Title
A Posterior Predictive Model Checking Method Assuming Posterior Normality for Item Response Theory

Author
Kuhfeld, Megan Rebecca

Publication Date
2016

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

A Posterior Predictive Model Checking Method Assuming Posterior Normality for Item Response Theory

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Megan Rebecca Kuhfeld

2016

Abstract of the Thesis Proposal

A Posterior Predictive Model Checking Method Assuming Posterior Normality for Item Response Theory

by

Megan Rebecca Kuhfeld

Master of Science in Statistics

University of California, Los Angeles,

Professor Qing Zhou, Chair

This study investigated the violation of local independence assumptions within unidimensional item response theory (IRT) models. IRT models assume that for a given value of the latent variable θ, the value of any observed variable is conditionally independent of all other variables. Violation of this assumption can bias item parameter estimates and latent trait scores. There are two existing classes of procedures to check for local dependence (LD): (a) frequentist model appraisal methods that rely on the expected and observed bivariate item frequencies, and (b) posterior predictive model checking (PPMC) methods, which are a flexible family of Bayesian model checking procedures. The advantages of the PPMC method is that it accounts for parameter estimation uncertainty and does not require asymptotic arguments. Given the current dominance of maximum likelihood approaches for the estimation of IRT models, I propose a posterior predictive model checking method for evaluating LD in IRT models that can be implemented using only byproducts of likelihood-based estimation. This approach, which relies on a posterior normality approximation, was found to be comparable to the fully Bayesian PPMC approach in terms of the sensitivity to local dependence in IRT models.

The thesis of Megan Rebecca Kuhfeld is approved.

Li Cai

Yingnian Wu

Qing Zhou, Committee Chair

University of California, Los Angeles

2016

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Statistical models are fit to data in an attempt to understand a set of complicated processes underlying the phenomena of interest. In applications within the social sciences, it is well known that any model is merely an approximation to reality. Psychometric models, such as factor analysis and item response theory models, are constructed around the hypothesis of a parsimonious set of latent factors that explain the relationship between observed scores or item responses. These models generally impose a large set of restrictions and assumptions regarding the distribution of observed variables and how variables are allowed to relate to each other. Given that these models are simplifications of underlying processes, it is important to evaluate the fit of the model in terms of the proposed structure of the data that the model implies. That is to say, the central focus of the assessment of data-model fit is characterizing the discrepancy between the observed data and the model-implied structure of the data.

Traditional item response theory (IRT) models specify a single continuous latent variable θ that explains the probability of endorsing a set of dichotomous or polytomous item responses, with the goal of representing individual differences on a psychological, behavioral, or educational construct. Accurate estimation of item parameters and the inferences made based on the estimates of θ depend on the degree to which the unidimensionality assumption holds in the observed item response data. The key implication of unidimensionality within item response theory is the conditional (or local) independence assumption. Local independence

means that for a given value of the latent variable θ, the joint probability of correct responses to an item pair is the product of the probabilities of correct responses to the two items,

$$P(y_j = 1, y_k = 1|\theta) = P(y_j = 1|\theta)P(y_k = 1|\theta), \tag{1.1}$$

where j and k index items, y_j is the observed response to item j.

However, there are many situations where this assumption is unlikely to hold. For example, items within a reading assessment that follow the same passage are likely to exhibit local dependence (LD). Clusters of items with overly similar meaning or phrasing may also violate the assumption of unidimensionality. Additionally, many psychological assessments are designed to measure an overall dimension (e.g., general health), but also are intended to measure sub-domains such as eating habits, sleep issues, and depression, which would result in LD if the sub-dimensions are not explicitly modeled.

The problem with locally dependent items is that ignoring local dependence can result in biased item parameter estimates (Chen & Thissen, 1997). As a result, local dependence can lead to poor estimation of the individual latent construct (Zenisky, Hambleton, & Sireci, 2002) and over estimation of the IRT information and test reliability (Sireci, Thissen, & Wainer, 1991).

There are two parallel lines of research for detecting local dependence that are used in the context of the two different IRT parameter estimation methods: frequentist estimation (e.g., using maximum likelihood) and Bayesian estimation. A full comparison of the theory, rationale, and assumptions of maximum likelihood and Bayesian estimation methods is beyond the scope of this paper (see Cai and Thissen (2015) and Fox (2010) for overviews). Instead, I focus on comparing and contrasting the model fit procedures specifically related to the examination of local dependence within each estimation framework.

Numerous indices and statistics for detecting local dependence have been proposed for IRT models estimated using maximum likelihood. Houts and Edwards (2013) provide an overview of the commonly-used IRT-based tests of local dependence. These include Yen's (1984) Q3 statistic, Chen and Thissen's (1997) use of the Pearson's χ^2 and the likelihood ratio statistic G^2, and Ip's (2001) suggestion to use the Mantel-Haenszel (MH) test with multiple testing corrections. These statistics are estimated based on a process that involves fitting a unidimensional IRT model to the data, and then comparing the observed responses with the model-expected responses.

Within Bayesian estimation, there is a promising set of model assessment techniques involving posterior predictive model checking (PPMC; Meng, 1994; Gelman, Meng, & Stern, 1996). The PPMC approach investigates the compatibility of a posited model to observed data by assessing the features of the observed data in relation to the model's implications. Replicated data can be simulated from a fitted model, and are referred to as posterior predictive data since values are drawn conditional on the observed data. PPMC compares the observed data with replicated data using a predefined test quantity, also known as discrepancy measure, which is a function of both the data and the parameters (Gelman et al., 1996). The assumption behind PPMC is that any systematic differences between the two data sets indicate a failure of the model to explain those aspects of the data.

This model-checking method requires the researcher to simulate draws from the full posterior distribution of the model parameters, which is straight-forward when one uses Bayesian sampling-based methods (e.g., Markov chain Monte Carlo (MCMC) estimation; Gilks, Richardson, & Spiegelhalter, 1996). Sinharay, Johnson, and Stern (2006) and Levy (2006), among others, have demonstrated the PPMC method to assess local dependence in item response theory models using various discrepancy measures, including the Pearson's χ^2, Mantel-Haenszel (MH)

statistic, and the likelihood ratio G^2.

While the same general set of discrepancy measures are used in the maximum likelihood and Bayesian examination of local dependence, the procedures and underlying assumptions are quite different. Frequentist approaches for assessing model fit evaluate the discrepancy between the observed data and the hypothesized model when the unknown model parameters are replaced by the best-fitting point estimates. There are two central problems with this approach. The use of point estimates for the unknown parameters in model fit assessment understates the uncertainty in the sampling distributions of discrepancy measures (Meng, 1994). Secondly, the null sampling distributions for most discrepancy statistics in likelihood model fit approaches are justified only asymptotically, which may not be tenable in real data situations.

By contrast, the PPMC method is a simulation-based model checking method that does not require asymptotic arguments. When the plausible values of the data can be drawn from the posterior predictive distribution, empirically constructing reference distributions of test quantities does not present a problem. Furthermore, the PPMC methods integrates parameter uncertainty into model fit assessment through the use of the posterior predictive distribution.

1.1 Rationale for this study

The PPMC method has multiple advantages over frequentist approaches to check for the local independence. However, maximum likelihood methods for parameter estimation and model fit testing currently dominate the field of in item response theory models, and the adoption of fully Bayesian approaches is likely to introduce many new complexities for applied users of IRT. Without additional training, both the specification of prior distributions and convergence monitoring of MCMC may present large barriers to an applied researcher. The goal of this study is

develop and evaluate a method that allows researchers to reap benefits of the (inherently Bayesian) posterior predictive model checking method while using likelihood estimation.

The current study expands upon an alternative PPMC method that was developed in the context of structural equation modeling (Lee, Cai, & Kuhfeld, 2016). This method, termed the Poor Person's PPMC method (PP-PPMC), employs only byproducts of maximum likelihood estimation (e.g., the ML parameter estimates and the associated asymptotic covariance matrix) in the estimation of the replicated data. This method is based on a well-known result in Bayesian literature that asymptotically the likelihood tends to dominate the posterior shape. Furthermore, for large samples, the estimated item parameters are approximately normally distributed around the ML estimates, with the inverse of the Fisher information matrix as the covariance matrix (Gelman, Carlin, Stern, & Rubin, 2003).

In this thesis, a posterior predictive model checking method assuming posterior normality is outlined in the context of item response theory modeling, and then employed in a series of Monte Carlo studies to test for local dependence when the true data structure is multidimensional but a unidimensional model is fit to the data. I chose to focus on LD indices, rather than overall model fit discrepancy measures, because LD indices provide more valuable feedback regarding sources of misfit than overall fit measures.

This work is significant because it is the first to bridge the disparate fields of local dependence assessment within maximum likelihood and Bayesian estimation. This study allows for the examination of the conditions under which a simulation-based model checking method improves upon inferences made about model fit using classical likelihood statistics. Additionally, I provide a comparison of the posterior predictive model checking assuming posterior normality (PPMC-N) method and the fully Bayesian PPMC method to demonstrate the extent to

which the multivariate normal approximation of the parameter posterior mirrors the estimated parameter posterior distributions.

The remainder of this thesis is organized as follows. First, item response theory models for dichotomous item response data are introduced. Next, background on the Bayesian PPMC method is provided, and two different likelihood-based posterior approximations are introduced. Subsequently, a set of discrepancy measures that can be used to detect LD within item pairs are described. A simulation study is outlined to examine the performance of the normality assumption PPMC approach compared to both a fully Bayesian PPMC approach and a frequentist approach to detect violations of unidimensionality in the estimation of item response theory models. The remaining sections describe and discuss the results and implications of the simulation study, followed by conclusions and future directions.

CHAPTER 2

Item Response Theory (IRT) models

In social and behavioral sciences, questionnaires or tests are often used to measure traits that are not directly observable. Item-level data from surveys and tests are generally categorical, with either dichotomous (yes/no, correct/incorrect) or polytomous (for example, strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) item response formats. When items follow this response format and the latent (unobserved) variables are assumed to be continuous, item response theory (IRT) models are a flexible set of models that can be used to make inferences about the latent variables (Mislevy, 1986; Bock, Gibbons, & Muraki, 1988).

First, some notation that is used throughout the chapter is introduced. Let there be $i = 1, ..., N$ independent respondents and $k = 1, ..., n$ items. Let the response from person i to item k be y_{ki}. It can be assumed that y_{ki} takes integer values from $\{0, 1, ..., C_k - 1\}$, where C_k is the number of response categories for item k. For this study, I only focus on dichotomous items ($C = 2$).

Let the $n \times 1$ vector of item responses for respondent i be $\mathbf{y}_i = (y_{1i}, ..., y_{ki}, ..., y_{ni})'$, and let $\mathbf{Y}$ be the matrix of all observed response patterns.

For this section, we assume that the assumption of unidimensionality holds for the set of n items. The item response theory (IRT) model specifies the conditional probability for the response to item k, given the individual's latent trait level θ_i. The two-parameter logistic (2PL) item response model is used specifies the conditional response probability curve (or traceline) of a correct response as a

function of the latent variable θ and the item parameters (a and c):

$$P(y_{ki} = 1|\theta_i, a_k, c_k) = \frac{1}{1 + \exp[-(a_k\theta_i + c_k)]} \tag{2.1}$$

where θ_i is the latent variable value for individual i, a_k is the slope for item k, and c_k is the intercept for item k. The slope parameter a_k describes the strength of the relationship between item k and the latent variable θ. The intercept parameter represent the boundary between the two response categories. For dichotomous items, the probability of an incorrect responses is $P(y_{ki} = 0|\theta, a_k, c_k) = 1 - P(y_{ki} = 1|\theta, a_k, c_k)$.

2.1 Local Independence Assumption

Unidimensional item response theory models assume that the trait value explains an examinee's performance on a test, so when the latent trait is accounted for in the model, item responses are independent. Pairwise local dependence can be expressed as

$$P(y_k = 1, y_{k'} = 1|\theta) = P(y_k = 1|\theta)P(y_{k'} = 1|\theta). \tag{2.2}$$

for all $k \neq k'$, where k and k' index two different items.

Previous researchers have distinguished between types of LD present among items, which have differential effects on parameter estimation (Chen & Thissen, 1997). The two types of LD are *underlying* local dependence (ULD), which results from unmodeled latent traits, and *surface* local dependence (SLD), which is due to highly similar content or placement (Thissen, Bender, Chen, Hayashi, & Wiesen, 1992). Surface LD is defined by the probability π_{LD} that the second item in the LD pair will have a response identical to the first item. That is to say, for a set portion of respondents given by π_{LD}, the second response is determined completely

by the first item within the item pair ($y_2 = 1$ if $y_1 = 1$, $y_2 = 0$ otherwise). SLD is generally seen as nuisance, while ULD can be seen as nuisance dimensions or as evidence of the need for latent dimensions to be added to the model.

2.2 Parameter Estimation

The estimation of IRT item parameters is commonly conducted using full-information estimation methods, which require the observed data to be arranged into a n-way contingency table, where each of the n dimensions represents in a response to a given item. Full-information approaches estimate the frequencies of each individual response pattern (i.e., the frequencies within each cell of the contingency table). The central methods for the estimation of IRT item parameters are full-information maximum likelihood (FIML) estimation and Bayesian estimation.

2.2.1 Full-Information Maximum Likelihood (FIML) Estimation

Full-information maximum likelihood (FIML) method is the most commonly-used method of estimation for item response theory models, and involves q-dimensional integration over a multivariate distribution, where q is the number of latent dimensions in the model (Bolt, 2005).

For all items, let $\boldsymbol{\gamma}$ be the vector of all item parameters. The goal of the FIML method is to find the set of parameters that would maximize the likelihood (or log-likelihood) given the observed data.

Let Y_{ki} be a Bernoulli (0-1) random variable representing individual i's response to item k, where y_{ki} is a realization of Y_{ki}. The conditional probability of the event $Y_{ki} = y_{ki}$ is equal to

$$P(Y_{ki} = y_{ki}|\theta, \boldsymbol{\gamma}) = P(\theta)^{y_{ki}}[1 - P(\theta)]^{1-y_{ki}}, \tag{2.3}$$

where $P(\theta)$ is the same as $P(y_{ki} = 1|\theta)$ defined in Equation 2.1. Given the assumption of independence of observations across cases, the probability of response pattern $\mathbf{y}_i = (y_{1i}, ..., y_{ki}, ..., y_{ni})'$ is a product over individual item response probabilities

$$P(\mathbf{y}_i|\theta, \boldsymbol{\gamma}) = \prod_{k=1}^{n} P(\theta)^{y_{ki}}[1 - P(\theta)]^{1-y_{ki}}. \tag{2.4}$$

The joint probability of the observed responses and latent variables θ is equal to the product of the conditional probability of the observed variables given the latent variable, multiplied by the prior probability $p(\theta)$ of the latent variable

$$P(\mathbf{y}_i, \theta|\boldsymbol{\gamma}) = \prod_{k=1}^{n} P(\theta)^{y_{ki}}[1 - P(\theta)]^{1-y_{ki}} p(\theta). \tag{2.5}$$

In IRT applications, it is customary to assume that the latent variable θ follows a standard normal distribution for parameter estimation, but this is not a requirement of the model. If we treat the item responses as fixed, the marginal likelihood function for all the item parameters in $\boldsymbol{\gamma}$, based on observed item response data can be expressed as:

$$L(\boldsymbol{\gamma}|\mathbf{Y}) = \prod_{i=1}^{N} \int \prod_{k=1}^{n} P(\theta)^{y_{ki}}[1 - P(\theta)]^{1-y_{ki}} p(\theta) d\theta. \tag{2.6}$$

Because the marginal likelihood given in Equation 2.6 does not depend on the unobserved θ values, it may be referred to as the *observed* data likelihood.

If we treat the item responses as fixed once observed, and also suppose the latent variable scores were observed, we can write the *complete* data likelihood as

$$L(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{n} P(\theta_i)^{y_{ki}}[1 - P(\theta_i)]^{1-y_{ki}} p(\theta_i), \tag{2.7}$$

where $\boldsymbol{\theta}$ is a vector containing all of the individual latent variable scores.

FIML estimation is typically done using either joint maximum likelihood (JML),

conditional maximum likelihood (CML), or marginal maximum likelihood (MML) procedures (Yen & Fitzpatrick, 2006). Of these, Bock and Aitkin's (1981) marginal maximum likelihood (MML) procedure is the most widely-used FIML estimation method, and involves the Expectation-Maximization (EM) algorithm. Bock and Aitkin (1981) approximate the marginal probability by replacing the integration with a summation over a set of Q quadrature points

$$P(\mathbf{y}_i;\boldsymbol{\gamma}) = \sum_{q=1}^{Q}\prod_{k=1}^{n} P(\theta_q)^{y_{ki}}[1-P(\theta_q)]^{1-y_{ki}}W_q, \tag{2.8}$$

where θ_q is a quadrature point, and W_q is the corresponding weight. If we assume that the quadrature points are defined at .1 intervals from -6 to 6, the weights can be estimated as set of normalized ordinates of the quadrature points from the standard normal population distribution

$$W_q = \frac{\phi(\theta_q)}{\sum_{q=1}^{Q}\phi(\theta_q)}. \tag{2.9}$$

In this approach, the latent variables are integrated out so that inferences are made based on the marginal likelihood function. The Bock-Aitkin approach uses the Expectation-Maximization (EM) algorithm developed by Dempster, Laird, and Rubin (1977). Bock-Aitkin alternates between the following two steps from a set of initial parameter estimates, say $\hat{\boldsymbol{\gamma}}^{(0)}$, and it generates a sequence of parameter estimates that converges under some very general conditions to the maximum likelihood estimate (MLE) of $\boldsymbol{\gamma}$ as the number of cycles b tends to infinity.

E-step. Given $\hat{\boldsymbol{\gamma}}^{(b)}$, evaluate the conditional expected complete data log-likelihood.

M-step. Maximize the conditional expected proportion of individuals to yield updated parameter estimates $\hat{\boldsymbol{\gamma}}^{(b+1)}$. Go back to E-step and repeat.

When the estimates from adjacent cycles have stabilized, the estimation process terminates.

2.2.2 Markov Chain Monte Carlo (MCMC) Estimation

Full information estimation may also be performed in a Bayesian framework, in which a prior distribution is assumed for the parameters being estimated. Bayesian estimation methods are based on Bayes' theorem, which states that the prior information combines with the information from the test ($\mathbf{Y}$) to produce the posterior distribution. The posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ given the observed data under the 2PL model is

$$P(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{Y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{\gamma})P(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int_{\boldsymbol{\theta}} \int_{\boldsymbol{\gamma}} p(\boldsymbol{\theta})p(\boldsymbol{\gamma})P(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\gamma})d\boldsymbol{\theta}d\boldsymbol{\gamma}}. \tag{2.10}$$

Bayesian estimation is conducted using Markov chain Monte Carlo (MCMC) estimation. MCMC is an iterative method that produces samples ("chain") that are drawn from each parameter's posterior distribution. The most commonly used MCMC estimation methods for IRT are the Metropolis-Hastings algorithm and Gibbs sampling (Albert, 1992; Patz & Junker, 1999). Edwards (2010) provides an overview of both algorithms within the context of item response theory, as well as a comparison of these estimation methods with MML-EM.

CHAPTER 3

Posterior Predictive Model Checking (PPMC) Method

The posterior predictive model checking (PPMC) method is a widely-used Bayesian model-checking method because it has a strong theoretical basis but is also easy to implement when conducting Bayesian estimation (Sinharay et al., 2006). The method primarily consists of comparing the observed data with replicated data (those predicted by the model) using a number of discrepancy measures. Systematic discrepancies between the observed data set and the replicated data sets indicate that the hypothesize model is failing to explain specific aspects of the observed data.

The observed data is referred to as $\mathbf{y}^{obs}$, while the replicated data is denoted $\mathbf{y}^{rep}$. In the Bayesian framework, we define the likelihood $p(\mathbf{y}^{obs}|\boldsymbol{\gamma})$ for a model applied to the data $\mathbf{y}^{obs}$, where $\boldsymbol{\gamma}$ denotes the set of parameters in the model. Let $p(\boldsymbol{\gamma})$ denote the prior distribution on the parameters. The posterior distribution of the parameters given hypothesized model H is

$$p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H) \propto p(\mathbf{y}^{obs}|\boldsymbol{\gamma}, H)p(\boldsymbol{\gamma}). \tag{3.1}$$

The hypothetical replicated data can be simulated from the posterior predictive distribution, which is the conditional distribution of the replicated data $\mathbf{y}^{rep}$ given the observed data $\mathbf{y}^{obs}$ and the model being tested, H. The general form of the

posterior predictive distribution is

$$p(\mathbf{y}^{rep}|\mathbf{y}^{obs}, H) = \int p(\mathbf{y}^{rep}|\boldsymbol{\gamma}, H)p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)d\boldsymbol{\gamma}. \tag{3.2}$$

In Equation 3.2, there are two components: (a) $p(\mathbf{y}^{rep}|\boldsymbol{\gamma}, H)$: the sampling distribution of the replicated data given the hypothesized model and parameters, and (b) $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$: posterior distribution of the model parameters under a given model for the observed data. By integrating over the unknown model parameters $\boldsymbol{\gamma}$, the entire posterior distribution of the model's parameters is accounted for in the model fit checking procedures.

A graphical overview of the PPMC procedure can be seen in Figure 3.1. On the left, we see the posterior distribution of the parameters, from which a large number (L) of sets of plausible parameters $\boldsymbol{\gamma}^1, ..., \boldsymbol{\gamma}^L$ are drawn. For each plausible parameter vector $\boldsymbol{\gamma}^\ell$, we draw one hypothetical replicate data set $\mathbf{y}^{rep,\ell}$, from the sampling distribution $p(\mathbf{y}^{rep,\ell}|\boldsymbol{\gamma}, H)$. We then have L pairs of draws from the joint posterior distribution of $\mathbf{y}^{rep}$ and $\boldsymbol{\gamma}$.

In order to measure the degree to which the observed data $\mathbf{y}^{obs}$ and the replicated data $\mathbf{y}^{rep}$ are discrepant, test quantities $T(\mathbf{y}, \boldsymbol{\gamma})$, also known as discrepancy measures, are defined to examine an aspect of discrepancy that is of interest. The arguments of the discrepancy measure are a data set (observed or replicated) and the model parameters. The test quantity $T(\mathbf{y}^{obs}, \boldsymbol{\gamma})$, where the data is fixed to the observed values but $\boldsymbol{\gamma}$ is draw from the parameter posterior $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$, is referred to as the *realized* test quantity. The *predictive* test quantity, $T(\mathbf{y}^{rep}, \boldsymbol{\gamma})$, is based on the joint posterior distribution of the replicated data $\mathbf{y}^{rep}$ and $\boldsymbol{\gamma}$. The *realized* and *predictive* test quantities are compared, with any significant difference between the test quantities indicating a failure of the model.

The PPMC method is primarily used to do graphical checks to compare the observed and replicated discrepancy measures. Figure 3.2 is an example scat-

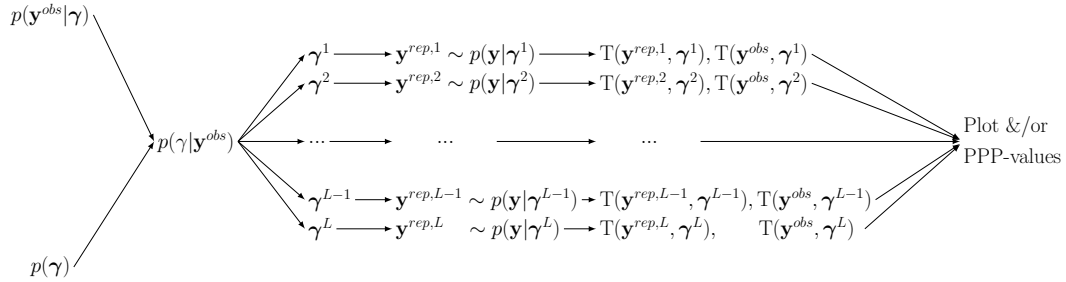


Figure 3.1: Graph describing the Posterior Predictive Model-Checking (PPMC) method, adapted from Sinharay et al. (2006)

terplot of the predictive test quantities (on the Y-axis) against the realized test quantities (on the X-axis). For correctly specified models, the points are expected to be spread evenly on either side of the 45-degree line, as seen in this figure. In the case of a model that fails to explain key features of the data, we would expect to see most or all of the points below the 45-degree line.

Additionally, a Bayesian counterpart of the classical p-value, known as the posterior predictive p-value, can be found by calculating the tail area probability of the distribution in Equation 3.3

$$P[T(\mathbf{y}^{rep}, \boldsymbol{\gamma}) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma})|\mathbf{y}^{obs}, H] = \int_{T(\mathbf{y}^{rep}, \boldsymbol{\gamma}) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma})} p(\mathbf{y}^{rep}|\boldsymbol{\gamma}, H) p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H) d\mathbf{y}^{rep} d\boldsymbol{\gamma}. \tag{3.3}$$

That is, the Bayesian posterior predictive p-value is defined as the probability that the replicated data, $\mathbf{y}^{rep}$, are more extreme than the observed data, $\mathbf{y}^{obs}$. When the fitted model is correct, the posterior predictive p-values (PPP-values) are not necessarily uniformly distributed, and previous researchers have found the PPP-values for a correct model tend to be closer to 0.5 (Sinharay & Stern, 2003). Extremely small probabilities (less than .05 or greater than .95) provide a strong suggestion that the model is not capturing the necessary features of the data.

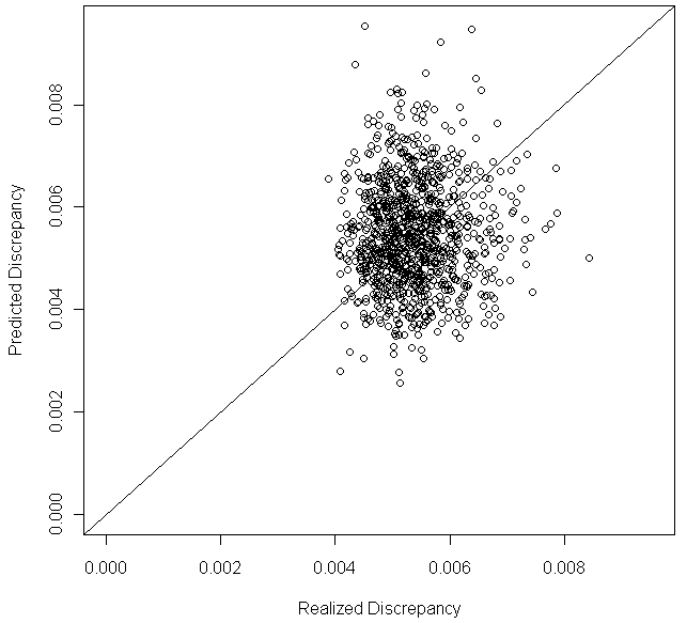


Figure 3.2: Example PPMC plot

3.1 Alternatives to Posterior Distribution in Predictive Model Checking

Various alternatives have been proposed to the parameter posterior distribution $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$ in the construction of the posterior predictive distribution (Gelfand, 1996). Box (1980) suggested the use of the *prior* distribution (prior predictive model checking), whereas Bayarri and Berger (2000) proposed the use of the *conditional posterior* distribution (partial or conditional posterior predictive model checking method).

Levy (2011) has previously discussed the similarities between posterior predictive, prior predictive, and partial predictive model checking. Furthermore, Lee et al. (2016) described a general expression under which all of the previously suggested predictive distributions can be combined. Building off of this general expression, these authors made use a well-known result in Bayesian literature that

asymptotically the likelihood tends to dominate the posterior shape. Furthermore, for large samples, the estimated item parameters are approximately normally distributed around the ML estimates, with the inverse of the Fisher information matrix as the covariance matrix (Gelman et al., 2003). Therefore, a normal approximation to the posterior was suggested for predictive model checking.

3.2 Posterior Predictive Model Checking - Normality Assumption

A predictive model checking method that relies on a posterior normality assumption was proposed by Lee et al. (2016), who referred to as a Poor Person's PPMC method. In place of $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$ in Equation 3.2, this method uses the multivariate normal distribution with its mean vector equal to the MLE $\hat{\boldsymbol{\gamma}}$ and dispersion matrix equal to the asymptotic covariance matrix of the maximum likelihood estimate $\hat{\mathbf{V}}$

$$\varphi_d(\boldsymbol{\gamma}) = |2\pi\hat{\mathbf{V}}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})'\hat{\mathbf{V}}^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})\right\}, \tag{3.4}$$

where $\varphi_d(\boldsymbol{\gamma})$ is the multivariate normal distribution with d dimensions. That is to say, this proposed posterior predictive model checking approach relies on posterior normality assumption to construct an approximate posterior predictive distribution using only by-products of full-information maximum likelihood estimation. For brevity, this proposed approach is subsequently referred to PPMC-N, where the "N" denotes the posterior normality assumption.

This approach does not require conducting Bayesian data analysis, but still incorporates parameter uncertainty in model assessment. Figure 3.3 provides a graphical depiction of the PPMC-N procedure. The only modification from the PPMC procedure that is displayed in Figure 3.1 comes on the left-most side,

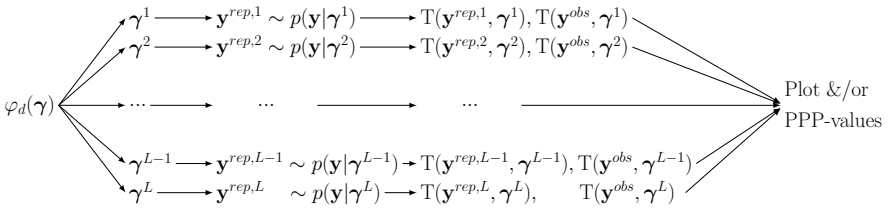


Figure 3.3: Graph describing the Posterior Predictive Model Checking - Normality approximation (PPMC-N) method

where the parameter posterior distribution is replaced with the multivariate normal approximation $\varphi_d(\boldsymbol{\gamma})$. The subsequent steps of drawing replicate datasets and estimating realized and predictive test quantities remain the same.

Using the predictive and realized test quantities estimated with the PPMC-N approach, the proportion of draws for which the predictive test quantity exceeds its corresponding realized test quantity can be calculated. These PPMC-N approximated Bayesian p-values are called PPMC-N *predictive* p-values:

$$\textit{predictive } p\text{-value} \approx \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}\left\{T(\mathbf{y}^{rep,\ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\right\}, \tag{3.5}$$

where $\mathbf{1}\left\{T(\mathbf{y}^{rep,\ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\right\}$ is an indicator function that takes a value of 1 only if $\left\{T(\mathbf{y}^{rep,\ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\right\}$ is true. This formulation is equivalent to counting the proportions of points lying above the 45-degree line in the scatterplot of the predictive and realized test quantities (such as Figure 3.2).

3.3 Posterior Predictive Model Checking - Normality Assumption with Re-sampling

A concern with the PPMC-N approach is that it is dependent upon a multivariate normality (MVN) assumption for the posterior of the item parameters

$p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$. It is not clear how realistic this assumption is with real data. The Sampling Importance Re-sampling (SIR) approach (Smith & Gelfand, 1992) is proposed as a potential method to address this concern. Importance sampling is a Monte Carlo method where a target distribution is approximated by a weighted average of random draws from another distribution. In this scenario, the idea is to represent the posterior density function by a set of random samples with associated weights from the normality approximation of the posterior.

The SIR approach necessitates the evaluation of the likelihood function given in Equation 2.7 based on the current set of $\boldsymbol{\gamma}^1, ..., \boldsymbol{\gamma}^L$ draws. For the PPMC-N method, the SIR approach breaks the posterior predictive simulations into several steps:

1. Draw L samples from multivariate normal distribution centered at MLE $\hat{\boldsymbol{\gamma}}$ and dispersion $\hat{\mathbf{V}}$ equal to inverse Fisher information matrix.

2. Evaluate the shape of the distribution and compute importance ratios.

   (a) Evaluate the likelihood $L(\boldsymbol{\gamma}^\ell|\mathbf{y}^{obs})$ at each of the L draws.

   (b) Evaluate the multivariate normal density $\varphi_d(\boldsymbol{\gamma}^\ell)$ at each of the L draws.

   (c) Calculate the ratio at each of the L draws.

$$\omega_\ell = \frac{L(\boldsymbol{\gamma}^\ell|\mathbf{y}^{obs})}{\varphi_d(\boldsymbol{\gamma}^\ell)} \tag{3.6}$$

   (d) Compute the L importance ratios and normalize

$$q_\ell = \frac{\omega_\ell}{\sum_{m=1}^{L} \omega_m}. \tag{3.7}$$

3. Draw a new set $\{\boldsymbol{\gamma}^{1*}, ..., \boldsymbol{\gamma}^{L*}\}$ from the discrete distribution over $\{\boldsymbol{\gamma}^1, ..., \boldsymbol{\gamma}^L\}$ placing mass q_ℓ on $\boldsymbol{\gamma}_\ell$. For each re-sampled parameter draw, produce a predictive draw of replicate data.

4. Estimate the predictive and realized test quantities.

In this approach, the likelihood $L(\boldsymbol{\gamma}^{\ell}|\mathbf{y}^{obs})$ is estimated in the numerator of Equation 3.6 in place of the posterior $p(\boldsymbol{\gamma}^{\ell}|\mathbf{y}^{obs})$. If a uniform prior $p(\boldsymbol{\gamma})$ is used, the posterior density and likelihood are proportional.

3.4 Discrepancy Measures

As was described previously, a set of functions called discrepancy measures are defined to capture the discrepancy between data and the model. A test quantity $T(\mathbf{y}, \boldsymbol{\gamma})$, or measure of discrepancy, is a function of both the data and the parameters. The discrepancy measures should be chosen to reflect important features of the model and reveals useful information about when model assumptions are violated. Levy (2006) highlighted various bivariate discrepancy measures that could be potentially used to detect multidimensionality in IRT models, which are described below.

Discrepancy measures that examine the pairwise association of items can be used to detect local item dependence. Many of the bivariate discrepancy measures involve observed and expected frequencies of item responses for two items (k and k') at a time (e.g., two-way tables for the frequencies of y_k and $y_{k'}$). The observed bivariate table for a dichotomous item is

		y_k = 1	y_k = 0
$y_{k'}$	1	n_{11}	n_{10}
	0	n_{10}	n_{00}

where n_{11} is the number of examinees with a correct response for both y_k and $y_{k'}$. The expected bivariate table is

		y_k	
		1	0
$y_{k'}$	1	$E(n_{11})$	$E(n_{10})$
	0	$E(n_{10})$	$E(n_{00})$

where $E(n_{10})$ is the expected frequency of a correct response to y_k and an incorrect response to $y_{k'}$ predicted by the item response theory model. Due to assumptions of local independence, $E(n_{10})$ can be estimated as

$$E(n_{10}) = N \int P_k(\theta) \left[1 - P_{k'}(\theta)\right] p(\theta) d\theta, \tag{3.8}$$

where $P_k(\theta)$ is the traceline for item k, $P_{k'}(\theta)$ is the traceline for item k', $p(\theta)$ is the population distribution, and N is the total number of examinees. This integral can be approximated numerically using a set of quadrature points.

The χ^2 and G^2 discrepancy measures for item-pairs (Chen & Thissen, 1997) are given, respectively, by

$$\begin{aligned} \chi^2_{kk'} &= \sum_{c=0}^{1} \sum_{c'=0}^{1} \frac{(n_{cc'} - E(n_{cc'}))^2}{E(n_{cc'})}, \\ G^2_{kk'} &= -2 \sum_{c=0}^{1} \sum_{c'=0}^{1} n_{cc'} ln \frac{E(n_{cc'})}{n_{cc'}}. \end{aligned} \tag{3.9}$$

Additionally, Yen's (1984, 1993) Q_3 statistic measures the correlation between a pair of items after accounting for the latent traits. The deviation between the observed and expected response for the ith examinee is calculated as

$$d_{ki} = y_{ki} - E_{y_{ki}}, \tag{3.10}$$

where y_{ki} is the observed response to item k and $E_{y_{ki}}$ is that examinee's expected response to item k. For dichotomous items, the expected response can be calculated as $E_{y_{ki}} = P(y_{ki} = 1 | \theta_i, a_k, c_k)$. The local dependence index Q_3 is computed

as the Pearson product-moment correlation between the deviation scores of items k and k'.

The model-based covariance (MBC) discrepancy measure, which also examines deviation between observed and expected responses, was proposed by Reckase (1997)

$$\text{MBC}_{kk'} = \frac{\sum_{i=1}^{N} \left(y_{ki} - E_{y_{ki}}\right)\left(y_{k'i} - E_{y_{k'i}}\right)}{N}. \tag{3.11}$$

For both Q_3 and Reckase's residual measure, a point estimate $\hat{\theta}$ is needed for each examinee. The *expected a posteriori* (EAP), or posterior mean, is a commonly used Bayesian estimator in IRT (Bock & Mislevy, 1982). The EAP estimator is calculated by taking the expectation over an individual's posterior distribution (e.g., product of the response pattern likelihood with a population distribution $p(\theta)$)

$$\hat{\theta}_i = \int_{\theta} \theta f(\theta|\mathbf{y}_i, \boldsymbol{\gamma}) d\theta = \frac{1}{f(\mathbf{y}_i|\boldsymbol{\gamma})} \int_{\theta} \theta L(\mathbf{y}_i|\theta, \boldsymbol{\gamma}) p(\theta) d\theta. \tag{3.12}$$

Typically, a standard normal distribution is used for $p(\theta)$ when estimating $\hat{\theta}$ scores. The sample covariance is another possible discrepancy measure, and is given by

$$\text{COV}_{kk'} = \frac{\left[(n_{11})(n_{00}) - (n_{10})(n_{01})\right]}{N^2}. \tag{3.13}$$

Similarly, residual item correlations can be estimated, which compare the observed covariance with the expected covariance for an item pair

$$\text{ResidCOV}_{kk'} = \frac{\left[(n_{11})(n_{00}) - (n_{10})(n_{01})\right]}{N^2} - \frac{\left[E(n_{11})E(n_{00}) - E(n_{10})E(n_{01})\right]}{N^2}. \tag{3.14}$$

Lastly, two odds ratio measures can be used as discrepancy measures. The odds ratio on the (natural) log scale is

$$\ln(\text{OR}_{kk'}) = \ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right]. \tag{3.15}$$

The standardized log ratio residual dependence formula given by Chen and Thissen (1997) is

$$\text{STDLN}(\text{OR}_{kk'}) = \frac{\ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right] - \ln\left[\frac{E(n_{11})E(n_{00})}{E(n_{10})E(n_{01})}\right]}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}}. \tag{3.16}$$

3.5 Previous Research on Local Dependence with IRT Models

Within the field of PPMC research, Yen's Q_3, the model-based covariance, and odds-ratio measures have been found to be effective in detecting multidimensionality and local dependence among item responses with PPMC (Levy, Mislevy, & Sinharay, 2009; Sinharay et al., 2006; Levy, 2006). The χ^2 and G^2 discrepancy measures, which are widely used in model fit evaluation within FIML estimation, were not found to be very sensitive to LD. However, many factors were found to influence the ability of the PPMC approach to detect multidimensionality, including the strength of dependence of the items on the nuisance latent dimensions, proportion of multidimensional items, and sample size (Levy, 2006).

Houts and Edwards (2013) compared the relative performance of a range of LD indices within FIML estimation by estimating power and Type I error rate, and found that the G^2 displayed both high power and reasonable Type I error rates across most conditions.

The previous field of work has examined a range of discrepancy measures across a wide set of LD conditions, but the work has mostly been siloed within the separate maximum likelihood and Bayesian fields of research. This study borrows from the insights of Bayesian predictive model checking with the goal of improving the detection of LD within a maximum likelihood estimation framework. In order to accomplish this goal, it is necessary to first establish the PPMC-N method as a viable alternative to the PPMC approach, and then to compare the simulation-based

predictive model checking methods with classical frequentist model appraisal approaches that are calculated based solely on observed and expected responses (calculated based on best-fitting (ML) parameter estimates).

3.6 Research Questions

The goal of the current work is to investigate the ability of the three outlined predictive model checking approaches to detect the presence of inadequately modeled multidimensionality under several data-generating conditions. The present study uses a Monte Carlo study to answer the following three research questions:

- **Research Question 1:** How do the three predictive model checking approaches (PPMC, PPMC-N, and PPMC-N with SIR) compare in the detection of violations of unidimensionality within the 2PL item response theory model?
- **Research Question 2:** Are some discrepancy measures more sensitive to the presence of local dependence than others?
- **Research Question 3:** Does the PPMC-N approach detect local dependence in situations where the traditional frequentist approach for LD described by Chen and Thissen (1997) are not sensitive?

CHAPTER 4

Methods

The purpose of this thesis is to investigate different predictive model checking approaches as well as the effectiveness of different discrepancy measures to detect the model misspecification within the two-parameter logistic (2PL) item response theory model, specifically focusing on detecting violations of unidimensionality.

There are $3 \times 5 = 15$ data generating conditions in this simulation study, which are as follows:

1. Sample size: 250, 500, 1000
2. Type of local dependence: null condition (no LD), surface LD (SLD; mild or strong), and underlying LD (ULD; mild or strong).

Fifty replications are conducted within each condition. The three data-generating models (no LD, SLD, and ULD) considered in this simulation study are summarized in Table 4.1. In all of the conditions, there are 20 items. The path diagrams for the data-generating models can be seen in Figure 4.1. The analysis model in each case is the unidimensional model shown in Panel 4.1a.

4.1 Data Generation

True latent ability (θ) were generated from a standard normal distribution in the unidimensional and SLD conditions, and a multivariate normal distribution

Table 4.1: Conditions for the Monte Carlo study

Analysis Model	Data-generating Model	Condition Number	Violated Assumption
Unidimensional model	Unidimensional model	1	None
	Unidimensional model with correlated item pairs	2	Surface LD (SLD)
	Case 1: probability $\pi_{LD} = .5$		
	Case 2: probability $\pi_{LD} = .8$		
	Bifactor item response model	3	Underlying LD (ULD)
	Case 1: Minor specific factor slopes (slope ratio: .5)		
	Case 2: Strong specific factor slopes (slope ratio: 1.5)		

in third data-generating condition. Item discrimination parameters ranged from values of 1 to 2, and item intercepts were set to values of -1, -.5, .5, 1, or 1.5.

In the null condition, data were generated following a unidimensional model. Programs were written in R (version 3.1) to simulate data in the Null, SLD, and ULD conditions. Surface LD was generated following the deterministic process outlined in Section 2.1. The probability that the second item within a pair will have the same response, π_{LD}, is manipulated to vary the degree of LD. The two values of π_{LD} considered are .5 and .8. There are four LD item pairs in the surface LD condition, and the remaining 16 items are locally independent.

For the underlying LD condition, data are generated following a bifactor model (Holzinger & Swineford, 1939). The bifactor model contains a general factor that explains the overall variance among the items, as well as specific factors that account for residual variance. The specific factors are uncorrelated with each other and the general factor. Each specific factor contains five items, with a total of four specific factors. For items within each specific factor, the degree of LD was defined by the ratio of the general slope to the specific factor slopes. The specific factor slopes are either set to be half of the general slopes (Case 1), or set to be

1.5 times the general factor slopes (Case 2). The variances of all of the latent variables in the bifactor model are set to 1.

4.2 Maximum Likelihood Estimation and PPMC-N

For the two posterior predictive model checking approaches that rely on normality approximations, parameters are estimated using Bock-Aitken EM algorithm. A unidimensional model fit to each simulated dataset using flexMIRT® (Cai, 2015) (see Appendix 1 for an example calibration file). flexMIRT® reports convergence statistics that check whether the solution is possible local maximum. The estimated item parameters $\hat{\boldsymbol{\gamma}}$ are reported in the output file, and corresponding asymptotic covariance matrix of the item parameters is requested as a setting in calibration. Parameter standard errors are calculated using empirical cross-product approximation (Houts & Cai, 2015).

To assess the fit of the model using the PPMC-N approach, the following three steps are repeated $L = 100$ times for each calibrated model:

1. Generate a draw of parameters $\boldsymbol{\gamma}^{\ell}$ from the multivariate normal distribution given by Equation 3.4.

2. Given the item parameters draw $\boldsymbol{\gamma}^{\ell}$, a replicated data set $\mathbf{y}^{rep,\ell}$ is drawn from the sampling distribution under the hypothesized model H (e.g., $p(\mathbf{Y}|\boldsymbol{\gamma}^{\ell}, H)$).

3. Compute the predictive $T(\mathbf{y}^{rep,\ell}, \boldsymbol{\gamma}^{\ell})$ and realized $T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^{\ell})$ discrepancy values for each of the specified test quantities.

An R program was written to simulate random normal draws, generate replicate datasets, and calculate the various discrepancy measures. The PPMC-N *predictive* p-values defined in Equation 3.5 are recorded within each replication for all of the discrepancy measures listed in Section 3.4.

Additionally, PPMC-N-SIR *predictive* p-values are calculated for each simulated dataset using the item parameters estimated by flexMIRT. As outlined in Section 3.3, the Sampling Importance Resampling (SIR) method requires a small additional step to the PPMC-N approach. Following Step 1 in the PPMC-N approach, the complete data likelihood given in Equation 2.7 is estimated separately for each set of item parameters (given the observed data $\mathbf{y}^{obs}$). Weighted re-sampling from the L draws is conducted, and then steps 2 and 3 of the PPMC-N approach are resumed. For the SIR approach, an initial sample of $L = 1000$ parameter sets are drawn, and then using the re-weighting, a final set of $L = 100$ draws are used for the PPMC-N-SIR procedure.

4.3 Bayesian Estimation and PPMC

Lastly, a fully Bayesian IRT calibration is conducted using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). A unidimensional 2PL model is estimated in WinBUGS (see Appendix 1 for model code) with the following prior distributions

$$\theta_i \sim N(0.0, 1.0)$$
$$b_k \sim N(0.0, 1.0)$$
$$a_k \sim N_+(1.0, 1.0)$$

where $N_+(1.0, 1.0)$ denotes the normal distribution truncated to the positive real line. The threshold parameter b_k is estimated in the Bayesian model rather than the intercept c_k parameter, which can be easily converted back to the intercept to match flexMIRT's parameterization ($c_k = -a_k * b_k$).

R2WinBUGS, a package in R that calls WinBUGS, was used to call WinBUGS and store latent proficiency and parameter posterior estimates (Sturtz, Ligges, &

Gelman, 2005). A total of 2,000 iterative simulations were done per chain (3 total chains), the first 400 iterations were discarded, and the iterations were thinned by a value of 2. After discarding burn-in and thinning the chain, the resulting iterations were pooled to produce 2,400 iterations for use in PPMC. Convergence of these models were measured using the effective sample size (sample size adjusted for autocorrelation across simulations) and Gelman and Rubin (1992) convergence diagnostic (printed in R2WinBUGS as Rhat). Additionally, plots of the MCMC sequences are examined for the parameter and θ estimates.

$L = 100$ draws of the item parameters $\boldsymbol{\gamma}$ and the vector of latent proficiency estimates $\boldsymbol{\theta}$ were sampled from the posterior distribution. Based on these estimates, replicate datasets were drawn and the predictive p-values were estimated for each of the discrepancy measures.

4.4 Evaluating the Predictive Model Checking Approaches

First, parameter recovery was examined in the null condition to ensure both estimation methods are recovering the true parameter estimates. As a part of this examination, I examine plots of the marginal parameter posterior distributions, and check for the closeness of the normality approximations.

Three criteria were used to examine the PPMC, PPMC-N, and PPMC-N-SIR performance during the null condition (e.g., no model misfit). The median predictive p-value in the null condition was calculated for each discrepancy measure within each model checking approach. Line plots, such as those seen in Levy (2006), that compare the distribution of the predictive p-values across replications for each discrepancy function and predictive model checking approach were also created. Additionally, the proportion of extreme predictive p-values (e.g., p-values $< .05$ or p-values $> .95$) was estimated for each method.

For the surface and underlying LD conditions, median and proportion ex-

treme predictive p-value for the item pairs demonstrating local dependence were calculated for each discrepancy measure. Additionally, various plots were used to demonstrate the median predictive p-values for item pairs that are expected to contain local dependence due to the generating model compared to item pairs that are not.

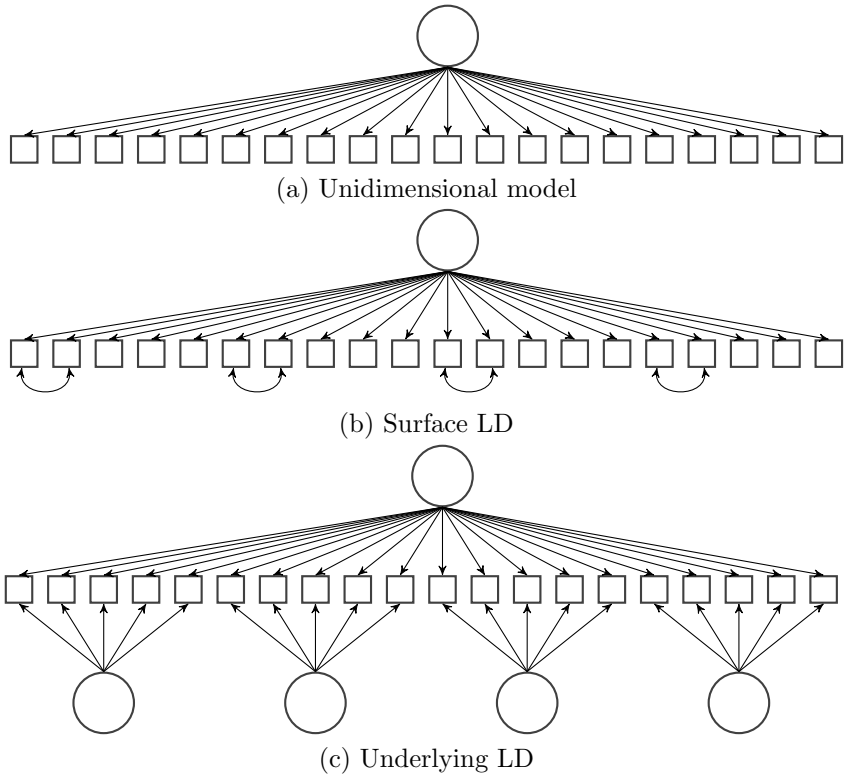
4.5 Comparing the Predictive Model Checking Approaches to Frequentist LD Approaches

Houts and Edwards (2013) have looked at the performance of LD indices using frequentist methods across a range of simulation conditions to test how sensitive these indices are to various forms of LD. These simulations focused on the Type I error rate and the power of the LD indices. Type I error rate is the incorrect classification of a non-locally dependent pair as being locally dependent. Power is defined as the correct identification of a LD item-pair by an LD index, and is reported as the average proportion of correctly flagged pairs out of the total number of replications. Values higher than 0.80 for power are generally considered acceptable. For indices with known distribution, observed Type 1 error rates at the nominal level are desired.

The third component of this study is to compare the performance of the predictive model checking approaches with the traditional frequentist approach for detecting LD. While the orientation of many PPMC researchers is to view the predictive model checking methods as diagnostic rather than as a test of data-model misfit (see, for example, Stern, 2000), researchers have previously conducted simulations that examine the Type I error and power of the Bayesian item fit measures for unidimensional IRT models (Sinharay, 2006). Therefore, I examine the Type I error rate and power of a set of LD indices under frequentist model appraisal methods and the PPMC-N method.

For this comparison, I focus on a subset of LD indices that have been commonly used in frequentist model appraisal: χ^2, G^2, and Q_3. These indices were chosen because they have been the focus of prior LD research, and because they are reported by existing IRT software. The equations for these discrepancy measures are given in Section 3.4. In the frequentist context, the χ^2 and G^2 were calculated based on observed and expected bivariate marginals. For dichotomous item pairs, χ^2 and G^2 asymptotically follow a chi-square distribution with 1 degree of freedom (df=1), and item pairs are flagged as locally dependent if the $\chi^2_{df=1}$ test is statistically significant at $\alpha = .05$ level. It does not follow a known distribution, and a cut-off of .2 is used to flag LD item pairs (Chen & Thissen, 1997).

For the predictive model checking method (PPMC-N), the χ^2, G^2, and Q_3 discrepancy measures were calculated as described in Section 4.2. A (two-tailed) hypothesis test with significance level $\alpha = .05$ is performed, where the null hypothesis of data-model fit is rejected if the p-value is less than $\alpha/2$ or is greater than $1 - \alpha/2$.



(a) Unidimensional model

(b) Surface LD

(c) Underlying LD

Figure 4.1: Data-generating IRT models

CHAPTER 5

Results

Across all of the conditions, all of the models estimated with flexMIRT converged to a possible local maximum. The unidimensional IRT model took an average of 0.9 seconds to estimate using flexMIRT, and the estimation time did not vary substantially by sample size. There were not convergence issues with the WinBUGS estimation across conditions. The average Gelman and Rubin convergence diagnostic (Rhat) across parameters and conditions was 1.00. The average estimation time was 156 seconds for $N = 250$, 320 seconds for $N = 500$, and 655 seconds for $N = 1,000$.

5.1 Comparing the Predictive Model Checking Approaches

The first research question is concerned with the comparability of three predictive model checking approaches (fully Bayesian PPMC, PPMC assuming posterior normality (PPMC-N), and PPMC-N with re-sampling method). I first present results examining the predictive model checking approaches under the null condition, before turning to the comparison across the misfit conditions.

5.1.1 Unidimensional Condition

Before turning to the results of the LD discrepancy measures, parameter recovery across the maximum likelihood and Bayesian estimation approaches were compared. There are three null conditions in which the data are unidimensional, cor-

responding to the three sample sizes ($N = 250, 500, 1000$). Plots of the true and estimated parameters across replications under the null condition from flexMIRT and WinBUGS for the $N = 250$ sample condition are shown in Figure 5.1. True values are plotted as red asterisks, ML estimates from flexMIRT are circles, and the parameter posterior mean estimates from WinBUGS are triangles. True parameters are well-recovered, with small variation in estimates across replications.

Figure 5.1: Comparison of parameter estimates with true (generating) parameters, null condition ($N = 250$)

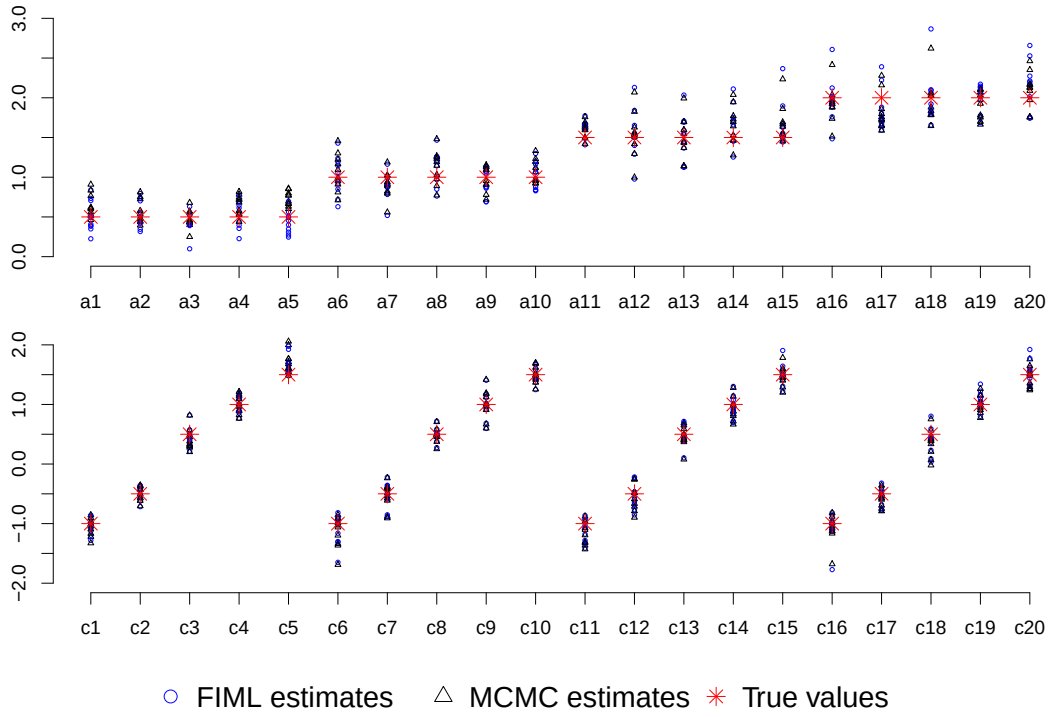


Figure 5.2 displays the estimated marginal posterior distributions and the two multivariate normal posterior approximations for the 20 item slope parameters for a single replication under the null condition ($N = 250$). The true generating item parameters are shown as a red asterisk on the x-axis, the estimated posterior density is plotted in black, the multivariate normal approximation using the MLE estimates and standard errors is displayed as a blue dashed line, and

the re-sampled density using the SIR approach is displayed as a red dotted line. Similarly, Figure 5.3 displays the estimated marginal posterior and the posterior approximations for the 20 item intercept parameters. As can be seen in these figures, the two posterior approximation distributions closely overlap with the marginal parameter posterior distributions estimated in WinBUGS, and all three approaches are generally doing an excellent job of recovering the true parameters. In the cases where the posterior and normal approximation are discrepant, such as the first and fifth slope parameter, the distributions for the SIR approach fall in between the marginal MVN approximation and the posterior distributions. Similar results are seen for the unidimensional model in the other sample size conditions, but the distributions are more narrow and peaked, demonstrating higher precision of the estimates. Given the results in these three figures, it is clear that maximum likelihood and Bayesian estimation methods are both accurately recovering the true parameters in the null condition.

The performance of the eight discrepancy measures in the null conditions was also examined. Given that the data were generated from a unidimensional model, all item-pairs reflect the same single latent dimension, and we can therefore pool results across item pairs following an exchangeability assumption (De Finetti, 1964). Given there are 20 items, there are $(20 * (20 - 1))/2 = 190$ item pairs. For the null condition, the predictive p-values from the 190 item-pairs for each bivariate discrepancy measure are pooled within a replication, and the results are then pooled across the 50 replications. Figure 5.4 contains eight panels, one for each discrepancy measure examined. In each panel, there are line plots for the distributions of the pooled p-values showing the relative frequency of each p-value (ranging from 0 to 1). The black line corresponds to the posterior predictive method, the blue dashed line corresponds to the PPMC-N method, and the red dotted line corresponds to the PPMC-N method with SIR. Figure 5.4 displays the results for the $N = 250$ condition, and Figure 5.5 displays the results for the

Figure 5.2: Comparison of three predictive model checking methods in the approximation of the item parameter marginal posterior distributions within a single replication, null condition ($N = 250$)

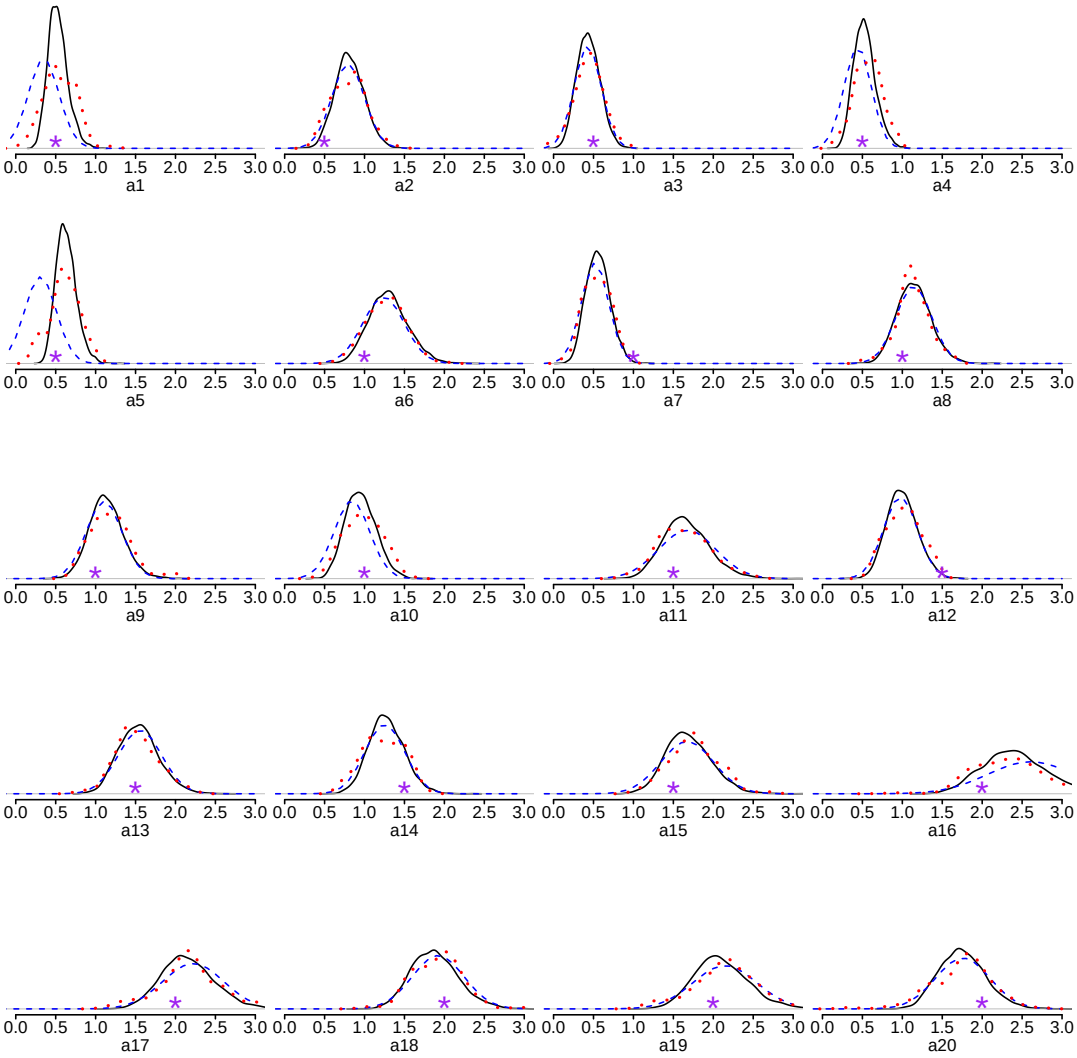


Figure 5.3: Comparison of three predictive model checking methods in the approximation of the item parameter marginal posterior distributions within a single replication, null condition ($N = 250$)

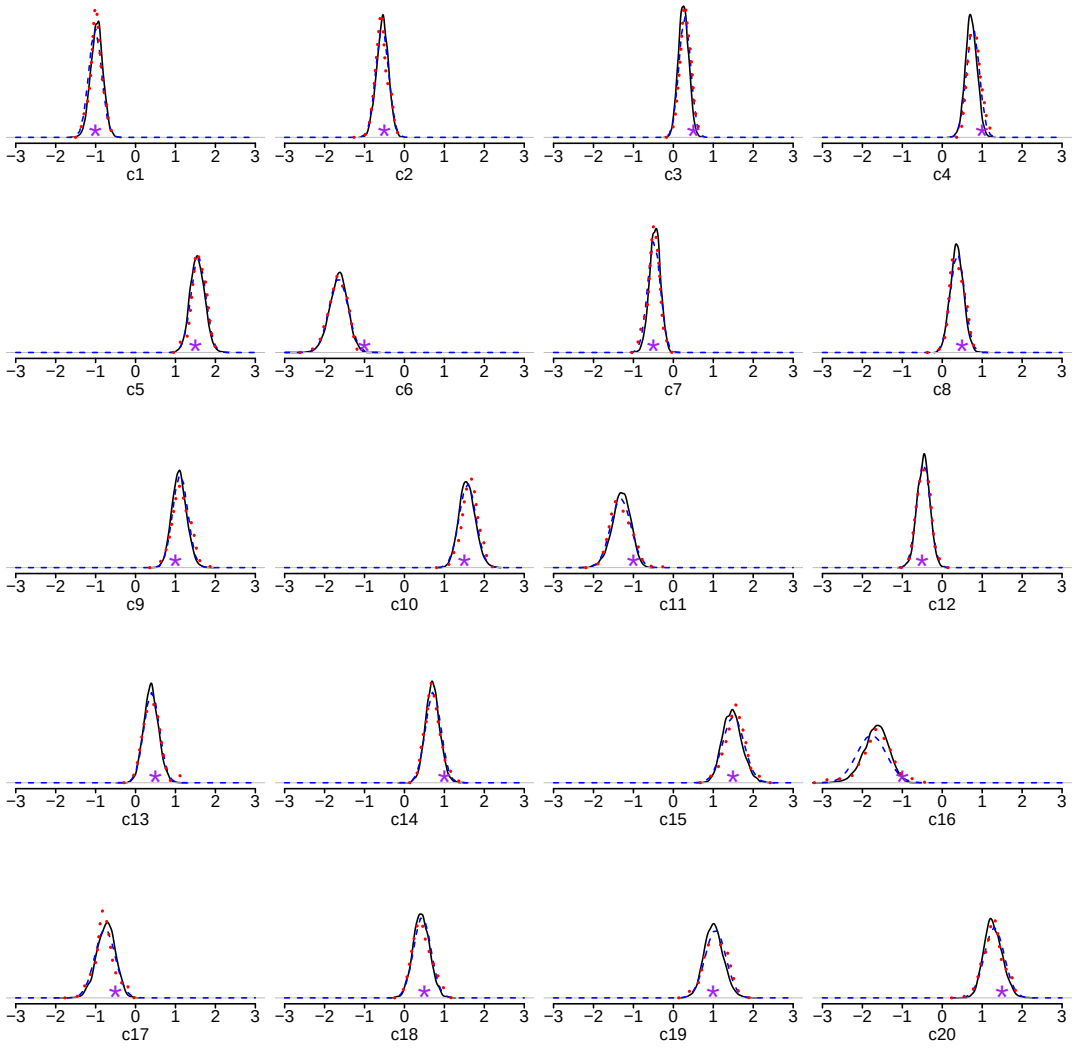
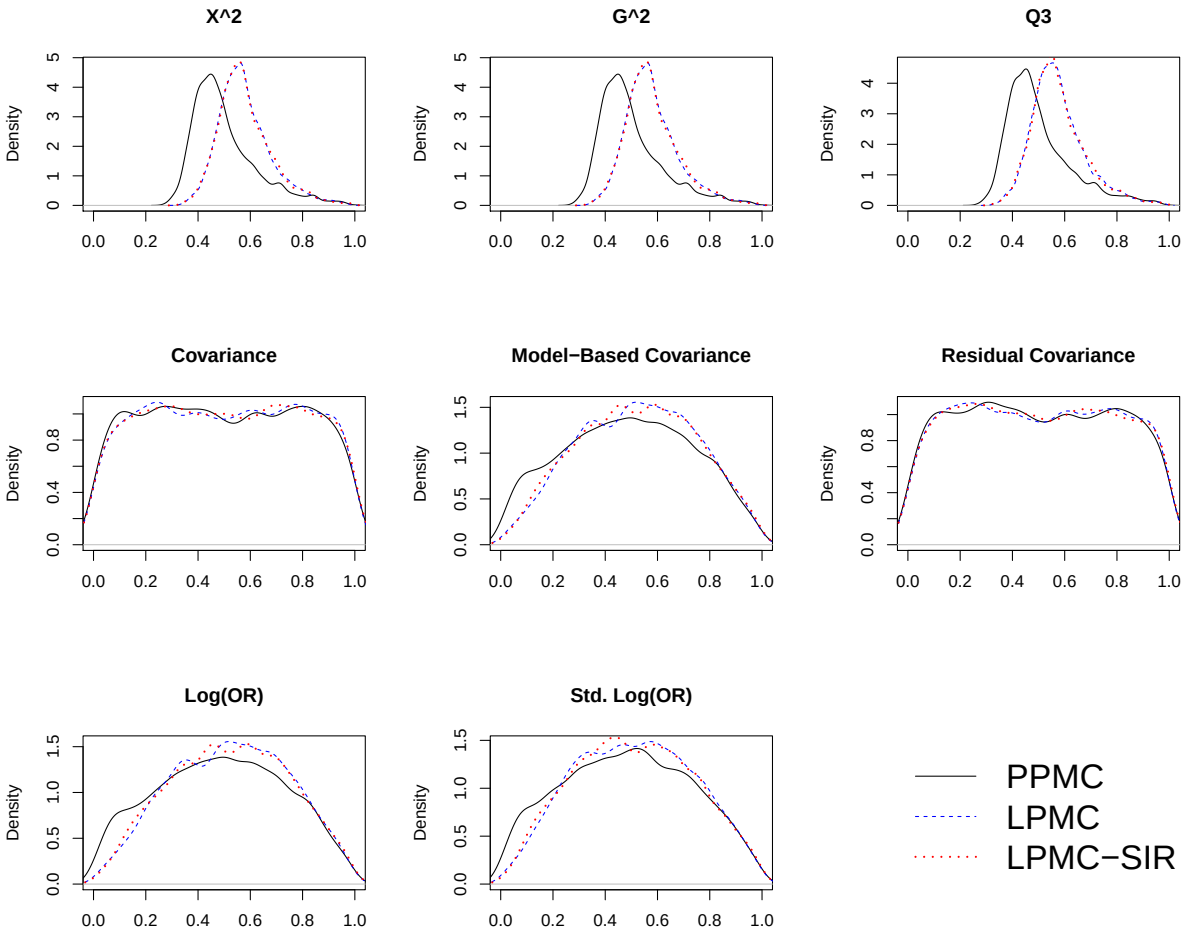


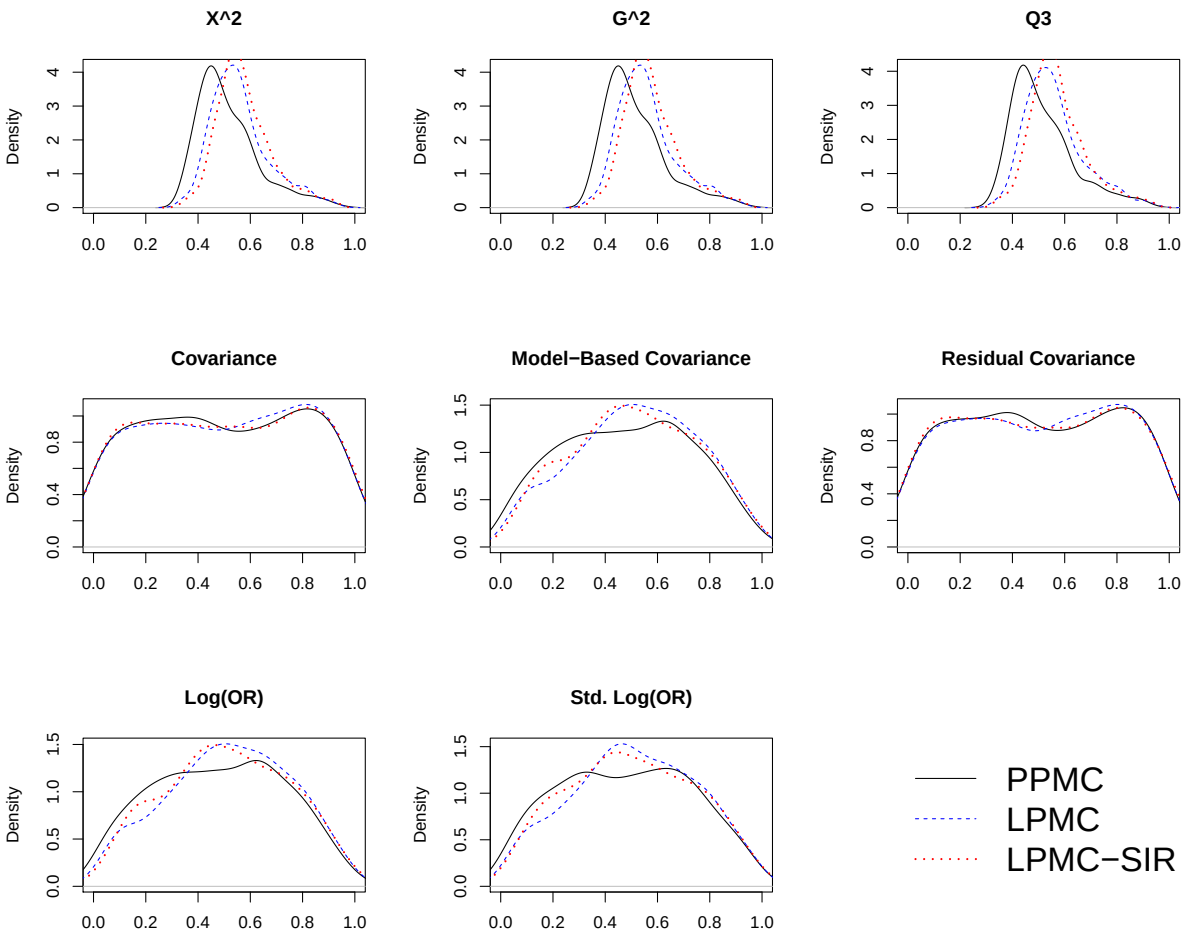
Figure 5.4: Comparison of three predictive model checking methods, null condition ($N = 250$)



$N = 1,000$ sample condition.

In both Figures 5.4 and 5.5, all the distributions of p-values for all three methods are symmetric around .5. The three predictive approaches produce mostly overlapping distributions, with the exception of χ^2, G^2, and Q_3. For these three discrepancy measures, the two PPMC-N approaches are shifted slightly closer to 1. As observed by others (Meng, 1994), the distributions of the predictive p-values for all of the discrepancy measures are less dispersed than a uniform distribution. This is seen particularly for χ^2, G^2, and Q_3, where the predictive p-values

Figure 5.5: Comparison of three predictive model checking methods, null condition ($N = 1000$)



are heavily concentrated between 0.3 and 0.7. The three covariance-based discrepancy measures and the two log-odds ratios perform similarly under the null condition across sample sizes.

Table 5.1: Median p-values for 8 discrepancy measures based on unidimensional data, null condition

PPMC method	Sample size	χ^2	G^2	$Q3$	Cov	MBC	Residual Cov.	Log(OR)	Std. Log(OR)
PPMC-N	250	0.57	0.57	0.56	0.51	0.53	0.50	0.53	0.51
PPMC-N-SIR	250	0.56	0.56	0.56	0.51	0.53	0.50	0.53	0.50
PPMC	250	0.47	0.47	0.47	0.49	0.49	0.48	0.49	0.48
PPMC-N	500	0.54	0.54	0.54	0.52	0.51	0.50	0.51	0.51
PPMC-N-SIR	500	0.54	0.54	0.54	0.50	0.51	0.50	0.51	0.50
PPMC	500	0.49	0.49	0.48	0.51	0.49	0.50	0.49	0.48
PPMC-N	1000	0.52	0.52	0.52	0.51	0.51	0.50	0.51	0.50
PPMC-N-SIR	1000	0.53	0.53	0.53	0.50	0.51	0.50	0.51	0.51
PPMC	1000	0.47	0.47	0.47	0.50	0.48	0.49	0.48	0.48

Table 5.1 shows the median predictive p-value within each predictive model checking approach across the three samples sizes. Within the null condition, median values around 0.5 are expected, which is what is observed across all of the discrepancy measures. The proportions of extreme predictive p-values for the discrepancy measures for each sample size are shown in Table 5.2. Extreme predictive p-values are defined as those below .05 or above .95. The results in Table 5.2 can be seen as Type I error rates using .05 and .95 as critical values (i.e., in a two-tailed test with $\alpha = .10$). The empirical Type I error rates for the covariance and residual covariance discrepancy measures are close to the nominal rate of .10. However, empirical Type I error rates for the other discrepancy measures are below .10, indicating that use of predictive p-values in hypothesis testing results in a conservative test. In summary, results from the null condition indicate very similar performance between the PPMC, PPMC-N, and PPMC-N-SIR approaches.

Table 5.2: Proportion of replications with extreme p-values (i.e., p-value $< .05$ or $> .95$) based on unidimensional data, null condition

PPMC method	Sample size	χ^2	G^2	$Q3$	Cov	MBC	Residual Cov.	Log(OR)	Std. Log(OR)
PPMC-N	250	0.00	0.00	0.00	0.08	0.02	0.08	0.02	0.02
PPMC-N-SIR	250	0.00	0.00	0.00	0.09	0.02	0.09	0.02	0.02
PPMC	250	0.00	0.00	0.00	0.08	0.04	0.08	0.04	0.04
PPMC-N	500	0.00	0.00	0.00	0.09	0.02	0.08	0.02	0.02
PPMC-N-SIR	500	0.00	0.00	0.00	0.09	0.02	0.09	0.02	0.02
PPMC	500	0.00	0.00	0.00	0.09	0.03	0.09	0.03	0.03
PPMC-N	1000	0.00	0.00	0.00	0.10	0.02	0.10	0.02	0.02
PPMC-N-SIR	1000	0.00	0.00	0.00	0.10	0.02	0.09	0.02	0.03
PPMC	1000	0.00	0.00	0.00	0.10	0.04	0.10	0.04	0.04

5.1.2 Local Dependence (LD) Conditions

We now turn to the results of the predictive model checking approaches where LD is present in a subset of items. As with the unidimensional data, predictive p-values obtained from each discrepancy measure on multiple item-pairs are pooled following exchangeability assumptions, as well as across the 50 replications within conditions. However, the p-values are pooled separately depending on whether the item-pair contains LD in the generating data structure (misfit item-pairs), or is locally independent (non-misfit item-pairs). In the LD conditions, predictive p-values near .5 indicate that the predictive model checking approaches have not detected the LD within item-pairs.

Figures 5.6 and 5.7 present the median predictive p-values across the three sample sizes for the Surface LD condition for item pairs that display LD and do not display LD, respectively. The three lines represent the three model checking approaches (PPMC, PPMC-N, and PPMC-N-SIR). The median predictive p-values for the misfit items in the strong SLD condition ($\pi_{LD} = .8$), which can be seen in the right panel of Figure 5.6, are equal to 1 for all of the discrepancy measures and across the three model checking approaches. For the mild SLD condition (left panel), none of the discrepancy measures appear sensitive to the

local dependence (median values between 0.4 and 0.5). The non-misfit item-pairs in the mild and strong surface LD conditions showed median predictive p-values that ranged between 0.4 and 0.6, which is to be expected. This indicates that the predictive model checking approaches are able to clearly distinguish between misfit and non-misfit items in the surface LD case.

Similarly, Figures 5.8 and 5.9 present the median predictive p-values for the Underlying LD condition for item pairs that display LD (misfit) and do not display LD (non-misfit), respectively. The results for misfit items under the strong Underlying LD ($a = 1.5$) condition mirror the strong SLD condition, where the median predictive p value is close or equal to 1 for all of the plotted discrepancy measures. However, in the mild ULD condition, the covariance and log-odds ratio discrepancy measures had median predictive p-values that are closer to 1, indicating that the discrepancy measures have somewhat succeeded in detecting LD (particularly for the large sample sizes). The median predictive p-values for the non-misfit items are displayed in Figure 5.9. The mild ULD condition displays median predictive p-values around .5, correctly indicating lack of local dependence. However, in the strong ULD case, the median predictive p-values stray from .5 in the larger sample sizes, indicating that the misfit in the model is affecting the LD indices for all of the items.

Based on both the unidimensional and the local dependence conditions, it is clear that the PPMC, PPMC-N, and PPMC-N-SIR approaches are performing similarly across sample sizes and across discrepancy measures. We now turn to comparing the performance of the various discrepancy measures in detecting LD.

5.2 Comparing the Discrepancy Measures

Table 5.3 displays the median predictive p-values for the misfit items for all of the eight examined discrepancy measures across the four LD conditions. The first

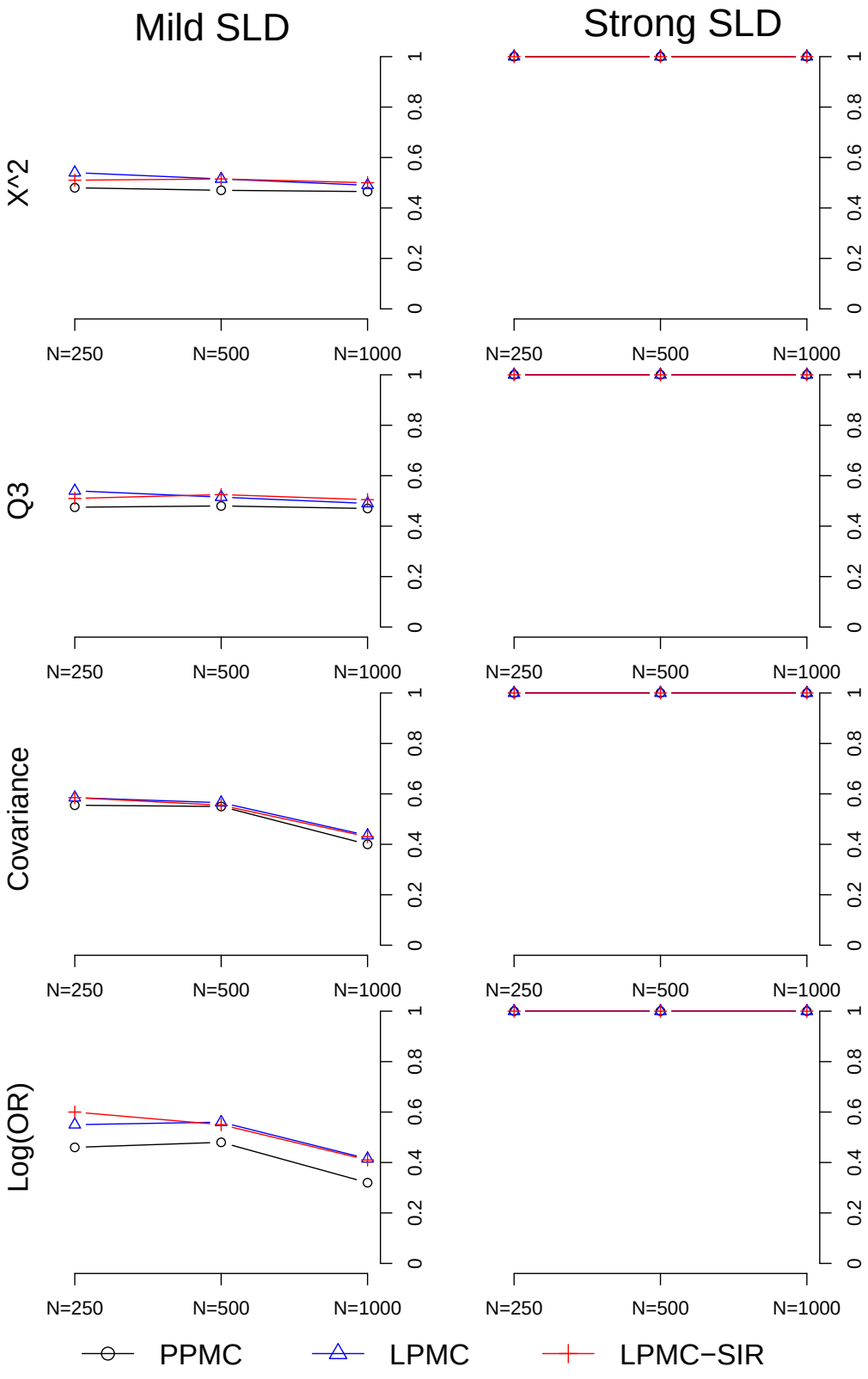


Figure 5.6: Comparison of the median predictive p-values among MISFIT items for the a set of discrepancy measure, Surface Local Dependence (SLD) conditions

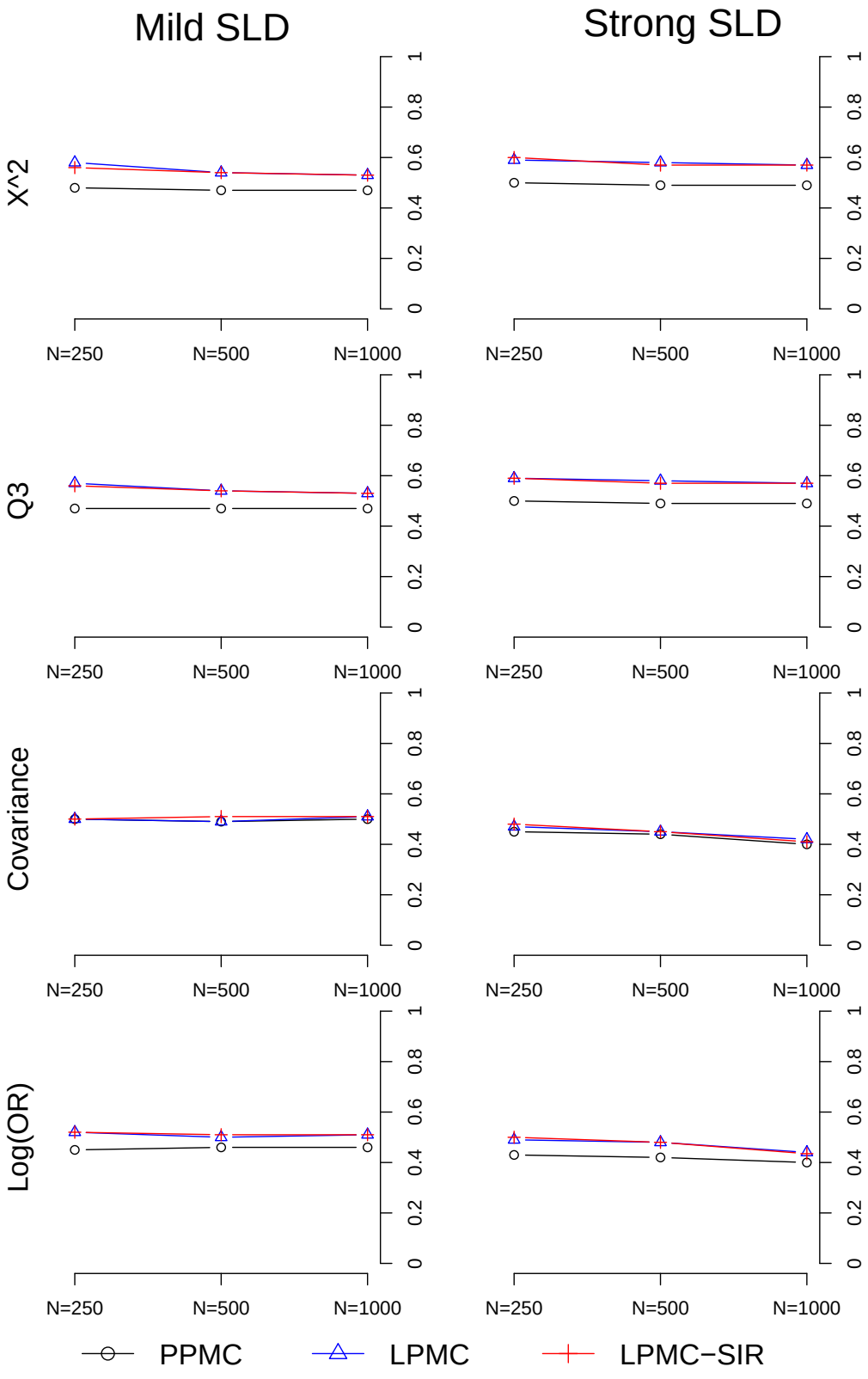


Figure 5.7: Comparison of the median predictive p-values among NON-MISFIT items for the a set of discrepancy measure, Surface Local Dependence (SLD) conditions

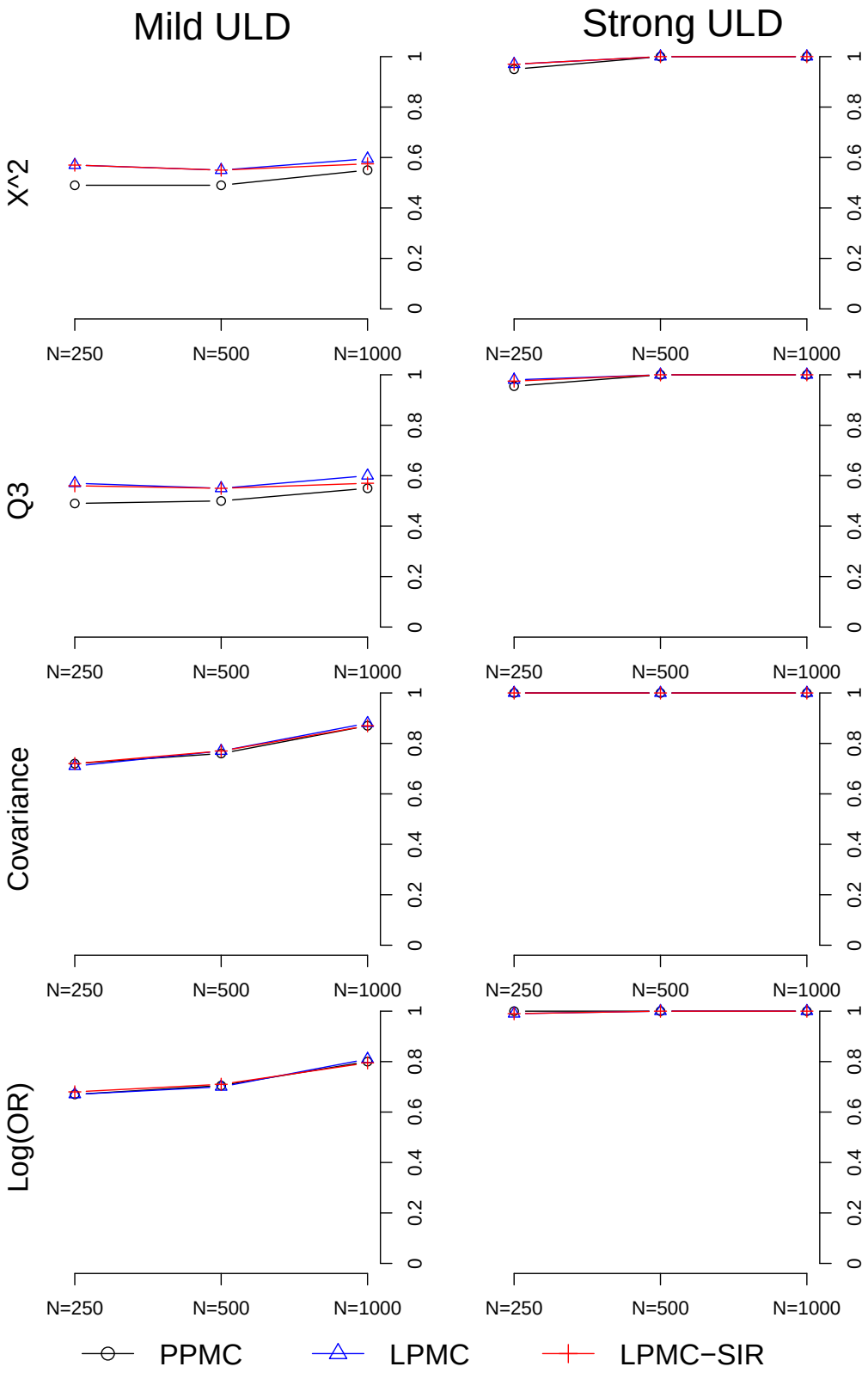


Figure 5.8: Comparison of the median predictive p-values among MISFIT items for the a set of discrepancy measure, Underlying Local Dependence (ULD) conditions

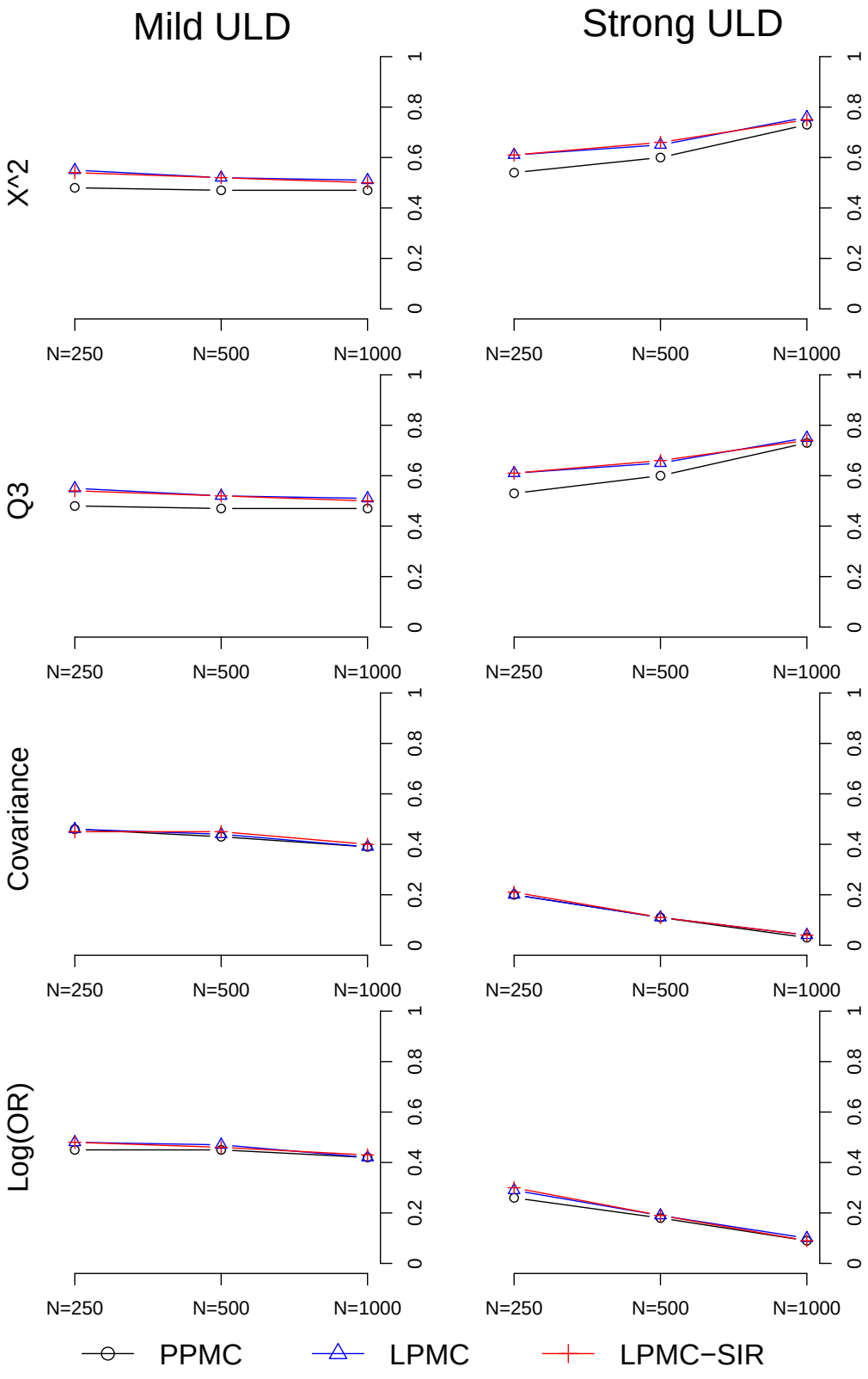


Figure 5.9: Comparison of the median predictive p-values among NON-MISFIT items for the a set of discrepancy measure, Underlying Local Dependence (ULD) conditions

two columns list the type of LD misspecification and the predictive model checking method. The performance of the various discrepancy measures within a LD condition are consistent across sample size and model checking approach. Results within a condition also appear fairly similar across the discrepancy measures, indicating little is gained by using all eight different discrepancy measures. The notable exception is that in the mild conditions, the median values of the χ^2, G^2, and Q_3 discrepancy measures are quite close to one another (and are frequently equal) and generally the lowest. The results for the covariance, model-based covariance (MBC), and the residual covariance are quite similar, and are typically the highest median p-values.

Table 5.4 displays the proportion of replications with extreme p-values for the misfit items (p-values $< .05$ or $> .95$) for all of the 8 studied discrepancy measures. Again, the performances of the different discrepancy measures within a LD condition are consistent across sample size and model checking approach. The covariance and the residual covariance discrepancy measures demonstrate the highest proportion extreme p-values within the misfit items, indicating they are most sensitive to LD misfit.

To compare the various discrepancy measures, it is also useful to compare the performance within single replications. In Figure 5.10, the estimated predictive p-values for four different discrepancy measures are plotted for the mild and strong ULD condition for item-pairs where violations of local independence are to be expected. In the data generating model for ULD, the first five items are nested in the first specific factor (see Figure 4.1c). Therefore, a total of 10 item-pairs are of interest within each specific factor: item 1 with 2, 3, 4, and 5; item 2 with 3, 4, and 5; item 3 with 4 and 5; and item 4 with 5. The estimated p-values are plotted in the same order for each of the four specific factors. Each specific factor consists of five items, and p-values are computed for the 10 item pairs.

Figure 5.10a displays the p-values for the forty total items of interest in the

Table 5.3: Median p-values for 8 discrepancy measures for the misfit items within the four LD conditions

LD condition	PPMC method	Sample size	χ^2	G^2	$Q3$	Cov	MBC	Residual Cov	Log (OR)	Std. Log(OR)
SLD - mild	PPMC-N	250	0.54	0.54	0.54	0.59	0.55	0.58	0.55	0.55
	PPMC-N-SIR	250	0.51	0.51	0.51	0.59	0.60	0.60	0.60	0.60
	PPMC	250	0.48	0.48	0.48	0.56	0.46	0.57	0.46	0.46
	PPMC-N	500	0.52	0.52	0.52	0.57	0.56	0.57	0.56	0.56
	PPMC-N-SIR	500	0.52	0.52	0.53	0.56	0.55	0.56	0.55	0.55
	PPMC	500	0.47	0.47	0.48	0.55	0.48	0.55	0.48	0.48
	PPMC-N	1000	0.49	0.49	0.49	0.44	0.42	0.43	0.42	0.42
	PPMC-N-SIR	1000	0.50	0.50	0.51	0.43	0.41	0.43	0.41	0.41
	PPMC	1000	0.47	0.47	0.47	0.40	0.32	0.40	0.32	0.32
SLD - strong	PPMC-N	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ULD - mild	PPMC-N	250	0.57	0.57	0.57	0.71	0.67	0.70	0.67	0.67
	PPMC-N-SIR	250	0.57	0.57	0.56	0.72	0.68	0.72	0.68	0.66
	PPMC	250	0.49	0.49	0.49	0.72	0.67	0.71	0.67	0.68
	PPMC-N	500	0.55	0.55	0.55	0.77	0.70	0.77	0.70	0.70
	PPMC-N-SIR	500	0.55	0.55	0.55	0.77	0.71	0.77	0.71	0.71
	PPMC	500	0.49	0.49	0.50	0.76	0.71	0.76	0.71	0.71
	PPMC-N	1000	0.60	0.60	0.60	0.88	0.81	0.88	0.81	0.80
	PPMC-N-SIR	1000	0.58	0.58	0.57	0.87	0.80	0.87	0.80	0.79
	PPMC	1000	0.55	0.55	0.55	0.87	0.80	0.87	0.80	0.81
ULD - strong	PPMC-N	250	0.97	0.97	0.98	1.00	0.99	1.00	0.99	0.99
	PPMC-N-SIR	250	0.97	0.97	0.98	1.00	0.99	1.00	0.99	1.00
	PPMC	250	0.95	0.95	0.96	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

mild underlying LD condition. Q_3 and the covariance discrepancy measure have p-values that are furthest from .5, and appear to display the most sensitivity to misfit. The χ^2 and G^2 are consistently the least sensitive. Figure 5.10b displays the p-values in the strong ULD condition, where all of the item p-values are very close or equal to 1 for every discrepancy measure.

In summary, all of the discrepancy measures behavior similarly in the strong LD conditions, while the covariance measures appear to be the most sensitive in the mild LD conditions.

5.3 Comparing the Predictive and Frequentist Approaches for LD Detection

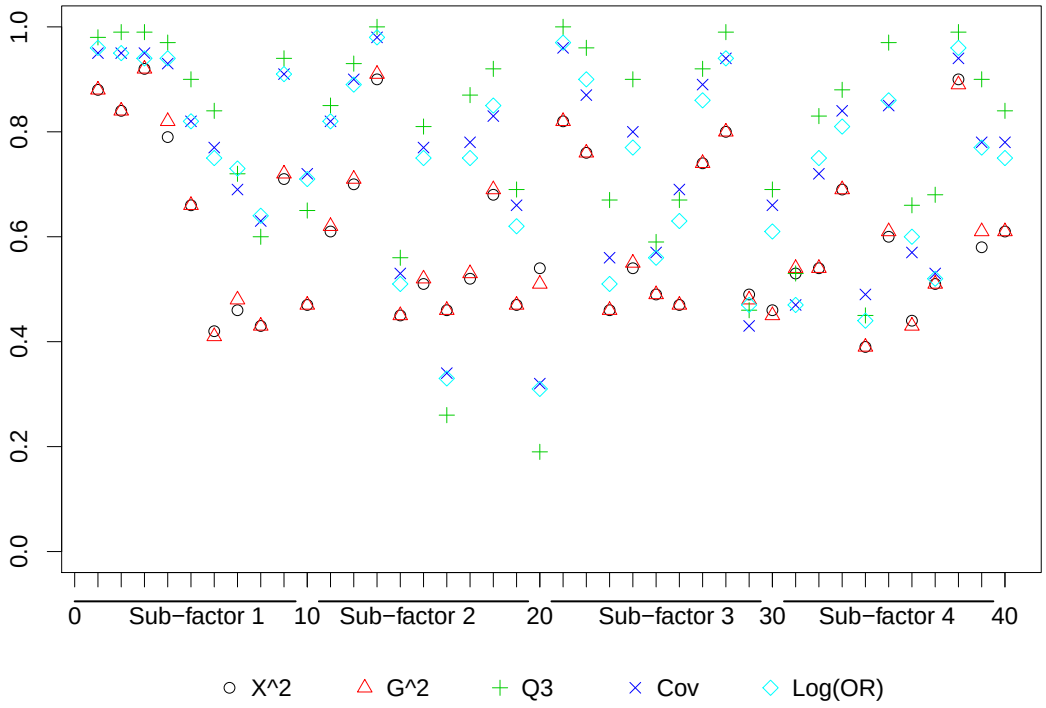
The results from the third component of this study compare the performance of the predictive model checking approaches with classical frequentist approach in terms of power and Type I error rate. In discussing these results, rules of thumb for power (0.80) and Type I error rates (0.05) are used.

Table 5.5 displays the Type I error rate for all of the model conditions (null and LD conditions), broken down by sample size and model-checking method. The overall Type I error rates in the null condition are low for both the frequentist and PPMC-N approaches. Within the PPMC-N method, only the Q_3 displays error rates at the nominal level. For the misfit conditions, the Type I error rates are low across the board for the χ^2 and G^2 discrepancy measures, with the exception of the ULD ($a = 1.5$) condition. For the frequentist LD approach, the Type I error rates are particularly high in this condition, implying that these measures are flagging a larger percentage of non-misfit items as displaying LD. A similar pattern of high Type I error is seen within the same ULD condition for PPMC-N method and Q_3 discrepancy measure.

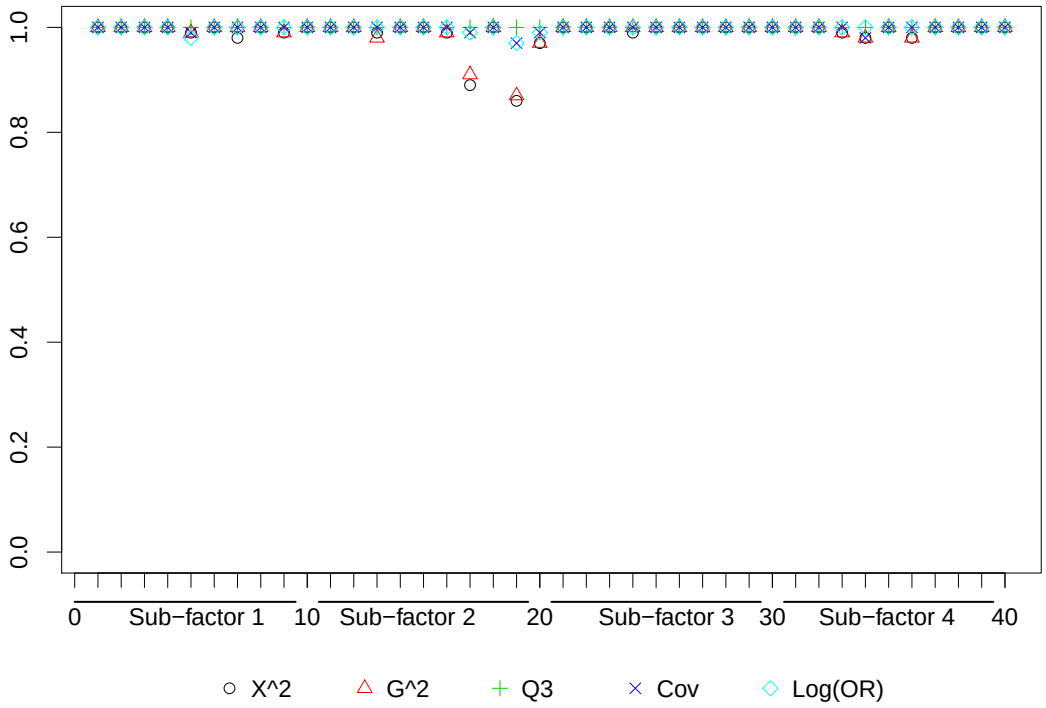
Table 5.6 displays the power for the LD conditions, broken down by sample

Table 5.4: Proportion of replications with extreme p-values (i.e., p-value $< .05$ or $> .95$) for the misfit items within the four LD conditions

LD condition	PPMC method	Sample size	χ^2	G^2	$Q3$	Cov	MBC	Residual Cov	Log (OR)	Std. Log(OR)
SLD - mild	PPMC-N	250	0.00	0.00	0.00	0.13	0.05	0.15	0.05	0.05
	PPMC-N-SIR	250	0.00	0.00	0.00	0.08	0.05	0.13	0.05	0.05
	PPMC	250	0.00	0.00	0.00	0.10	0.15	0.10	0.15	0.15
	PPMC-N	500	0.00	0.00	0.00	0.15	0.13	0.15	0.13	0.13
	PPMC-N-SIR	500	0.00	0.00	0.00	0.18	0.13	0.15	0.13	0.10
	PPMC	500	0.03	0.03	0.03	0.15	0.13	0.15	0.13	0.10
	PPMC-N	1000	0.00	0.00	0.00	0.15	0.10	0.10	0.10	0.05
	PPMC-N-SIR	1000	0.00	0.00	0.00	0.13	0.08	0.13	0.08	0.08
	PPMC	1000	0.00	0.00	0.00	0.13	0.08	0.13	0.08	0.08
SLD - strong	PPMC-N	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ULD - mild	PPMC-N	250	0.01	0.01	0.01	0.15	0.05	0.14	0.05	0.06
	PPMC-N-SIR	250	0.00	0.00	0.00	0.16	0.04	0.15	0.04	0.05
	PPMC	250	0.00	0.00	0.00	0.14	0.05	0.14	0.05	0.04
	PPMC-N	500	0.02	0.02	0.01	0.16	0.08	0.17	0.08	0.07
	PPMC-N-SIR	500	0.02	0.02	0.02	0.20	0.07	0.19	0.07	0.08
	PPMC	500	0.01	0.01	0.01	0.18	0.07	0.18	0.07	0.08
	PPMC-N	1000	0.03	0.03	0.03	0.29	0.13	0.28	0.13	0.13
	PPMC-N-SIR	1000	0.02	0.02	0.03	0.28	0.14	0.28	0.14	0.14
	PPMC	1000	0.02	0.02	0.02	0.30	0.13	0.30	0.13	0.15
ULD - strong	PPMC-N	250	0.63	0.63	0.66	0.94	0.85	0.94	0.85	0.85
	PPMC-N-SIR	250	0.60	0.60	0.62	0.93	0.84	0.93	0.84	0.84
	PPMC	250	0.47	0.47	0.50	0.94	0.87	0.94	0.87	0.87
	PPMC-N	500	0.94	0.94	0.94	1.00	0.99	1.00	0.99	0.99
	PPMC-N-SIR	500	0.95	0.95	0.95	1.00	0.98	1.00	0.98	0.99
	PPMC	500	0.90	0.90	0.91	1.00	0.98	1.00	0.98	0.98
	PPMC-N	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00



(a) Mild ULD



(b) Strong ULD

Figure 5.10: Comparison of predictive model checking p-values for the a set of discrepancy measures for a single replication, Underlying Local Dependence (ULD) condition ($N = 500$)

size and model-checking method. The predictive (PPMC-N) and frequentist LD approaches perform very similarly, with very low levels of power in the mild LD conditions and very high power in the strong LD conditions. The only notable difference in the performance of the Q_3 discrepancy measure between the PPMC-N and frequentist method. The Q_3 measure performs well in the PPMC-N method, but has lower power throughout the conditions in using the frequentist measure.

Based on these results, it appears that the PPMC-N and frequentist method have fairly similar power and Type I error rates for these three focal discrepancy measures.

Table 5.5: Type I error rates for the null and misfit conditions for the PPMC-N and frequentist approaches

		N=250			N=500			N=1000		
Method	Degree of LD	χ^2	G^2	Q_3	χ^2	G^2	Q_3	χ^2	G^2	Q_3
PPMC-N method	None	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.05
	SLD									
	$\pi = .5$	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.05
	$\pi = .8$	0.00	0.00	0.06	0.00	0.00	0.07	0.00	0.00	0.08
	ULD									
	a =0.5	0.00	0.00	0.05	0.00	0.00	0.06	0.00	0.00	0.06
	a =1.5	0.00	0.00	0.13	0.01	0.01	0.26	0.04	0.04	0.44
Frequentist method	None	0.02	0.02	0.00	0.02	0.02	0.00	0.01	0.02	0.00
	SLD									
	$\pi = .5$	0.01	0.02	0.00	0.01	0.02	0.00	0.01	0.01	0.00
	$\pi = .8$	0.02	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.00
	ULD									
	a =0.5	0.02	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.00
	a =1.5	0.07	0.06	0.00	0.15	0.15	0.00	0.32	0.32	0.00

Table 5.6: Power for the misfit conditions for the PPMC-N and frequentist approaches

		N=250			N=500			N=1000		
Method	Degree of LD	χ^2	G^2	Q_3	χ^2	G^2	Q_3	χ^2	G^2	Q_3
PPMC-N method	SLD									
	$\pi = .5$	0.00	0.00	0.05	0.00	0.00	0.08	0.00	0.00	0.05
	$\pi = .8$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	ULD									
	a =0.5	0.00	0.00	0.08	0.01	0.01	0.11	0.01	0.01	0.21
	a =1.5	0.47	0.49	0.90	0.88	0.90	0.99	1.00	1.00	1.00
Frequentist method	SLD									
	$\pi = .5$	0.01	0.01	0.00	0.06	0.06	0.00	0.03	0.03	0.00
	$\pi = .8$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	ULD									
	a =0.5	0.03	0.03	0.00	0.06	0.07	0.00	0.10	0.11	0.00
	a =1.5	0.83	0.84	0.34	0.98	0.98	0.32	1.00	1.00	0.31

CHAPTER 6

Summary and Conclusions

The primary goal of this study was to demonstrate the performance of the Posterior Predictive Model Checking assuming posterior normality (PPMC-N) method for the detection of model misfit (and more specifically, local dependence) in Item Response Theory (IRT). The examination of local dependence (LD) in IRT has a long history in both the frequentist (e.g., Yen, 1984; Chen & Thissen, 1997; Houts & Edwards, 2013) and Bayesian literature (e.g., Levy, 2006; Sinharay et al., 2006). The PPMC-N method for IRT models is proposed to take advantage of the flexibility of the previously-studied Posterior Predictive Model Checking (PPMC) approach (that is used in tandem with Bayesian estimation) when maximum likelihood estimation methods are used. The advantages of the PPMC method over frequentist approaches include the fact that the PPMC methods integrates parameter uncertainty into model fit assessment through the use of the posterior predictive distribution, as well as the fact that it does not rely on asymptotically-defined distributions for the discrepancy statistics.

Using a set of discrepancy measures that have been previously identified for the detection of multidimensionality in IRT, this study compared the performance (in terms of the detection of LD) of the PPMC method with two normality approximations: (a) the PPMC-N method, in which an the posterior predictive distribution is approximated using a multivariate normal distribution that is centered around the maximum likelihood estimates of the parameters, and (b) the PPMC-N-SIR method, which relies on an additional Sampling Importance Re-sampling step to

decrease the reliance on the multivariate normal approximation of the posterior. The three predictive model checking methods were found to perform very similarly across a range of LD conditions and sample sizes.

The second research question asked whether any of the studied bivariate discrepancy measures are more useful than others in detecting LD. The studied LD indices included χ^2, G^2, Q_3, sample covariance, model-based covariance, residual covariance, log-odds ratio, and the standardized log-odds ratio. All of the discrepancy measures performed similarly in the strong LD conditions, showing excellent detection of the LD items. The most effective measures in the mild LD conditions were the covariance, model-based covariance, and residual covariance, which performed almost identically. Additionally, in the null condition, these three discrepancy measures displayed distributions of predictive p-values that are closest to uniform.

Lastly, I compared the PPMC-N method with the frequentist approach, focusing on the power and Type I error rates of each method using a subset of the most commonly-used discrepancy measures. The two approaches were found to be fairly similar, with high power in the strong LD conditions and low power in the mild LD conditions. This may seem to imply that nothing is gained by using a predictive method over frequentist approaches with this specific set of discrepancy measures. There are two important notes regarding this comparison. First, it is already well-known that PPMC predictive p-values lead to conservative inferences, due to the fact that the data is used in predictive model checking for both estimation and model checking (Sinharay et al., 2006). This has led many researchers to treat the predictive p-values as pieces of for data-model (mis)fit, rather than for use in significance tests. Secondly, we do see evidence that the best performing discrepancy measures within the predictive model checking approach (specifically, the covariance-based measures) do show some sensitivity to LD in the mild LD conditions, where the commonly-used discrepancy measures in the fre-

quentist framework (χ^2, G^2, and Q_3) were not sensitive. These covariance-based measures are suggested for further study within the PPMC-N framework.

The results from this simulation study are generalizable only to the extent that the design variables are similar to real world data conditions. Sample sizes were chosen to be reflective of small to large samples that may be used in IRT calibration and model appraisal, but the simulated data had no missing responses, which is likely in real-world applications. Further work is necessary to examine the PPMC-N approach with shorter item banks and other item response models (e.g., graded response model for polytomous item responses).

The purpose of this study is to provide additional ways for applied users of IRT, who typically rely on maximum likelihood methods of estimation available in widely-used software, to detect violations of model assumptions. However, subjective judgments on the part of the user are still required when deciding whether the discrepancies between the data and hypothesized are important enough to address. When misfit is detected, possible actions include (a) retaining the hypothesized model to make inferences but acknowledging the limitations and source of misfit, (b) discarding locally dependent items and re-testing the model with the revised data, or (c) using a more general model that accounts for the extra dimensionality (such as a bifactor or testlet model). These decisions remain as an important piece of model appraisal, which can be best made with strong information regarding the specific aspects of the model that are discrepant with the data. The PPMC-N method is a flexible alternative to both frequentist and Bayesian model checking methods, and retains many of the advantages of the fully Bayesian approach while remaining a viable options for those who choose the use maximum likelihood estimation methods.

APPENDIX A

Program files

Example flexMIRT calibration file:

```
<Project>
   Title = "Null condition calibration file (N=250, n=20)";
   Description = "True Dim=1; Estimated Dim=1";

<Options>
   Mode = Calibration;
   Algorithm = BAEM;
   Quadrature = 121,6.0;
   Score = EAP;
   SaveSCO = YES;
   SavePRM = Yes;
   SaveCov = Yes;

<Groups>
%G%
   File = "2PL_UniD_5_data.dat";
   Missing = 9;
   Varnames = v1-v20;
   N = 250;
   Select = v1-v20;
```

```
  Model(v1-v20) = Graded(2);
  Ncats(v1-v20) = 2;
  Dimensions = 1;

<Constraints>
```

Example WinBUGS calibration file:

```
model{
  for (i in 1:N){
    for (k in 1:n){
      Y[i, k] ~ dbern(prob[i, k])
      logit(prob[i, k]) <- a[k]*(theta[i] - b[k])
    }
    theta[i] ~ dnorm(0.0, 1.0)
  }
  ## Priors on item parameters
  for (k in 1:n){
    b[k] ~ dnorm(m.b, precision.b)
    a[k] ~ dnorm(m.a, precision.a) I(0, )
  }
  precision.b <- pow(s.b, -2)
  precision.a <- pow(s.a, -2)
}
```

Bibliography

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association*, *95*(452), 1127–1142.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Bolt, D. M. (2005). Limited-and full-information estimation of item response theory models. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Earlbaum.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383–430.

Cai, L. (2015). *flexMIRT version 3: Flexible multilevel multidimensional item analysis and test scoring.* Seattle, WA: Vector Psychometric Group.

Cai, L., & Thissen, D. (2015). Modern approaches to parameter estimation in item response theory. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment.* New York: Routledge.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using

item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

De Finetti, B. (1964). Foresight: its logical laws in subjective sources. In H. E. Smokler & H. E. Kyburg (Eds.), *Studies in Subjective Probability* (pp. 93–158).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38.

Edwards, M. C. (2010). A markov chain monte carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.

Fox, J.-P. (2010). *Bayesian Item Response Modeling.* New York, NY: Springer New York.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice.* CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, *6*(4), 733–760.

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

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (pp. 1–19). Springer.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: the stability of a bi-factor solution. *Supplementary Educational Monographs*.

Houts, C. R., & Cai, L. (2015). *flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring User's Manual Version 3.0* (Computer soft-

ware manual). Vector Psychometric Group, LLC.

Houts, C. R., & Edwards, M. C. (2013, October). The Performance of Local Dependence Measures With Psychological Data. *Applied Psychological Measurement*, *37*(7), 541–562.

Ip, E. H.-s. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109–132.

Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2).

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and bayesian networks* (Unpublished doctoral dissertation). University of Maryland, College Park, College Park, MD.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663–685.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*(7), 519–537.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325–337.

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

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics*, *11*(1), 3–31.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366.

Reckase, M. D. (1997). A Linear Logistic Multidimensional Model for Dichotomous Item Response Data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271–286). Springer New York.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321.

Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, *111*(1), 209–221.

Sireci, S. G., Thissen, D., & Wainer, H. (1991, September). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Smith, A. F., & Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, *46*(2), 84–88.

Stern, H. S. (2000). Asymptotic Distribution of P Values in Composite Null Models: Comment. *Journal of the American Statistical Association*, *95*(452), 1157–1159.

Sturtz, S., Ligges, U., & Gelman, A. E. (2005). R2winbugs: A package for running WinBUGS from R. *Journal of Statistical software*, *12*(3), 1–16.

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report (Research Memorandum 922)* (Tech. Rep.). Chapel Hill, NC: LL Thurstone Laboratory, University of North Carolina at Chapel Hill.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating

performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 111–153). Westport, CT: American Council on Education and Praeger.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002, December). Identification and Evaluation of Local Item Dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291–309.