model selection strategies like AIC, BIC, and the likelihood ratio test behave inappropriately with the LLTM. The preceding findings occurred even when the unrealistic assumptions of the LLTM were approximately met ($R^2 = .9$). Instead, the LLTM-E2S is recommended for the accuracy of its parameter and standard error estimates. It does exhibit biased estimates for τ with marginal maximum likelihood estimation, but if this is concerning then alternative estimators, such as REML, could be considered. Second, LOCO-CV is recommended for model selection over AIC or AIC_c with the LLTM-E2S. The simulation indicated that correct penalties for the LLTM-E2S depend on R^2, or by extension, τ, and neither AIC nor AIC_c address this phenomenon.

Seemingly contradictory findings arose from the simulation study; choices of modeling strategies and selection strategies had clear implications for model selection, but this did not lead to substantial differences in predictive accuracy. Though Model 2 made the best predictions on average across all conditions, it often performed only slightly better than Model 3. As the simulation demonstrated, competing models may have similar predictive utility, and furthermore predictive accuracy is limited by the residual standard deviation of item difficulty.

If the purpose of model selection is to choose a best model for generating predictions (with the specific parameter estimates from the available data), then holdout validation with new items is the recommended strategy as holdout validation is the only approach that is conditional on the parameter estimates. In this way, inferences regarding predictive utility are based on the specific prediction parameter estimates that would be used. In contrast, single dataset approximations like AIC and LOCO-CV rely on expectations over hypothetical data. This may still be useful if the goal is to merely identify a preferred model with expected predictive accuracy as a benchmark. In particular, LOCO-CV is recommended as it is less reliant on assumptions regarding the penalty term.

A stronger method of prediction would combine predictions from several models. The super learner (van der Laan, Polley, & Hubbard, 2007) accomplishes this by assigning weights to the predictions from the candidate models, and asympotically such predictions are as good as those from the candidate model that minimizes the loss function with respect to the true probability distribution. However, the application of the super learner to the kind of data considered in this chapter is not straightforward. The super learner has been applied to propensity scores (Pirracchio, Petersen, & van der Laan, 2015), which like item response data involve a binary outcome variable, but the super learner would also need to be extended to clustered data.

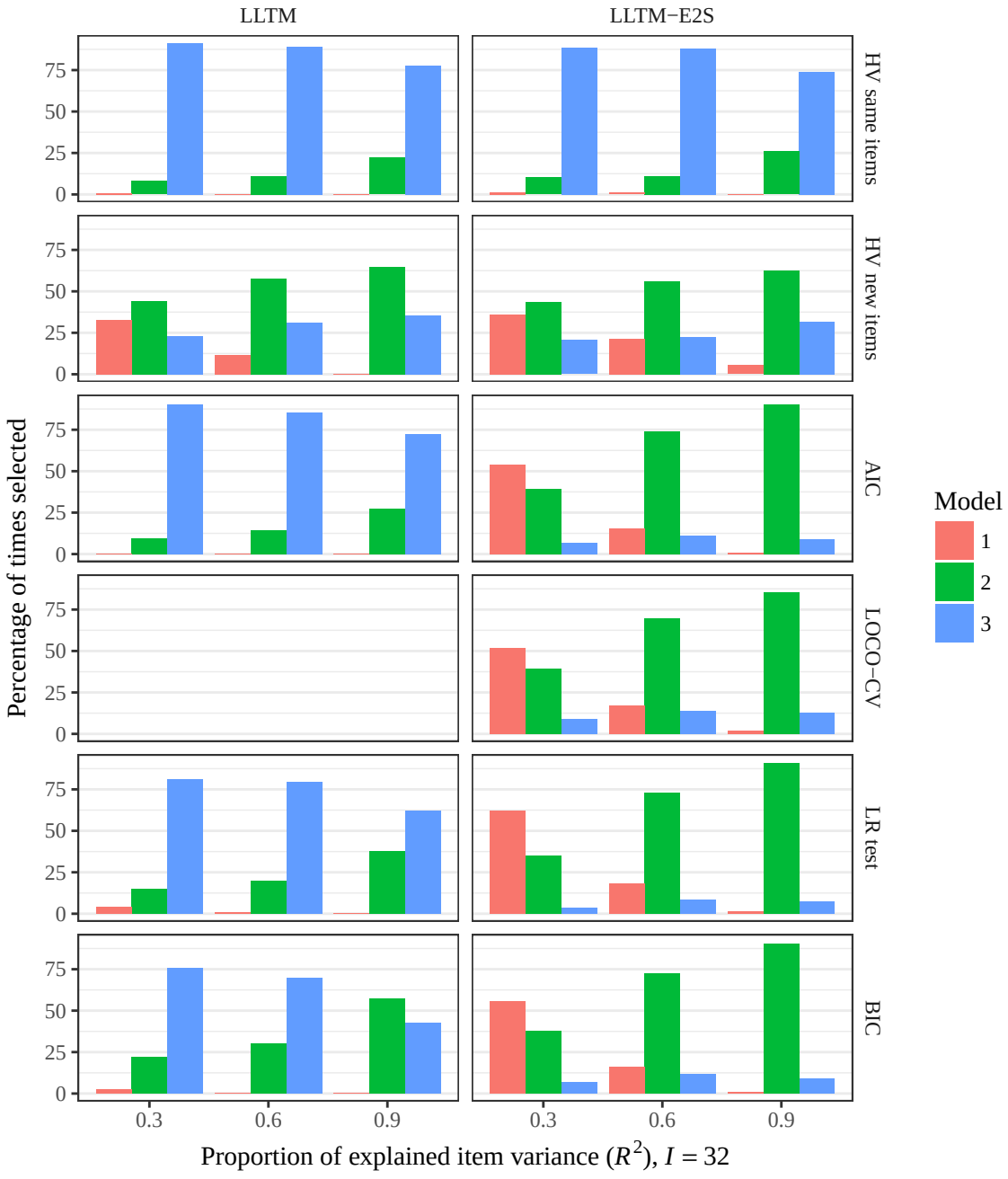


Figure 2.3: Percentages of times each model was chosen by the various selection criteria for simulation replications in which the proportion of explained item variance (R^2) is varied. Standard AIC is used for the LLTM, while AIC_c is used for the LLTM-E2S. LOCO-CV was applied only to the LLTM-E2S.

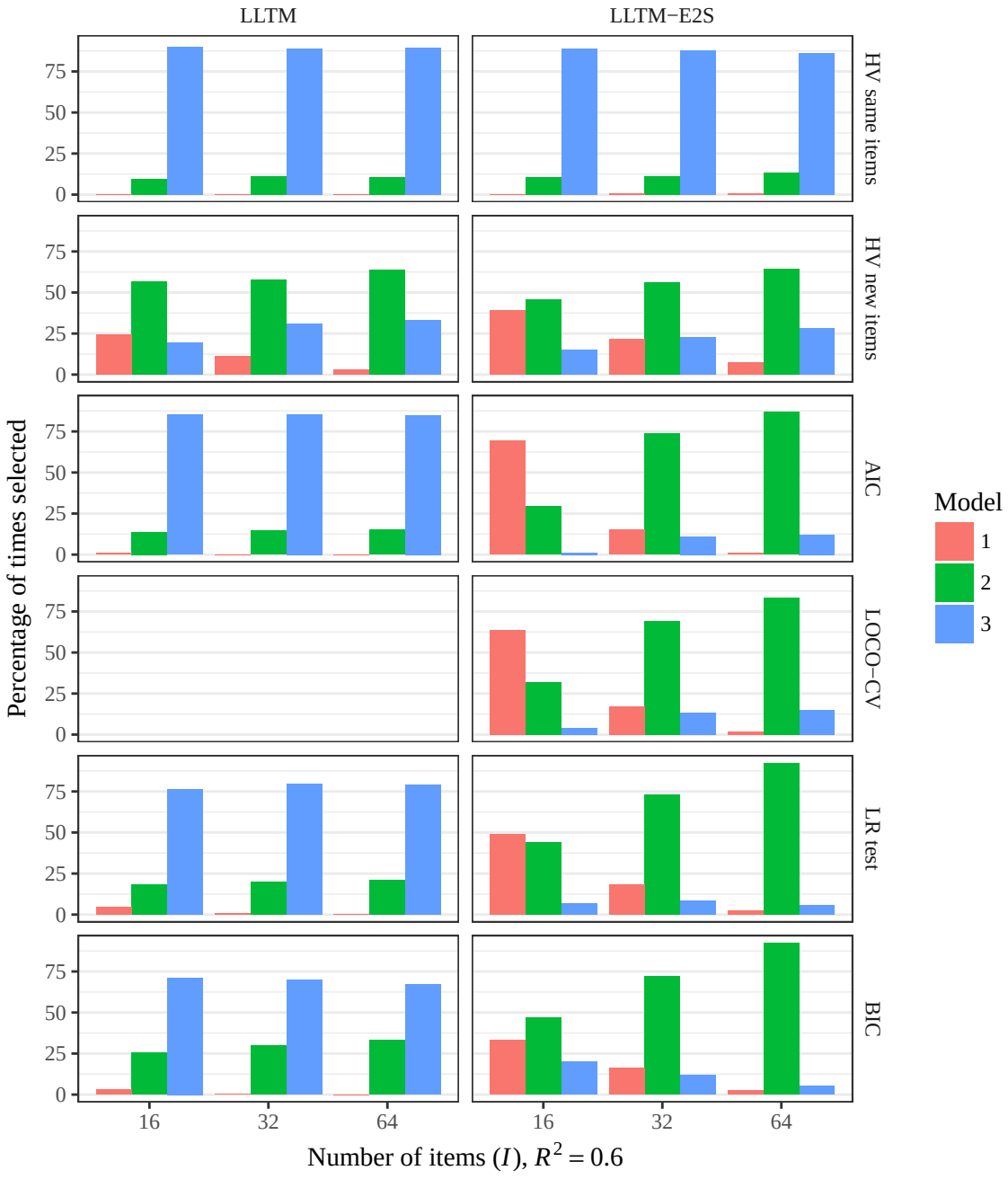


Figure 2.4: Percentages of times each model was chosen by the various selection criteria for simulation replications in which the number of items is varied. Standard AIC is used for the LLTM, while AIC_c is used for the LLTM-E2S. LOCO-CV was applied only to the LLTM-E2S.

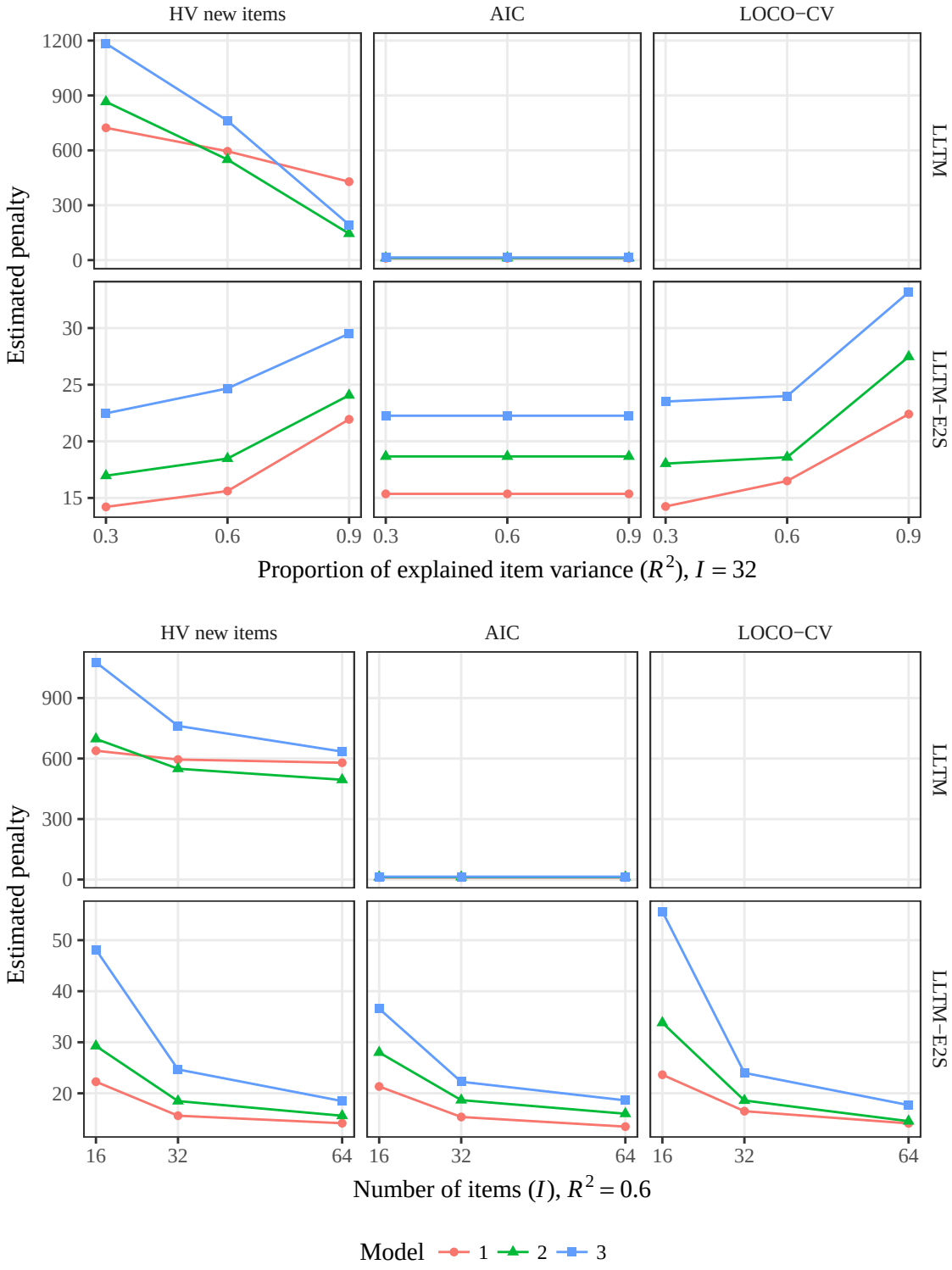


Figure 2.5: Estimated penalties implied by holdout validation with new items ($\hat{d}_{\text{HV}}$), AIC (d_{AIC} or d_{AIC_c}), and LOCO-CV ($\hat{d}_{\text{CV}}$) for the LLTM and LLTM-E2S. The y-axes vary. Standard AIC is used for the LLTM, while AIC_c is used for the LLTM-E2S. LOCO-CV was not performed with the LLTM.

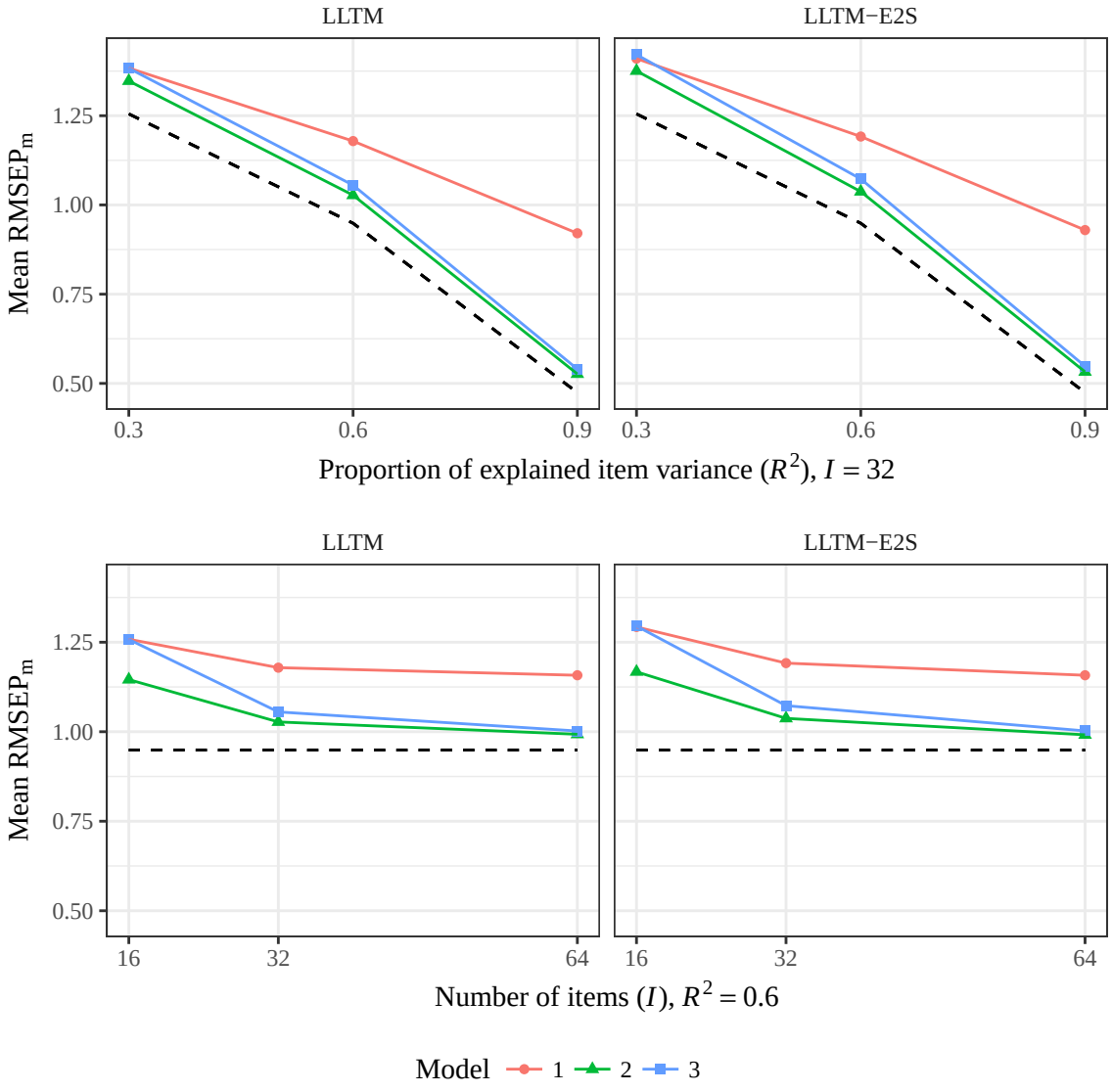


Figure 2.6: Mean for the root mean squared error of prediction (RMSEP$_m$) for each model across simulation conditions. The dashed line represents the residual item standard deviation (τ), which is the limit of prediction accuracy.

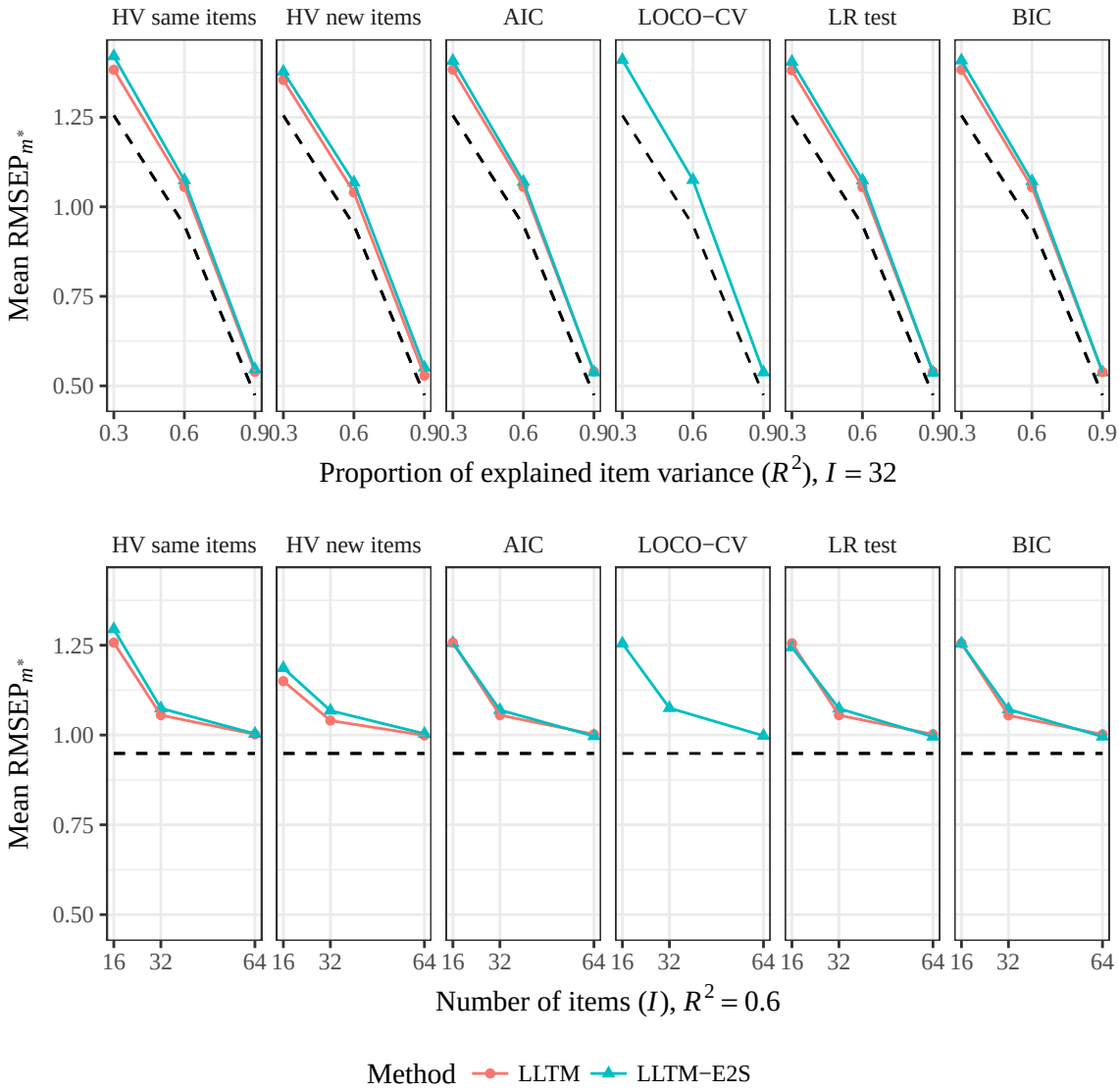


Figure 2.7: The mean for the mean root mean squared error of prediction for the selected model (RMSE_{m*}) across simulation replications for the various selection strategies. The dashed line represents the residual item standard deviation (τ), which is the limit of prediction accuracy. Standard AIC is used for the LLTM, while AIC_c is used for the LLTM-E2S. LOCO-CV was applied only to the LLTM-E2S.

Chapter 3

Bayesian approaches to cross-validation for person-explanatory models

3.1 Introduction

Model comparison using indices motivated by cross-validation requires consideration of how new observations would arise. For the case of independent observations, new observations come about in a straightforward way, that being simply a new collection of exchangeable units. For the case of clustered observations, new observations may come from within the existing clusters or instead from within new clusters. Spiegelhalter, Best, Carlin, and van der Linde (2002) described this distinction as the focus of the model. A model may be focused on the direct parameters associated with clusters, often referred to as latent variables. Focus on the direct parameters implies that a new data collection would entail a new sample of units from the existing set of clusters. Alternatively, the focus of the model may be the hyperparameters for the distribution of the direct parameters. In this case, the implied new data collection involves a new sample of clusters, which of course also provide previously unobserved units.

A researcher may be interested in obtaining estimates of the out-of-sample prediction accuracy, either to assess a single model or to compare several models. The deviance information criterion (DIC; Spiegelhalter et al., 2002), widely applicable information criterion (WAIC; Watanabe, 2010), and Pareto-smoothed importance sampling estimates of leave-one-out cross-validation (PSIS-LOO; Vehtari & Gelman, 2016) are methods of obtaining such estimates. Each of these depend on evaluating the likelihood given posterior draws from Markov chain Monte Carlo (MCMC) simulation. The usual way of specifying Bayesian models in programs like BUGS, JAGS, or Stan bases the likelihood on the direct parameters, and so the "default" implementation of information criteria results in an inference with the focus being on the direct parameters. As such, the estimates of predictive accuracy are for

predictions for new units arising from the existing set of clusters.

For assessing predictive accuracy with a focus on the hyperparameters, or in other words for predictions involving a new sample of clusters, it is necessary to obtain cluster-level likelihoods that are *marginal* over the direct parameters. Such likelihoods are common in the frequentist tradition; generalized structural equation models and generalized linear mixed models are usually fit using marginal maximum likelihood estimation, for example. In this chapter, I advocate for obtaining posterior draws of *marginal* likelihoods after the usual MCMC simulation for use with DIC, WAIC, or PSIS-LOO when a prediction inference involving new clusters in needed.

3.2 A simple hierarchical Bayesian model

The posterior for a simple hierarchical Bayesian model is

$$p(\omega, \psi, \zeta|y) \propto \prod_{j=1}^{J} \prod_{i=1}^{I} p(y_{ij}|\omega, \zeta_j) \prod_{j=1}^{J} p(\zeta_j|\psi) p(\omega, \psi), \tag{3.1}$$

where $y_{ij} \in y$ is the response for observation i $(i = 1 \dots I)$ in cluster j $(j = 1 \dots J)$, ω is a vector of parameters common across clusters, ζ_j is a cluster-specific parameter, and ψ is a hyperparameter for the prior distribution of ζ_j. Further, let ζ represent the vector containing all ζ_j. The likelihood $p(y_{ij}|\omega, \zeta_j)$ is conditional on ζ_j and therefore does not directly involve ψ. The prior distributions include the hierarchical prior for ζ and the joint prior for ω and ψ, and the latter may be rewritten as $p(\omega)p(\psi)$ if independent priors are assigned.

For a focus on $\{\omega, \psi\}$, a marginal, cluster-level likelihood may be obtained in order to make predictive inferences regarding data from new clusters. Otherwise, for a focus on $\{\omega, \zeta\}$, the usual unit-level likelihood may be used for prediction inferences for new data from the original clusters. I will refer to this likelihood as the "conditional" likelihood as it conditions on ζ.

Any hierarchical prior distribution may be assumed for ζ_j, and depending on the type of distribution, ψ may be a vector representing multiple parameters associated with the chosen distribution. Further, ζ_j may be a vector of several cluster-specific parameters. However, the adaptive quadrature method proposed in Section 3.4 requires a scalar ζ_j having a normal prior, $\zeta_j \sim \mathrm{N}(0, \psi^2)$, where ψ is a standard deviation. Here, ζ_j may be thought of as a residual, and then ω determines the mean structure of y while ψ represents the standard deviation of the residuals. As an aside, the proposed adaptive quadrature method could be generalized to a multivariate normal distribution to accommodate ζ_j being a vector.

3.3 Information criteria for hierarchical Bayesian models

In this section, the notation of Vehtari, Gelman, and Gabry (2016) regarding out-of-sample pointwise predictive accuracy is modified for the specific case of hierarchical models. For the conditional approach, predictions are conditional on ζ_j, while for the marginal approach ζ_j is integrated out to obtain predictions. The notation of Vehtari et al. (2016) implicitly takes the conditional approach, which is made explicit here by the inclusion of the c subscript on the relevant quantities.

Conditional forms of out-of-sample predictive accuracy

Target quantity for the conditional case

Inference regarding predictive performance of the model for new data relies on the posterior predictive distribution (Rubin, 1984; Marshall & Spiegelhalter, 2007). The posterior predictive distribution for a new observation $\tilde{y}_{i'j}$ is

$$p(\tilde{y}_{i'j}|y) = \iiint p(\tilde{y}_{i'j}|\omega,\zeta_j)p(\zeta_j|\omega,\psi,y)p(\omega,\psi|y)\ \mathrm{d}\omega\mathrm{d}\psi\mathrm{d}\zeta_j, \tag{3.2}$$

where i' is a new unit within a previously existing cluster j and $p(\tilde{y}_{i'j}|\omega,\zeta_j)$ is similar to $p(y_{ij}|\omega,\zeta_j)$ but for unobserved (hypothetical) $\tilde{y}_{i'j}$. Implicitly, $p(\tilde{y}_{i'j}|y)$ in the above also depends on the choice of model, not just the data. The expected log pointwise predictive density for a new dataset, in which new observations arise from the existing clusters, is

$$\mathrm{elpd}_c = \sum_{j=1}^{J}\sum_{i=1}^{I}\int p_t(\tilde{y}_{i'j}) \log p(\tilde{y}_{i'j}|y)\ \mathrm{d}\tilde{y}_{i'j}, \tag{3.3}$$

where p_t is the true (unknown) data generating distribution. The c subscript in elpd_c denotes that it is based on the likelihood conditional on ζ_j. Information criteria and cross-validation are means of approximating the expected log pointwise predictive density.

Conditional widely applicable information criteria

WAIC is a form of Bayesian information criterion that requires only the log pointwise predictive density, which is:

$$\mathrm{lpd}_c = \sum_{j=1}^{J}\sum_{i=1}^{I}\log\iiint p(y_{ij}|\omega,\zeta_j)p(\zeta_j|\omega,\psi,y)p(\omega,\psi|y)\ \mathrm{d}\omega\mathrm{d}\psi\mathrm{d}\zeta_j. \tag{3.4}$$

The lpd_c is the log of the full data likelihood integrated over the posterior for all parameters. An estimate for it is obtained from the draws of MCMC simulation as

$$\widehat{\mathrm{lpd}}_c = \sum_{j=1}^{J}\sum_{i=1}^{I}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_{ij}|\omega^s,\zeta_j^s)\right), \tag{3.5}$$

where ω^s and ζ_j^s are the s-th posterior draw from MCMC simulation, $s = 1 \dots S$. The $\widehat{\text{lpd}}_c$ is a naively optimistic, in-sample estimate of elpd_c, given that the same data are used both to fit the model and obtain $\widehat{\text{lpd}}_c$. Similar to many forms of information criteria, WAIC adds a "penalty" to a naive estimate like $\widehat{\text{lpd}}_c$ to obtain an estimate of expected fit to new data. This penalty, which approximates the "optimism" of lpd_c, may also be referred to as the estimated effective number of parameters. For WAIC, this estimated effective number of parameters is

$$\hat{p}_{\text{WAIC},c} = \sum_{j=1}^{J} \sum_{i=1}^{I} V_{s=1}^{S} \log p(y_{ij}|\omega^s, \zeta_j^s) \tag{3.6}$$

where $V_{s=1}^{S}$ represents the sample variance across the S posterior draws. Then the expected log pointwise predictive density for WAIC is

$$\widehat{\text{elpd}}_{\text{WAIC},c} = \widehat{\text{lpd}}_c - \hat{p}_{\text{WAIC},c}. \tag{3.7}$$

The value normally reported for WAIC is on the deviance scale,

$$\text{WAIC}_c = -2\widehat{\text{elpd}}_{\text{WAIC},c}. \tag{3.8}$$

WAIC is asymptotically equal to Bayesian cross-validation with the deviance as the loss function (Vehtari et al., 2016, p. 2). It is unreliable when the variance of $\log p(y_{ij}|\omega^s, \zeta_j^s)$ (from Equation 3.6) exceeds .4 for a given observation (Vehtari et al., 2016, p. 11). The penalty in WAIC may be viewed as an approximation to the number of unconstrained parameters (Gelman, Hwang, & Vehtari, 2014, p. 1003). An alternative formula for the WAIC effective number of parameters based on the mean of $p(y_{ij}|\omega^s, \zeta_j^s)$ is available, but is less numerically stable (Gelman, Hwang, & Vehtari, 2014, p. 1002).

Conditional approximate leave-one-out cross-validation

The Bayesian leave-one-out expected log pointwise predictive density is

$$\text{elpd}_{\text{LOO},c} = \sum_{j=1}^{J} \sum_{i=1}^{I} \log p(y_{ij}|y_{-i,j}) \tag{3.9}$$

where $y_{-i,j}$ is all observations except for the i-th observation in cluster j and

$$p(y_{ij}|y_{-i,j}) = \iiint p(y_{ij}|\omega, \zeta_j) p(\zeta_j|\psi, y_{-i,j}) p(\omega, \psi|y_{-i,j}) \, \mathrm{d}\omega \mathrm{d}\psi \mathrm{d}\zeta_j, \tag{3.10}$$

which bears resemblance to the posterior predictive distribution for a new response. An estimate of $\text{elpd}_{\text{LOO},c}$ may be obtained from the posterior draws as

$$\widehat{\text{elpd}}_{\text{PSIS-LOO},c} = \sum_{j=1}^{J} \sum_{i=1}^{I} \log \left(\frac{\sum_{s=1}^{S} w_{ij}^s p(y_{ij}|\omega^s, \zeta_j^s)}{\sum_{s=1}^{S} w_{ij}^s} \right), \tag{3.11}$$

where w_{ij}^s is a weight specific to the observation and posterior draw. Vehtari and Gelman (2016) introduced Pareto smoothed importance sampling weights, which are calculated separately for every observation as follows: first, raw importance ratios are calculated as $p(y_{ij}|\omega^s, \zeta_j^s)^{-1}$ for each posterior draw; second, the generalized Pareto distribution is fit to the 20% largest raw importance ratios; third, the 20% largest raw importance ratios are replaced by the expected values of the order statistics of the fitted generalized Pareto distribution; and fourth, the weights are truncated at $S^{\frac{3}{4}}\bar{w}_{ij}$, where $\bar{w}_{ij}$ is the average of the S smoothed weights. The calculation of the raw importance weights was developed by Gelfand, Dey, and Chang (1992), the truncation of the weights was developed by Ionides (2008), and the Pareto smoothing was developed by Vehtari and Gelman (2016). On the deviance scale,

$$\text{PSIS-LOO}_c = -2\widehat{\text{elpd}}_{\text{PSIS-LOO},c}. \tag{3.12}$$

Because PSIS-LOO_c is calculated from a single fit of a model on one dataset, it may be considered a form of information criterion. Lastly, an estimate of the effective number of parameters associated with PSIS-LOO may be obtained after the fact:

$$\hat{p}_{\text{PSIS-LOO},c} = \widehat{\text{elpd}}_{\text{PSIS-LOO},c} - \widehat{\text{lpd}}_c. \tag{3.13}$$

PSIS-LOO, like WAIC, is an approximation to elpd (Vehtari et al., 2016, p. 3) that may be computed from the lpd. PSIS-LOO is more robust than WAIC in cases with weak priors or influential observations (Vehtari et al., 2016, p. 2), and so it may be used when the requirements related to the penalty term in WAIC are not met. Still, it becomes unreliable when the estimated shape parameter for the generalized Pareto distribution exceeds 1 for a given observation (Vehtari et al., 2016, p. 11).

Conditional deviance information criteria

While conditional WAIC and PSIS-LOO approximate elpd_c, the target quantity for DIC differs. The target is

$$\text{elpd}_{\text{DIC},c} = \sum_{j=1}^{J}\sum_{i=1}^{I}\int p_{\text{t}}(\tilde{y}_{i'j})p(\tilde{y}_{i'j}|\bar{\omega}, \bar{\zeta}_j)\ d\tilde{y}_{i'j}, \tag{3.14}$$

where $\bar{\omega}$ and $\bar{\zeta}_j$ represent the posterior means for their respective parameters. While elpd_c, presented earlier, involves the full posterior predictive distribution $p(\tilde{y}_{i'j}|y)$, $\text{elpd}_{\text{DIC},c}$ instead relies on the predictive distribution $p(\tilde{y}_{i'j}|\bar{\omega}, \bar{\zeta}_j)$, which is based on point estimates. The use of point estimates differentiates DIC from WAIC and PSIS-LOO.

The naive, in-sample equivalent of $\text{elpd}_{\text{DIC},c}$ is the log-likelihood of the observed y evaluated at the posterior mean, which is

$$\text{lpd}^*_{\text{DIC},c} = \sum_{j=1}^{J}\sum_{i=1}^{I}\log p(y_{ij}|\bar{\omega}, \bar{\zeta}_j). \tag{3.15}$$

Estimates for $\bar{\omega}$ and $\bar{\zeta}_j$ are plugged into the above to obtain $\widehat{\text{lpd}}^*_{\text{DIC},c}$, the estimated log-likelihood evaluated at the posterior mean. Obtaining the estimated effective number of parameters requires this quantity as well as the mean of the log-likelihood taken over the posterior, which is

$$\text{lpd}_{\text{DIC},c} = \sum_{j=1}^{J}\sum_{i=1}^{I} \iiint \log p(y_{ij}|\omega,\zeta_j)p(\zeta_j|\omega,\psi,y)p(\omega,\psi|y)\ \text{d}\omega\text{d}\psi\text{d}\zeta_j. \tag{3.16}$$

This $\text{lpd}_{\text{DIC},c}$ differs from lpd_c in that it is the log-likelihood integrated over the posterior rather than the log of the likelihood integrated over the posterior. It is estimated in MCMC simulation as

$$\widehat{\text{lpd}}_{\text{DIC},c} = \sum_{j=1}^{J}\sum_{i=1}^{I}\left(\frac{1}{S}\log\sum_{s=1}^{S} p(y_{ij}|\omega^s,\zeta_j^s)\right), \tag{3.17}$$

where ω^s and ζ_j^s represent parameter values at posterior draw s, $s = 1\cdots S$. Then the estimated effective number of parameters is

$$\hat{p}_{\text{DIC},c} = 2(\widehat{\text{lpd}}^*_{\text{DIC},c} - \widehat{\text{lpd}}_{\text{DIC},c}). \tag{3.18}$$

The approximation for the expected log pointwise predictive density is

$$\widehat{\text{elpd}}_{\text{DIC},c} = \widehat{\text{lpd}}^*_{\text{DIC},c} - \hat{p}_{\text{DIC},c}, \tag{3.19}$$

and the value reported for DIC is usually on the deviance scale:

$$\text{DIC}_c = -2\widehat{\text{elpd}}_{\text{DIC},c}. \tag{3.20}$$

This conditional DIC corresponds to the DIC_7 of Celeux, Forbes, Robert, and Titterington (2006).

The reliance on point estimates in calculating $\widehat{\text{lpd}}^*_c$ results in DIC not being invariant to reparameterization (Spiegelhalter, Best, Carlin, & van der Linde, 2014, p. 4). For example, results for DIC will differ depending on whether a parameter like ψ represents a standard deviation or a variance. Further, owing to the reliance on point estimates, the posterior distribution must be reasonably summarized by its mean (Gelman, Hwang, & Vehtari, 2014, p. 1015), and it is possible for $\hat{p}_{\text{DIC},c}$ to be negative if the posterior mean is far from the mode (Gelman, Carlin, et al., 2014, p. 172). An alternative for $\hat{p}_{\text{DIC},c}$ based on the variance of $p(y_{ij}|\omega^s,\zeta_j^s)$ is guaranteed to be positive, but is less numerically stable (Gelman, Carlin, et al., 2014, p. 173). Lastly, DIC_c will only be a good approximation of $-2\text{elpd}_{\text{DIC},c}$ when $\hat{p}_{\text{DIC},c}$ is much less than the number of units (Plummer, 2008, p. 535), which may not be the case for DIC_c given the cluster-specific parameters.

Marginal forms of out-of-sample predictive accuracy

Target quantity for the marginal case

Marginal likelihoods may be used with information criteria instead of the "default" conditional likelihoods. In this section, marginal equivalents of the conditional quantities in the previous section are described. The marginal likelihood, which integrates out ζ_j, is

$$p(y_j|\omega,\psi) = \int p(\zeta_j|\omega,\psi) \prod_{i=1}^{I} p(y_{ij}|\omega,\zeta_j) \ \mathrm{d}\zeta_j, \tag{3.21}$$

where y_j is the vector of responses for cluster j, and $p(\zeta_j|\psi)$ is the prior distribution of ζ_j given ψ. This prior is not directly influenced by the data, in contrast to the posterior $p(\zeta_j|y_j,\omega,\psi)$. Though the conditional likelihood is normally used to specify the Bayesian model in software for MCMC, the marginal likelihood may be calculated after MCMC simulation, which is the approach taken here.

The predictive distribution for $\tilde{y}_{j'}$, which is a new response vector arising from a new cluster j', is

$$p(\tilde{y}_{j'}|y) = \iint p(\tilde{y}_{j'}|\omega,\psi)p(\omega,\psi|y) \ \mathrm{d}\omega\mathrm{d}\psi, \tag{3.22}$$

where $p(\tilde{y}_{j'}|\omega,\psi)$ is similar to $p(y_j|\omega,\psi)$ (in Equation 3.21) but for unobserved $\tilde{y}_{j'}$. The density $p(\tilde{y}_{j'}|y)$ may be referred to as a mixed predictive distribution (Gelman et al., 1996; Marshall & Spiegelhalter, 2007); it involves the posterior for ω and ψ but the prior $p(\tilde{\zeta}_{j'},\psi)$ for $\tilde{\zeta}_{j'}$. The expected log pointwise predictive density for a new dataset, containing a new sample of clusters, is

$$\mathrm{elpd}_m = \sum_{j=1}^{J} \int p_t(\tilde{y}_{j'}) \log p(\tilde{y}_{j'}|y) \ \mathrm{d}\tilde{y}_{j'}, \tag{3.23}$$

where p_t again is the true data generating distribution. The m subscript indicates that elpd_m results from the marginal likelihood. It is important to note that here the meaning of "point" is redefined to refer to a cluster rather than a single unit within a cluster.

Marginal widely applicable information criteria

Marginal WAIC is calculated in much the same way as the conditional version by substituting $p(y_j|\omega^s,\psi^s)$ for $p(y_{ij}|\omega^s,\zeta_j^s)$ in the calculations and defining the points to be clusters. For completeness, the modified equations are presented. The marginal form for the log pointwise predictive density is

$$\mathrm{lpd}_m = \sum_{j=1}^{J} \log \iint p(y_j|\omega,\psi)p(\omega,\psi|y) \ \mathrm{d}\omega\mathrm{d}\psi \tag{3.24}$$

and may estimated from the draws of MCMC simulation as

$$\widehat{\text{lpd}}_m = \sum_{j=1}^{J} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(y_j|\omega^s, \psi^s) p(\omega^s, \psi^s|y) \right], \tag{3.25}$$

where $p(y_j|\omega^s, \psi^s)$ is similar to $p(y_j|\omega, \psi)$ in Equation 3.21 but for a given posterior sample s. It is expected that lpd_m will be less than lpd_c (Trevisani & Gelfand, 2003). The effective number of parameters for marginal WAIC is

$$\hat{p}_{\text{WAIC},m} = \sum_{j=1}^{J} V_{s=1}^{S} \log p(y_j|\omega^s, \psi^s), \tag{3.26}$$

the expected log pointwise predictive density is

$$\widehat{\text{elpd}}_{\text{WAIC},m} = \widehat{\text{lpd}}_m - \hat{p}_{\text{WAIC},m}, \tag{3.27}$$

and the final value on the deviance scale is

$$\text{WAIC}_m = -2\widehat{\text{elpd}}_{\text{WAIC},m}. \tag{3.28}$$

The integrated WAIC of Li, Qiu, Zhang, and Feng (2016) bears some relation to this marginal WAIC. Integrated WAIC was developed for the case in which there is a vector of direct parameters (ζ_j) for each observation. The data in this case are not clustered. Li et al. (2016) use Monte Carlo sampling to approximate the integration over the ζ_j vector. On one hand, integrated WAIC is less general than the marginal WAIC proposed in this chapter in that it does not handle clustered data. On the other hand, it is more general in that it allows for cluster-specific parameters vectors that are correlated between clusters.

Marginal approximate leave-one-out cross-validation

The marginal Bayesian leave-one-out expected log pointwise predictive density is

$$\text{elpd}_{\text{LOO},m} = \sum_{j=1}^{J} \log p(y_j|y_{-j}) \tag{3.29}$$

where y_{-j} is the response vectors of all clusters except for the j-th cluster and

$$p(y_j|y_{-j}) = \iint p(y_j|\omega, \psi) p(\omega, \psi|y_{-j}) \, \text{d}\omega \text{d}\psi, \tag{3.30}$$

which bears resemblance to the mixed predictive distribution for $\tilde{y}_{j'}$. An estimate of $\text{elpd}_{\text{LOO},m}$ may be obtained from the posterior draws as

$$\widehat{\text{elpd}}_{\text{PSIS-LOO},m} = \sum_{j=1}^{J} \log \left(\frac{\sum_{s=1}^{S} w_j^s p(y_j|\omega^s, \psi^s)}{\sum_{s=1}^{S} w_j^s} \right). \tag{3.31}$$

The raw importance ratios are obtained as $p(y_j|\omega^s, \psi^s)^{-1}$, and these are adjusted by smoothing and truncating as before to obtain weights w_j^s. On the deviance scale,

$$\text{PSIS-LOO}_m = -2\widehat{\text{elpd}}_{\text{PSIS-LOO},m}. \tag{3.32}$$

Marginal deviance information criteria

For marginal DIC, the target quantity is

$$\text{elpd}_{\text{DIC},m} = \sum_{j=1}^{J} \int \log p_{\text{t}}(\tilde{y}_{j'}) p(\tilde{y}_{j'}|\bar{\omega}, \bar{\psi}) \; d\tilde{y}_{j'}. \tag{3.33}$$

The marginal log-likelihood evaluated at the posterior means of the parameters is

$$\text{lpd}^{*}_{m} = \sum_{j=1}^{J} \log p(y_j|\bar{\omega}, \bar{\psi}), \tag{3.34}$$

and plugging in the estimated posterior means for $\bar{\omega}$ and $\bar{\psi}$ yields $\widehat{\text{lpd}}^{*}_{m}$. The log-likelihood integrated over the posterior is

$$\text{lpd}_{\text{DIC},m} = \sum_{j=1}^{J} \iint \log p(y_j|\omega, \psi) p(\omega, \psi|y) \; \text{d}\omega\text{d}\psi, \tag{3.35}$$

which is estimated in MCMC simulation as

$$\widehat{\text{lpd}}_{\text{DIC},m} = \sum_{j=1}^{J} \frac{1}{S} \sum_{s=1}^{S} \left[\log p(y_j|\omega^s, \psi^s) p(\omega^s, \psi^s|y)\right]. \tag{3.36}$$

The estimated effective number of parameters for marginal DIC is

$$\hat{p}_{\text{DIC},m} = 2(\widehat{\text{lpd}}^{*}_{m} - \widehat{\text{lpd}}_{m}), \tag{3.37}$$

the expected log pointwise predictive density is

$$\widehat{\text{elpd}}_{\text{DIC},m} = \widehat{\text{lpd}}^{*}_{m} - \hat{p}_{\text{DIC},m}, \tag{3.38}$$

and the final value on the deviance scale is

$$\text{DIC}_m = -2\widehat{\text{elpd}}_{\text{DIC},m}. \tag{3.39}$$

This marginal DIC corresponds to the DIC_1 of Celeux et al. (2006). It has not been used much in the literature, partly due to the difficulty in integrating out the cluster-specific parameters.

3.4 Adaptive Gaussian quadrature for marginal likelihoods

For models with normally distributed y_{ij}, obtaining $\widehat{\text{lpd}}_{m}$ by way of Equation 3.54 provides an exact and computationally efficient result. For cases where an analytical form for the

integration is unavailable, such as logistic models, Gaussian quadrature may be used to perform numerical integration. (Both methods depend on the integration being performed over a normal prior distribution.) Rabe-Hesketh et al. (2002) applied the adaptive quadrature scheme developed by Naylor and Smith (1982) to generalized linear mixed models. In this chapter, that approach is extended to the individual posterior draws from Markov chain Monte Carlo simulation.

The proposed adaptive quadrature method relies on M standard Gaussian quadrature node locations $G_{\text{std},m}$ and weights $W_{\text{std},m}$, $m = 1 \cdots M$, as well as the posterior mean and standard deviation of each ζ_j. The posterior mean of ζ_j is

$$\hat{\mu}_j = \hat{E}(\zeta_j|y) = \frac{1}{S}\sum_{s=1}^{S}\zeta_j^s, \tag{3.40}$$

and the posterior standard deviation is

$$\hat{\tau}_j = \sqrt{\widehat{\text{var}}(\zeta_j|y_j)} = \sqrt{V_{s=1}^{S}\zeta_j^s}. \tag{3.41}$$

The posterior means and standard deviations are marginal over ω and ψ, whereas adaptive quadrature for maximum likelihood estimation would use conditional quantities. The adaptive quadrature node locations are

$$\text{G}_{jm} = \hat{\mu}_j + \hat{\tau}_j \times \text{G}_{\text{std},m}, \tag{3.42}$$

and their weights are

$$\text{W}_{jm}^s = \sqrt{2\pi} \times \hat{\tau}_j \times \exp\left(\frac{\text{G}_{jm}^2}{2}\right) \times \phi\left(\text{G}_{jm}; 0, \psi^{2,s}\right) \times W_{\text{std},m}. \tag{3.43}$$

The adaptive quadrature node locations will differ between clusters, while the weights will differ between both clusters and MCMC iterations because they depend on ψ^s. The marginal likelihood for cluster j at posterior draw s is approximated as

$$p(y_j|\omega^s, \psi^s) \approx \sum_{m=1}^{M}\left[\text{W}_{jm}^s \prod_{i=1}^{I} p(y_{ij}|\omega^s, \text{G}_{jm})\right], \tag{3.44}$$

where $p(y_{ij}|\omega^s, \text{G}_{jm})$ is similar to the conditional likelihood $p(y_{ij}|\omega^s, \zeta_j^s)$ except that G_{jm} is substituted for ζ_j^s.

3.5 Circular block bootstrap for estimating Monte Carlo error

Straightforward expressions exist for estimating the Monte Carlo error for means or variances of functions of parameters but not for more complicated quantities like DIC, WAIC, and

PSIS-LOO. Instead, the moving block bootstrap (Kunsch, 1989; Liu & Singh, 1992), a bootstrap technique for autocorrelated data, may be used to estimate the Monte Carlo error for these quantities. The moving block bootstrap may be used in this context by concatenating the independent MCMC chains into a single chain having length S. Then blocks of consecutive draws are drawn with replacement and concatenated into a new chain of the same length, S. A quantity of interest, in this case information criterion, is recorded for the new chain. The process is repeated a large number of times, and the standard deviation of the results provides a bootstrap estimate of the Monte Carlo error. This approach may further be improved by using the circular block bootstrap (Politis & Romano, 1992), which joins the ends of the chain, forming a circle. In this way, a sampled block may wrap around from the last observations to the first observations, solving the problem in the moving block bootstrap of under sampling the early and late observations.

Sampling from the original draws in blocks preserves the autocorrelation structure when forming a bootstrap chain, except at the "seams" where the blocks are stitched together. Some care is needed in selecting the size of the blocks to maintain the autocorrelation structure while also obtaining sufficiently different block bootstrap samples. Data-dependent means of selecting a block length have been proposed (Hall, Horowitz, & Jing, 1995; Bühlmann & Künsch, 1999; Politis & White, 2004; Patton, Politis, & White, 2009), but these methods are not used in this chapter. Instead, a simulation is conducted in which a wide range of block sizes are chosen with the circular block bootstrap in order to study how the results may depend on block size. As an aside, both Hall et al. (1995) and Bühlmann and Künsch (1999) suggest a block size of $S^{\frac{1}{3}}$ as a starting point.

3.6 Simulation study of adaptive quadrature and circular block bootstrap

Data and model

Data are generated and analyzed using a linear random intercept model because for this model marginal likelihoods may be obtained by exact integration, which will serve as a benchmark to test the adaptive Gaussian quadrature approximation. The model is

$$y_{ij}|x_j, \beta, \zeta_j, \sigma^2 \sim \mathrm{N}(x_j'\beta + \zeta_j, \sigma^2) \tag{3.45}$$

$$\zeta_j \sim \mathrm{N}(0, \psi^2) \tag{3.46}$$

$$\beta \sim \mathrm{N}(0, 4) \tag{3.47}$$

$$\sigma \sim \mathrm{Exp}(.1) \tag{3.48}$$

$$\psi \sim \mathrm{Exp}(.1), \tag{3.49}$$

where $i = 1 \ldots I$ indexes observations within cluster j, $j = 1 \ldots J$. Further,

$$x_j'\beta = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{1j} x_{2j} + \beta_5 x_{2j} x_{3j}. \tag{3.50}$$

The generating parameters are: $\sigma = 1$, $\psi = 1$, and $\beta = \{-0.5, 0.5, 0.5, 0.5, -0.5, 0\}$. One dataset is simulated for each cluster size $I \in \{25, 50, 100\}$. Each dataset has $J = 200$ clusters, and the covariates in x_j are random draws from a standard normal distribution.

All MCMC simulations within the simulation study use 4 chains of 1000 iterations. The first 500 iterations of each chain are discarded, leaving a total of 2000 posterior draws across the chains. Convergence is monitored using the $\hat{R}$ statistic of Gelman and Rubin (1992); when $\hat{R} < 1.1$ for each parameter and for the log posterior, convergence may be inferred.

The linear random intercept model is a special case of the general model described previously. Let $\omega = \{\beta, \sigma\}$, and then ω, ζ_j, and ψ directly correspond to the parameters of the general model. The likelihood for the linear random intercept model is

$$p(y_{ij}|x_j, \omega, \zeta_j) = \phi(y_{ij}; x_j'\beta + \zeta_j, \sigma^2), \tag{3.51}$$

where ϕ is the normal density function. The conditional log pointwise predictive density is

$$\widehat{\text{lpd}}_c = \sum_{j=1}^{J}\sum_{i=1}^{I} \log \frac{1}{S}\sum_{s=1}^{S} \phi(y_{ij}; x_j'\beta^s + \zeta_j^s, \sigma^{s,2}). \tag{3.52}$$

The simulation focuses on the marginal likelihood, which has a simple form because of the normally distributed y_{ij} and ζ_j:

$$p(y_j|x_j, \omega, \psi) = \Phi\left(y_j; x_j'\beta, \Omega\right), \tag{3.53}$$

where Φ is the multivariate normal density function and Ω is an I-by-I covariance matrix with elements on the diagonal equal to $\psi^2 + \sigma^2$ and elements on the off-diagonal equal to ψ^2. Then marginal log pointwise predictive density is

$$\widehat{\text{lpd}}_m = \sum_{j=1}^{J} \log \frac{1}{S}\sum_{s=1}^{S} \Phi(y_j; x_j'\omega^s, \Omega^s). \tag{3.54}$$

Marginal information criteria may be calculated from this $\widehat{\text{lpd}}_m$ without resorting to a quadrature approximation.

Conditional and marginal information criteria estimates

Information criteria for the three simulated datasets are presented in Table 3.1. The multivariate normal density (as in Equation 3.54) is used to evaluate the marginal likelihood. Values for WAIC and PSIS-LOO ($-2\widehat{\text{elpd}}$) are very close to one another in both the conditional and marginal cases, while those for DIC are slightly lower.

The effective number of parameters $\hat{p}$ may be compared against the number of parameters in focus. In the conditional focus, there are 207 parameters (contained in $\{\beta, \zeta, \sigma\}$), but $\hat{p}_c$ is less than this count in all cases. The reason for this discrepancy is that the prior on ζ is

informative and as such partially constrains ζ, reducing the effective number of parameters. As cluster size increases, $\hat{p}_c$ decreases, reflecting the increase of data available for estimating each ζ_j and the resulting decline in influence of the prior. In the marginal focus, there are 8 parameters (contained in $\{\beta, \sigma, \psi\}$), and $\hat{p}_m$ approximately matches this count.

		Conditional		Marginal		Minimum N. nodes	
I	Criterion	$-2\widehat{\text{elpd}}_c$	$\hat{p}_c$	$-2\widehat{\text{elpd}}_m$	$\hat{p}_m$	Absolute	Relative
25	DIC	14454.35	193.21	14923.10	7.93	25	37
25	WAIC	14457.70	189.48	14922.71	7.18	17	25
25	PSIS-LOO	14457.98	189.62	14922.80	7.22	17	25
50	DIC	28541.09	197.91	29149.04	7.40	37	55
50	WAIC	28542.97	195.88	29148.92	7.00	17	25
50	PSIS-LOO	28542.99	195.89	29148.95	7.02	17	25
100	DIC	56641.63	199.66	57377.89	7.59	83	125
100	WAIC	56642.74	198.71	57378.07	7.42	37	55
100	PSIS-LOO	56642.64	198.66	57378.10	7.43	37	55

Table 3.1: Conditional and marginal DIC, WAIC, and PSIS-LOO for the simulated datasets using the multivariate normal density. In the conditional focus, there are a total of 207 model parameters, whereas there are 8 in the marginal focus. Shown on far right are the minimum numbers of adaptive quadrature nodes (among those considered) needed to obtain an absolute and relative error less than .01. Absolute error refers to the absolute difference between the adaptive quadrature approximation and exact results. Relative error refers to the difference between the adaptive quadrature approximation and the same approximation using one-third fewer nodes.

Adaptive quadrature approximation

Marginal DIC, WAIC, and PSIS-LOO are calculated on the same three simulated datasets using 7, 11, 17, 25, 37, 55, 83, and 125 adaptive quadrature nodes. Each number of nodes is 50% greater than the preceding number, rounded to the nearest odd number. With an odd number of nodes, one node is placed on the mean of the distribution, which does not happen for an even number. The absolute difference between approximate results from adaptive quadrature and exact results from the multivariate normal density are shown in Figure 3.1. In this chapter, an absolute difference less than .01 is judged to be a sufficiently close approximation (shown as horizontal dashed lines), which is somewhat arbitrary but should be conservative unless the information criteria values are very close. The figure indicates that more nodes are required as cluster size increases and that DIC requires more nodes than WAIC or PSIS-LOO.

Marginal DIC required more nodes than WAIC or PSIS-LOO to obtain the desired accuracy, which appears to be due to the difficulty in obtaining accurate values for $\widehat{\text{lpd}}_m^*$

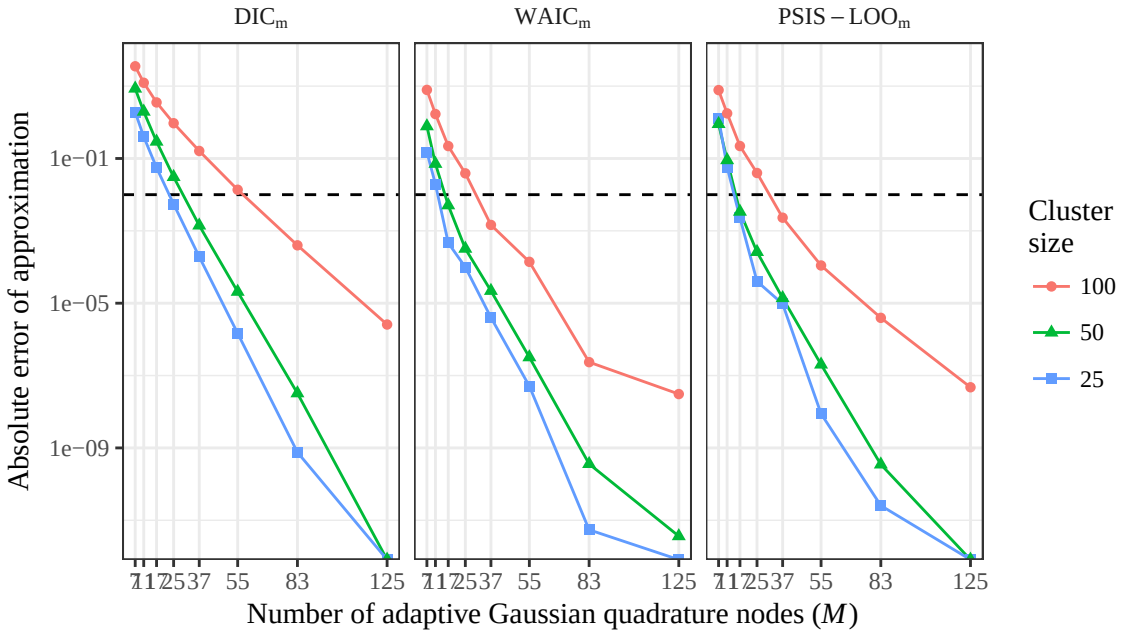


Figure 3.1: Differences in marginal information criteria ($-2 \times \widehat{\text{elpd}}_m$) between calculations using adaptive quadrature (approximate method) and the multivariate normal density function (exact method). The y-axis uses a log scale. The dashed line is drawn at .01.

(Equation 3.34) with adaptive quadrature. To illustrate, consider the most difficult simulation condition, which involved using 7 nodes to estimate marginal quantities for cluster sizes of $I = 100$. In this condition, the difference between the adaptive quadrature approximation for $\widehat{\text{lpd}}^*_m$ and the exact value was 14.08, while the corresponding difference for $\widehat{\text{lpd}}_m$ was -1.81. Given that adaptive quadrature calculations for $p(y_j|\omega^s, \psi^s)$ and $p(y_j|\bar{\omega}, \bar{\psi})$ are identical, the disparity in accuracy is likely due to the fact that $p(y_j|\omega^s, \psi^s)$ is averaged over many posterior draws to obtain $\widehat{\text{lpd}}_m$, but no such averaging is involved in using $p(y_j|\bar{\omega}, \bar{\psi})$ to obtain $\widehat{\text{lpd}}^*_m$. It may be that errors resulting from adaptive quadrature cancel out, at least partially, when they are averaged over posterior draws.

In real applications, adaptive quadrature will only be used when an exact calculation is not available, and so a sufficient number of nodes cannot generally be determined by comparison to an exact calculation. In this case, results may be obtained and compared for different numbers of nodes to calculate a *relative* error of approximation. The relative error of approximation may be calculated for M nodes by finding the absolute value of difference between the information criteria result using M nodes and using $\frac{2}{3}M$, which corresponds to the node counts used in this study. In this chapter, an approximation using M nodes is judged sufficient when the relative error of approximation is less than .01, which again is somewhat arbitrary but is expected to be conservative. Table 3.1 provides the minimum

number of nodes needed to obtain the desired absolute (that is, in comparison to the exact result) and relative precision. The sufficient number of nodes as judged by relative precision tends to be the next higher number than that indicated by absolute precision. In this way, the simulation results support the conservativeness of this approach in testing the number of nodes used.

Circular block bootstrap

Focusing on WAIC, the circular block bootstrap will be used to obtain standard error estimates. Before that, however, a "brute force" approach is used as a reliable means of studying the variability of WAIC that arises from Monte Carlo error. To this end, 200 independent sets of MCMC chains are simulated, and for each the conditional and marginal versions of lpd_{WAIC}, p_{WAIC}, and WAIC are estimated. The same simulated dataset is used throughout, having 200 clusters each having 25 units. The standard deviation of the brute force results may serve as a benchmark against which to compare the bootstrap standard error estimates. The means and standard deviations related to the brute force replications are presented in Table 3.2. Conditional WAIC (WAIC_c) is substantially more variable than marginal WAIC (WAIC_m), as shown by its larger standard deviation. For either the conditional or marginal case, most of the variability appears tied to p_{WAIC} rather than lpd_{WAIC}, which has a relatively low standard deviation across replications.

	Conditional		Marginal	
	Mean	SD	Mean	SD
lpd_{WAIC}	-6981.01	0.14	-7393.95	0.04
$\hat{p}_{\text{WAIC}}$	189.85	0.75	8.44	0.22
WAIC	14341.73	1.55	14804.78	0.43

Table 3.2: Means and standard deviations for the "brute force" WAIC results for the simulation. Results are from 200 independent sets of MCMC chains using the same simulated dataset with cluster size $I = 25$.

Next, the circular block bootstrap is applied once each to the 200 independent sets of MCMC chains. Block sizes ranging from 1 to 100 are used, and the results are presented in Figure 3.2. The bootstrap substantially underestimates the Monte Carlo error for both conditional and marginal WAIC. While Monte Carlo error estimates for lpd_{WAIC} and p_{WAIC} are both too small using the bootstrap, the discrepancy for p_{WAIC} is much worse. Further, block size does not appear to have an effect on estimated Monte Carlo error, possibly because autocorrelations with Stan are typically low. This bootstrap scheme may also be applied to DIC and PSIS-LOO, but there is no reason to expect better performance.

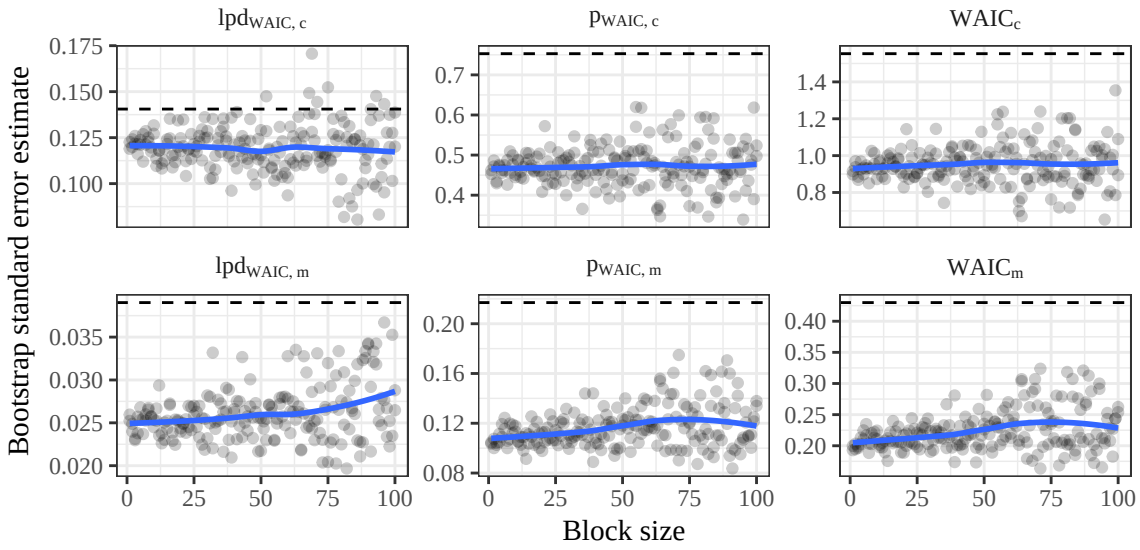


Figure 3.2: Circular block bootstrap standard error estimates for WAIC by block size for the simulation. Points are bootstrap estimates from 200 independent sets of MCMC chains using the same simulated dataset, and the solid line is a loess curve fit to the points. The dashed line is the "brute force" standard error estimate. The y-axes vary.

3.7 Applied example

Data and models

A latent regression Rasch model is fit to a dataset on verbal aggression (Vansteelandt, 2000) that consists of $J = 316$ persons and $I = 24$ items. Participants were instructed to imagine four frustrating scenarios, and for each they responded to items regarding whether they would react by cursing, scolding, and shouting. They also responded to parallel items regarding whether they would *want* to engage in the three behaviors, resulting in a total six items per scenario (cursing/scolding/shouting $\times$ doing/wanting). An example item is, "A bus fails to stop for me. I would want to curse." The response options for all items were "yes", "perhaps", and "no." The items have been dichotomized for this example by combining "yes" and "perhaps" responses. Two person-related covariates are included: the respondent's trait anger score (Spielberger, 1988), which is a raw score from a separate measure taking values ranging from 11 to 39 in the data, and an indicator for whether the respondent is male, which takes the values 0 and 1.

The model is

$$y_{ij}|w_j, \lambda, \zeta_j, \delta_i \sim \text{Bernoulli}\left(\text{logit}^{-1}(w_j'\lambda + \zeta_j - \delta_i)\right) \tag{3.55}$$

$$\delta_1 \ldots \delta_{I-1} \sim \text{N}(0, 9) \tag{3.56}$$

$$\zeta_j \sim \mathrm{N}(0, \sigma^2) \tag{3.57}$$

$$\sigma \sim \mathrm{Exp}(.1) \tag{3.58}$$

$$\lambda \sim t_1(0, 1), \tag{3.59}$$

where $y_{ij} = 1$ if the response for person j to item i is correct and $y_{ij} = 0$ otherwise, w_j is a vector of person-related covariates, λ is a vector of latent regression coefficients, ζ_j is a person residual, and δ_i is an item difficulty parameter. One element of w_j is one for the intercept, and the last item difficulty is constrained, $\delta_I = -\sum_i^{(I-1)} \delta_i$. The priors for λ match those recommended by Gelman, Jakulin, Pittau, and Su (2008) for logistic regression. First, the covariates (w_j) are transformed. Continuous covariates are mean-centered and then rescaled to have a standard deviation of .5. Binary covariates are also mean-centered and then are rescaled by dividing by the difference between their maximum and minimum values, which results in a range of 1. A constant supplied for the model intercept is left to equal 1. With these transformations, the same prior is applied to all coefficients, $\lambda \sim t_1(0, 1)$, where t_1 is the Student's t distribution with one degree of freedom. A transformation may be applied to λ to find what the regression coefficients would be on the original scale of the covariates.

Focus is placed on ζ_j for the conditional approach, which yields a prediction inference involving new responses from the same persons (and items). The marginal approach, perhaps more realistically, places focus on σ, implying a prediction inference involving new responses from a new sample of persons. Five competing models are considered, differing only in what person covariates are included: Model 1 includes no covariates, Model 2 has the trait anger score, Model 3 has the indicator for male, Model 4 has both covariates, and Model 5 has both covariates and their interaction. All models include an intercept term.

Results

The five models are estimated with Stan using 5 chains of 2,500 draws with the first 500 draws of each discarded, resulting in a total of 10,000 kept posterior draws. The larger number of posterior draws is chosen here due to the anticipated Monte Carlo errors, but such a large number is not ordinarily necessary for estimating, for example, the posterior means and standard deviations for parameters. Marginal DIC, WAIC, and PSIS-LOO are computed using adaptive quadrature to integrate out ζ. Arbitrarily focusing on Model 4, the process described in the previous section was used to determine that 17 nodes are sufficient to obtain an accurate approximation. All evaluations of the marginal likelihood that follow use this number of nodes.

In order to obtain an indication of the variability of results owing to Monte Carlo error, 10 independent sets of MCMC chains are run for each model. The circular block bootstrap is not used for this purpose, as it was found to substantially underestimate the Monte Carlo error. Figure 3.3 provides the estimated effective number of parameters ($\hat{p}$) for each model and focus, as well as the count of parameters associated with each focus. For conditional information criteria, the $\hat{p}$ are substantially less than the counts of parameters, owing mainly

to the fact that each ζ_j, with its hierarchical prior, contributes less than one to the effective number of parameters. On the other hand, $\hat{p}$ for marginal information criteria are close to the counts of model parameters, and for WAIC and PSIS-LOO it tends to be slightly larger than the count of parameters.

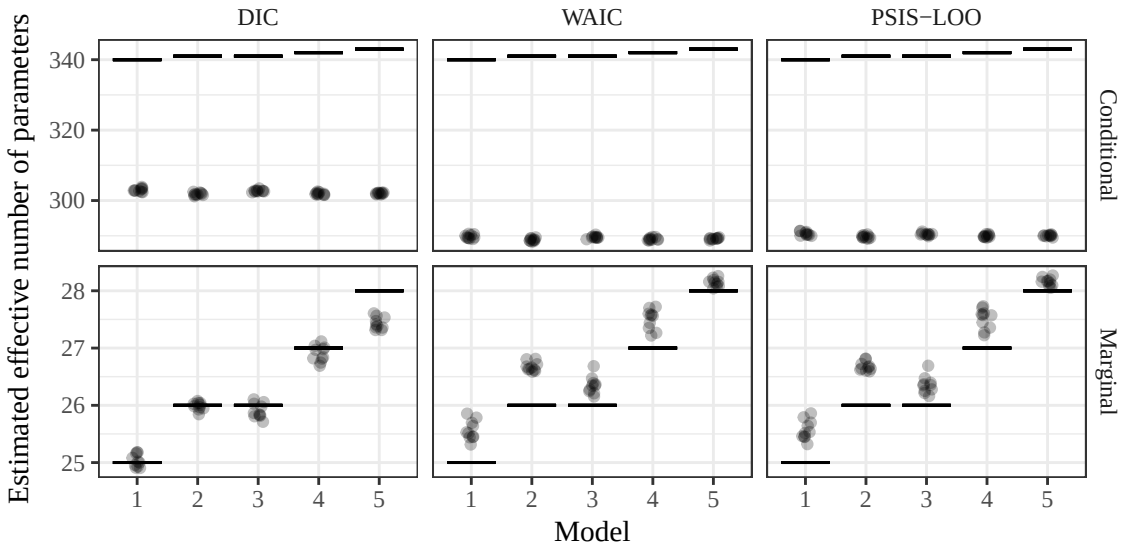


Figure 3.3: Estimated effective number of parameters ($\hat{p}$) for the five latent regression Rasch models. Points represent the results of the 10 independent MCMC simulations per model. A small amount of horizontal jitter is added to the points. The horizontal lines represent the counts of parameters associated with each model and focus. The y-axes vary by focus.

Figure 3.4 provides the estimates for the information criteria values themselves ($-2\widehat{\text{elpd}}$). The different conditional information criteria differ from each other for any given model, though they seem to show a similar pattern between models. The high degree of Monte Carlo error in the conditional focus renders differentiating the predictive performance of the models difficult. In the marginal focus the amount of Monte Carlo error is less but still poses a degree of difficulty in making close comparisons. Models 1 and 2 clearly provided poorer predictions in comparison to the others using marginal information criteria, and there is some evidence supporting Model 4 as the best among the candidates.

As discussed in their respective sections, WAIC and PSIS-LOO are associated with criteria to support the reliability of their estimates. For WAIC any given point should contribute less than .4 to $\hat{p}$, and for PSIS-LOO a point should not have a Pareto shape parameter greater than one. As mentioned earlier, "point" is defined either as a unit or a cluster depending on whether the focus is conditional or marginal, respectively. Table 3.3 provides the counts of problematic observations by model, averaging over the 10 repeated MCMC simulations. Conditional WAIC has a small number problematic observations with each model, while marginal WAIC does not exhibit any such issues. No problematic observations are found for

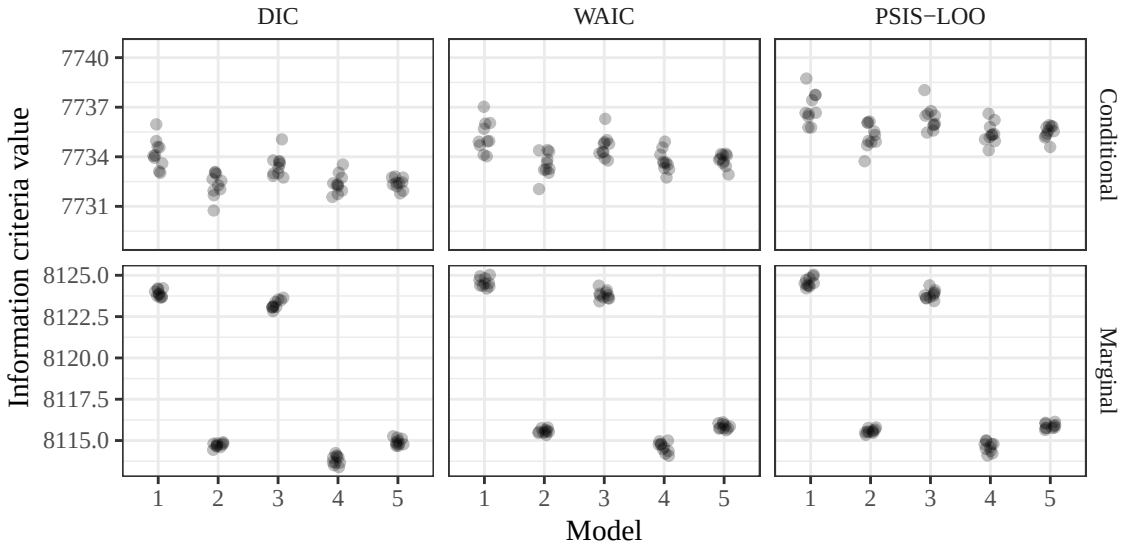


Figure 3.4: Information criteria values ($-2\widehat{\text{elpd}}$) for the five latent regression Rasch models. Points represent the results of the 10 independent MCMC simulations per model. A small amount of horizontal jitter is added to the points. The y-axes vary by focus.

either form of PSIS-LOO.

	Conditional		Marginal	
Model	WAIC	PSIS-LOO	WAIC	PSIS-LOO
1	3.0	0.0	0.0	0.0
2	3.3	0.0	0.0	0.0
3	2.0	0.0	0.0	0.0
4	3.2	0.0	0.0	0.0
5	3.1	0.0	0.0	0.0

Table 3.3: Counts of problematic observations for WAIC and PSIS-LOO by model, averaged over the 10 repeated MCMC simulations. For WAIC, this is the count of observations that contribute more than .4 to $\hat{p}$, and for PSIS-LOO this is the number of observations having a Pareto shape parameter greater than one. In the conditional focus, observations are defined at the unit level, whereas they are defined at the cluster level for the marginal focus.

3.8 Discussion

The choice of conditional or marginal focus should depend on the prediction inference to be made, but the marginal approach was found to have some advantages over the conditional

approach. Marginal information criteria were found to have less Monte Carlo error and to be more robust in terms of pointwise diagnostics for WAIC and PSIS-LOO. Marginal information criteria are easily obtained for linear models when the cluster-specific parameters to be integrated out are assigned a normal prior. For non-linear models, analysis revealed adaptive Gaussian quadrature to be a viable means of obtaining the necessary marginal likelihoods. The methods described in this chapter may be extended to models with cluster-specific vectors of parameters having a multivariate normal prior.

The preceding analyses also demonstrated the existence of non-ignorable Monte Carlo error in both marginal and conditional WAIC, PSIS-LOO, and DIC, though the issue is substantially worse for the conditional information criteria. Caution is therefore advised when using these methods for model comparison. Disappointingly, the circular block bootstrap did not provide reasonable estimates for Monte Carlo error of information criteria. The approach taken in the applied example of simply rerunning the MCMC simulation many times is cumbersome and merely suggestive of the amount of Monte Carlo error, but it could provide accurate results if the time and computing power is available to conduct a larger number of replications.

The conditional information criteria are more easily obtained than the marginal, as the conditional information criteria depend on quantities easily generated from MCMC software. In fact, BUGS and JAGS provide conditional DIC by default, perhaps accounting for the popularity of DIC. Researchers may rely on these defaults without an awareness of the marginal alternatives, running the risk of obtaining inappropriate prediction inferences. Further, consideration is not usually given to the degree of Monte Carlo error associated with information criteria, which as demonstrated in the applied example may be substantial even with a large number of weakly correlated posterior draws. In short, the naive application of these techniques leaves a great deal of room for obtaining misleading results.

Bibliography

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Borjas, G. J. & Sueyoshi, G. T. (1994). A two-stage estimator for probit models with structural group effects. *Journal of Econometrics*, *64*(1), 165–182.

Bühlmann, P. & Künsch, H. R. (1999). Block length selection in the bootstrap for time series. *Computational Statistics & Data Analysis*, *31*(3), 295–310.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–673.

Chen, M.-H., Ibrahim, J. G., Shah, A. K., Lin, J., & Yao, H. (2012). Meta-analysis methods and models with applications in evaluation of cholesterol-lowering drugs. *Statistics in medicine*, *31*(28), 3597–3616.

Cho, S.-J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: item modeling for explanation and item generation. *Psychometrika*, *79*(1), 84–104.

Dawid, A., Cowell, R., Lauritzen, S., & Spiegelhalter, D. (1999). *Probabilistic networks and expert systems*. New York: Springer-Verlag.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, *64*(4), 407–433.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Psychology Press.

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. *Journal of Data Science*, *9*(1), 15–21.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–243). New York, NY: Springer.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). *Model determination using predictive distributions with implementation via sampling-based methods*. Defense Technical Information Center. Fort Belvoir, VA.

Gelfand, A. E., Sahu, S. K., & Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, *82*(3), 479–488.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd). Texts in Statistical Science. Boca Raton, FL: Chapman & Hall.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–807.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gierl, M. J. & Haladyna, T. M. (2013). *Automatic item generation: theory and practice*. New York: Routledge.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*(2), 430–431.

Greven, S. & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*(4), 773–789.

Hall, P., Horowitz, J. L., & Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, *82*(3), 561–574.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, *17*(2), 295–311.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. D. Boeck & M. Wilson (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Jones, K. S., Nakagawa, S., & Sheldon, B. C. (2009). Environmental sensitivity in relation to size and sex in birds: meta-regression analysis. *The American Naturalist*, *174*(1), 122–133.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*(1), 140–155.

Knowles, S. C., Nakagawa, S., & Sheldon, B. C. (2009). Elevated reproductive effort increases blood parasitaemia and decreases immune function in birds: a meta-regression approach. *Functional Ecology*, *23*(2), 405–415.

Kuha, J. (2004). AIC and BIC: comparisons of assumptions and performance. *Sociological Methods & Research*, *33*, 188–229.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241.

Li, L., Qiu, S., Zhang, B., & Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, *26*(4), 881–897.

Liang, H., Wu, H., & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, *95*, 773–778.

Liu, R. Y. & Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In R. LePage & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 255–248). New York: Wiley.

Marshall, E. C. & Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: A simulation-based approach. *Bayesian Analysis*, *2*(2), 409–444.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*(392), 993–997.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*(3), 281–296.

Naylor, J. C. & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C*, *31*(3), 214–225.

Patton, A., Politis, D. N., & White, H. (2009). Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Reviews*, *28*(4), 372–375.

Pinheiro, J. C. & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), 12–35.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, *181*(2), 108–119.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523–539.

Politis, D. N. & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 263–270). New York: Wiley.

Politis, D. N. & White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, *23*(1), 53–70.

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

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323.

Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral statistics*, *19*(4), 337–350.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. & Bryk, A. S. (1985). Empirical bayes meta-analysis. *Journal of Educational and Behavioral Statistics*, *10*(2), 75–98.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205.

Ritz, J. & Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, *13*(4), 309–323.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Saefken, B., Kneib, T., van Waveren, C.-S., Greven, S., et al. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics*, *8*(1), 201–225.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Skrondal, A. & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A*, *172*(3), 659–687.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B*, *76*(3), 485–493.

Spielberger, C. (1988). *State-trait anger expression inventory research edition.* Psychological Assessment Resources. Odessa, FL.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B*, *39*(1), 44–47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite sample corrections. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26.

Trevisani, M. & Gelfand, A. E. (2003). Inequalities between expected marginal log-likelihoods, with implications for likelihood-based model complexity and comparison measures. *Canadian Journal of Statistics*, *31*(3), 239–250.

Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370.

van der Laan, M. J. & Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples*. Division of Biostatistics, University of California, Berkeley.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Vansteelandt, K. (2000). *Formal models for contextualixed personality psychology* (Unpublished doctoral dissertation, University of Leuven, Leuven, Belgium).

Vehtari, A. & Gelman, A. (2016). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 1–20.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

Wilson, M. (2004). *Constructing measures: an item response modeling approach*. Mahway, New Jersey: Lawrence Erlbaum.

Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In M. Wilson & P. De Boeck (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 43–74). New York: Springer.

Yu, D. & Yau, K. K. (2012). Conditional Akaike information criterion for generalized linear mixed models. *Computational Statistics & Data Analysis*, *56*(3), 629–644.

Yu, D., Zhang, X., & Yau, K. K. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *Journal of Multivariate Analysis*, *116*, 245–262.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589–600.

Appendix A

Software details

A.1 Software details for chapter 2

Stata 14.2 was used to conduct the simulation. The Rasch and LLTM models were fit using the `melogit` command. For example, if `y` is the response variable and `item` and `person` are both factor variables, then the Rasch model may be estimated as follows:

```
melogit y ibn.item, noconstant || person:
```

A new dataset may then be created having one row per item. Suppose that the variable `delta` is the estimated item difficulties and `delta_se` is the standard errors for those estimates. Also, suppose that `x1`, `x2`, and `x3` are item covariates. The LLTM-E2S may be fit as follows:

```
gsem (delta <- c.delta_se#c.M@1 x1 x2 x3), variance(M@1)
```

Figures and tables were prepared using R 3.4.0 with the ggplot2 2.2.1 and xtable 1.8.2 packages, respectively. The following R packages were additionally used: grid 3.4.0, gridExtra 2.2.1, readstata13 0.8.5, reshape2 1.4.2.

A.2 Software details for chapter 3

All analysis was conducted using R 3.4.0. MCMC simulation was carried out using the rstan 2.15.1 package, which is the R implementation of Stan. WAIC and PSIS-LOO estimates were obtained using loo 1.1.0. The block bootstrap was carried out using the `tsboot` function in boot 1.3.19. The example application data was from edstan 1.0.6. Figures and tables were prepared using the ggplot2 2.2.1 and xtable 1.8.2, respectively. The following R packages were additionally used: doParallel 1.0.10, foreach 1.4.3, matrixStats 0.52.2, mvtnorm 1.0.6, reshape2 1.4.2, statmod 1.4.29.

The Stan code used for the random intercept model is printed below. The transformed data and generated quantities blocks are needed only to sample from the posterior log-likelihoods. In the generated quantities block, `cll_ij` is the conditional log-likelihood for an observation, and `mll_j` is the marginal log-likelihood for a cluster.

```
data {
  int<lower=1> I;                 // # obs per cluster
  int<lower=1> J;                 // # clusters
  int<lower=1, upper=I> ii[I*J];  // obs for n (1:I for each cluster)
  int<lower=1, upper=J> jj[I*J];  // cluster for n
  vector[I*J] y;                  // measurement for n
  int<lower=1> L;                 // # covariates
  matrix[J, L] X;                 // covariate matrix
}
transformed data{
  vector[I] y_vecs[J];
  for(n in 1:(I*J)) y_vecs[jj[n]][ii[n]] = y[n];
}
parameters {
  vector[L] beta;
  real<lower=0> sigma;
  real<lower=0> psi;
  vector[J] zeta;
}
model {
  vector[J] eta;
  eta = X*beta;
  beta ~ normal(0, 2);
  sigma ~ exponential(.1);
  psi ~ exponential(.1);
  zeta ~ normal(0, psi);
  y ~ normal(eta[jj] + zeta[jj], sigma);
}
generated quantities {
  vector[J] eta;
  vector[I*J] cll_ij;
  vector[J] mll_j;
  eta = X*beta;
  for(n in 1:I*J)
    cll_ij[n] = normal_lpdf(y[n] | eta[jj[n]] + zeta[jj[n]], sigma);
  {
    matrix[I, I] Omega;
```

```
    Omega = rep_matrix(psi^2, I, I);
    for(i in 1:I)
      Omega[i,i] = psi^2 + sigma^2;
    for(j in 1:J)
      mll_j[j] = multi_normal_lpdf(y_vecs[j] | rep_vector(eta[j], I), Omega);
  }
}
```

The Stan code used for the latent regression Rasch model is below. This is a modified version of the Rasch model provided by edstan. It adds in sampling from the posterior of `zeta` and `theta_fix`, which are required for approximating the marginal log-likelihood.

```
functions {
  matrix obtain_adjustments(matrix W) {
    real min_w;
    real max_w;
    int minmax_count;
    matrix[2, cols(W)] adj;
    adj[1, 1] = 0;
    adj[2, 1] = 1;
    if(cols(W) > 1) {
      for(k in 2:cols(W)) {                        // remaining columns
        min_w = min(W[1:rows(W), k]);
        max_w = max(W[1:rows(W), k]);
        minmax_count = 0;
        for(j in 1:rows(W))
          minmax_count = minmax_count + W[j,k] == min_w || W[j,k] == max_w;
        if(minmax_count == rows(W)) {       // if column takes only 2 values
          adj[1, k] = mean(W[1:rows(W), k]);
          adj[2, k] = (max_w - min_w);
        } else {                            // if column takes > 2 values
          adj[1, k] = mean(W[1:rows(W), k]);
          adj[2, k] = sd(W[1:rows(W), k]) * 2;
        }
      }
    }
    return adj;
  }
}
data {
  int<lower=1> I;               // # questions
  int<lower=1> J;               // # persons
```

```
  int<lower=1> N;                // # observations
  int<lower=1, upper=I> ii[N];   // question for n
  int<lower=1, upper=J> jj[N];   // person for n
  int<lower=0, upper=1> y[N];    // correctness for n
  int<lower=1> K;                // # person covariates
  matrix[J,K] W;                 // person covariate matrix
}
transformed data {
  matrix[2,K] adj;                // values for centering and scaling covariates
  matrix[J,K] W_adj;              // centered and scaled covariates
  adj = obtain_adjustments(W);
  for(k in 1:K) for(j in 1:J)
      W_adj[j,k] = (W[j,k] - adj[1,k]) / adj[2,k];
}
parameters {
  vector[I-1] delta_free;
  vector[J] theta;
  real<lower=0> sigma;
  vector[K] lambda_adj;
}
transformed parameters {
  vector[I] delta;
  delta[1:(I-1)] = delta_free;
  delta[I] = -1*sum(delta_free);
}
model {
  target += normal_lpdf(delta | 0, 3);
  theta ~ normal(W_adj*lambda_adj, sigma);
  lambda_adj ~ student_t(3, 0, 1);
  sigma ~ exponential(.1);
  y ~ bernoulli_logit(theta[jj] - delta[ii]);
}
generated quantities {
  vector[K] lambda;
  vector[J] theta_fix;
  vector[J] zeta;
  lambda[2:K] = lambda_adj[2:K] ./ to_vector(adj[2,2:K]);
  lambda[1] = W_adj[1, 1:K]*lambda_adj[1:K] - W[1, 2:K]*lambda[2:K];
  theta_fix = W_adj*lambda_adj;
  zeta = theta - theta_fix;
}
```