Posterior Predictive Assessment for Item Response Theory Models: A Proposal Based on the Hellinger Distance

Valutazione Predittiva A Posteriori per i Modelli di Item Response Theory: Una Proposta Basata sulla Distanza di Hellinger

Mariagiulia Matteucci and Stefania Mignani

Abstract Bayesian posterior predictive assessment has received considerable attention for investigating specific aspects of fit of item response theory models. In fact, this approach is easy to apply within Markov chain Monte Carlo estimation, it is flexible and free from distributional assumptions. In its classical implementation, the method is based on graphical analysis and the estimation of posterior predictive *p*-values to investigate the degree to which observed data are expected under the model, given a discrepancy measure. In this work, we propose to quantify the distance between the realized and the predictive distributions of the discrepancy measure based on the Hellinger distance. The results show that this measure is able to provide clear recommendations about the investigated aspects of model fit.

Abstract *Lo studio di aspetti specifici dell'adattamento dei modelli di item response theory è stato affrontato di recente con successo in ambito bayesiano usando strumenti della valutazione predittiva a posteriori. Questo approccio infatti è di facile applicazione quando si utilizza il metodo Markov chain Monte Carlo, è flessibile e non dipende da assunzioni distributive. Nella sua implementazione classica, il metodo si basa sull'analisi grafica e sulla stima dei p-value predittivi a posteriori basati su una particolare misura di discrepanza. In questo lavoro, si propone di quantificare la distanza tra la distribuzione realizzata e quella predittiva della misura di discrepanza utilizzando la distanza di Hellinger. I risultati mostrano che questa misura di distanza è in grado di fornire indicazioni chiare circa i particolari aspetti dell'adattamento considerati.*

[1] Mariagiulia Matteucci, University of Bologna; email: m.matteucci@unibo.it
Stefania Mignani, University of Bologna; email: stefania.mignani@unibo.it

Key words: posterior predictive model checks, item response theory models, Hellinger distance.

1 Introduction

In educational and psychological measurement, item response theory (IRT) models (see, e.g., van der Linden and Hambleton, 1997) are commonly used to estimate the characteristics of both the categorical items and the test takers. Several IRT unidimensional and multidimensional models have been proposed to account for different data structures. While unidimensional models assume the presence of a single latent variable underlying the response process, the multidimensional ones allow for multiple abilities. In this setting, the issue of model goodness-of-fit is crucial to investigate both absolute and relative fit.

Due to the increasing model complexity, a considerable amount of literature has been recently focused on Bayesian estimation of IRT models via Markov chain Monte Carlo (MCMC) methods due to its flexibility. Starting from a MCMC output, one possibility for examining model fit is using Bayesian posterior predictive model checks (PPMC; Rubin, 1984). Considerable advantages of the method are that it does not rely on distributional assumptions, and it is relatively easy to implement, given that the entire posterior distribution of all parameters of interest is obtained through MCMC algorithms.

The first proposals on the use of PPMC for IRT models deal with differential item functioning, person fit, fit of unidimensional models and item fit (see, e.g., Sinharay, 2006). Later, there was an increasing interest in checking specifically for the behavior of unidimensional models fitted to potential multidimensional data (see, among others, Sinharay, Johnson, and Stern, 2006; Levy, Mislevy, and Sinharay, 2009; Levy and Svetina, 2011). In these studies, PPMC has been implemented with graphical analyses and the estimation of the posterior predictive *p*-values (PPP-values) to investigate the degree to which observed data are expected under the model, given a discrepancy measure. Moreover, Wu, Yuen, and Leung (2014) proposed the use of relative entropy (RE) within PPMC to quantify the information the realized distribution loses when it is approximated by the predictive distribution.

The aim of this study is to propose the Hellinger distance, based on the Hellinger integral (Hellinger, 1909), to measure the distance between the realized and the predictive distributions. Unlike the relative entropy, the Hellinger distance is symmetric, it does obey the triangle inequality and it goes from zero to one. The use of the Hellinger distance is investigated for detecting the misfit of an IRT unidimensional model when response data are multidimensional with both simulated and real data.

2 Posterior Predictive Assessment of IRT Models

PPMC techniques are based on the comparison of observed data with replicated data generated or predicted by the model by using a number of diagnostic measures that are sensitive to model misfit (Sinharay, Johnson, and Stern, 2006). Substantial differences between the posterior distribution based on observed data and the posterior predictive distribution indicate poor model fit.

Given the data **y**, let p(**y**|**ω**) and p(**ω**) be the likelihood for a model depending on the set of parameters **ω** and the prior distribution for the parameters, respectively. In the IRT context, **ω** consists of the item parameters, person parameters, and trait correlations. To examine the differences between the observed and the replicated data, the latter are drawn from the posterior predictive distribution (PPD) of replicated data $\mathbf{y}^{\mathrm{rep}}$

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int_{\boldsymbol{\omega}} p(\boldsymbol{y}^{rep}|\boldsymbol{\omega})p(\boldsymbol{\omega}|y)\partial\boldsymbol{\omega}. \quad (1)$$

From a practical point of view, one should define a suitable discrepancy measure D(·) and compare the posterior distribution of D(**y**,**ω**), based on observed data, to the posterior predictive distribution of D($\mathbf{y}^{\mathrm{rep}}$,**ω**). Discrepancy measures should be chosen to capture relevant features of the data and differences among data and the model. As a first step in PPMC, a graphical analysis is conducted to investigate the differences among realized and replicated discrepancy measures. Then, the PPP-value is defined as

$$\text{PPP-value} = p\big(D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega}|\boldsymbol{y})\big). \quad (2)$$

The PPP-value is estimated by computing the proportion of MCMC replications which satisfy Equation (2). The PPP-values provide a measure of the degree to which observed data would be expected under the model: values close to 0 or 1 mean that the realized values fall far in the tails of the distribution of the discrepancy measure based on PPD, indicating misfit; conversely, values of approximately 0.5 mean that the realized values fall in the middle of the distribution, indicating good fit. As underlined by Levy, Mislevy, and Sinharay (2009), PPMC has several advantages over traditional techniques. The method is easy to apply and flexible because the reference distribution is built empirically and it does not require regularity conditions or asymptotic results. Moreover, PPMC relies on Bayesian estimation, which is based on the full posterior distribution: compared with maximum likelihood techniques, which are based on a point estimate, the method is able to directly incorporate uncertainty into the estimation. However, using PPMC is not equivalent to conducting a classical hypothesis test, and the method should be treated as a diagnostic tool (Gelman, Meng, and Stern, 1996; Sinharay, Johnson, and Stern, 2006).

The choice of a suitable discrepancy measure is crucial in PPMC. Effective diagnostic measures in checking for unidimensionality or multidimensionality are based on the association or on covariance/correlation among item pairs. In the first group, the Mantel-Haenszel (MH) statistic is based on the odds ratio conditionally to the rest score *s*, i.e., the raw test score obtained by excluding the two items. For each couple of items *j* and *j*', with *j*, *j*'=1,…,*k*, the MH statistic is defined as

$$\mathrm{MH}_{jj'} = \frac{\sum_s n_{11s} n_{00s}/n_s}{\sum_s n_{10s} n_{01s}/n_s}, \tag{3}$$

where $n_{tt's}$ is the number of subjects with rest score *s* who score *t* on item *j* and *t'* on item *j'*, with *t*, *t'*=0,1, and n_s is the number of subjects with rest score *s*. In the second group, the model-based covariance (MBC) is defined as follows

$$\mathrm{MBC}_{jj'} = \frac{\sum_{i=1}^{n}(Y_{ij}-E(Y_{ij}))(Y_{ij'}-E(Y_{ij'}))}{n}, \tag{4}$$

where Y_{ij} is the response variable for individual *i* to item *j*, with *i*=1,…,*n* and *j*=1,…,*k*, and E(Y_{ij}) is its expected value depending on the specific IRT model and the estimated parameters.

2.1 The Hellinger Distance for PPMC

While the PPP-value counts the number of replications for which the predictive discrepancy exceeds the realized one, the researcher may be interested in measuring the size of the difference itself. For this reason, Wu, Yuen, and Leung (2014) proposed the use of the relative entropy (RE), also known as Kullback-Leibler divergence or information, to evaluate the magnitude of the differences between the realized and the predictive measures with limited information statistics based on low-order margins. However, the RE is asymmetric and it is not upper bounded so it is difficult to establish proper threshold levels for assessing absolute model fit or making comparisons.

To overcome these limitations, we propose the use of the Hellinger distance which is symmetric, it does obey the triangle inequality and its range is 0-1. Since the Hellinger distance is used to quantify the distance between two probability measures, it can be used to measure the distance between the realized and the predictive distribution within PPMC as follows

$$\mathrm{H}(P,Q) = \sqrt{1 - \int \sqrt{p\big(D(\boldsymbol{y},\boldsymbol{\omega})\big)p\big(D(\boldsymbol{y}^{rep},\boldsymbol{\omega})\big)d\boldsymbol{y}d\boldsymbol{\omega}}.} \tag{5}$$

The direct calculation of (5) is computationally demanding and it is usually done via MCMC. Specifically, it is calculated by using the normal kernel density estimates to represent the probability density functions of the realized and the predictive discrepancy measures, given the MCMC replications. In order to check for model unidimensionality, we propose the use of the Hellinger distance with the MBC discrepancy measure, which is based on both data and model parameters, to take into account a fit measure for each item pair. A MATLAB code was written by the Authors to implement the proposal.

3 Main Results

A simulation study is conducted to investigate the performance of the PPP-values and the Hellinger distance at detecting the misfit of a unidimensional IRT model when the data structure is multidimensional. Two different multidimensional IRT models are considered, namely the multi-unidimensional and additive models (see Sheng and Wikle, 2009). Within a confirmatory approach, the multi-unidimensional model relates each item response to a single latent variable, by allowing for trait correlations. In the additive model, a further overall latent trait is assumed underlying all item responses. All traits may be correlated as well. The corresponding IRT unidimensional model is built under the assumption of unidimensionality.

In the study, response data for tests with 10 items and 1,000 respondents are simulated. The trait correlations are manipulated. A number of 5,000 MCMC iterations are conducted, where 1,000 are used for PPMC. Finally, 100 replications are done for each simulation condition. The MH statistic and the MBC are used as discrepancy measures, with 45 item pairs to be considered. Given the mean of the PPP-values for each item pair over the replications, the proportion of extreme PPP-values (below 0.05 or above 0.95) is estimated pooling the results on all item pairs. Given the mean of the Hellinger distance for the MBC (MBC-H) for each item pair, some summary descriptive measures are computed by pooling the results on all item pairs.

Data are generated from the multidimensional models with a bidimensional structure and then analysed with the unidimensional approach. Due to space limit, the results are not reported in the paper but only briefly discussed in the following. The results on the PPP-values show that, for both discrepancy measures, the proportion of extreme values is above 0.75 for most conditions suggesting bad fit. In particular, the MH statistic outperforms the MBC. In the cases of strong and very strong trait correlations, data are conceived as unidimensional and, consequently, the proportion of extreme PPP-values decreases. The average MBC-H is estimated above 0.8 for most cases showing bad fit. The results are coherent and easily interpretable, in fact, the more the MBC-H is close to one, the bad the fit is. For strongly correlated traits, the MBC-H is estimated on average around 0.65. This means that the data are conceived as unidimensional but the distance measure is still able to catch the discrepancy between the generating model and the one used to analyse response data.

Data coming from a survey conducted by the University of Bologna (Bernini, Matteucci, and Mignani, 2015) to investigate the residents' perceptions toward tourism in terms of perceived benefits and costs are used. A total of 5 items on benefits and 5 items on costs are administered, suggesting a bidimensional latent structure. The unidimensional, multi-unidimensional and additive models are fitted. The results on the PPP-values show that the unidimensional approach is associated to a proportion of about 80% of extreme values. On the contrary, about 30% and 16% of extreme

PPP-values are reported for the multi-unidimensional and the additive model, respectively. Clearly, the additive model shows the best fit. These results are confirmed by the analysis with the MBC-H. On average, the estimated distances are about 0.8, 0.5, and 0.4 for the unidimensional, multi-unidimensional and additive model, respectively.

The approach based on the Hellinger distance seems to be promising to evaluate model fit within posterior predictive assessment. In particular, all measures could be used to investigate misfit due to specific items. A more comprehensive simulation study is needed to check the performance of the method for different simulation conditions.

References

1. Bernini, C., Matteucci, M., Mignani, S.: Investigating heterogeneity in residents' attitudes toward tourism with an IRT multidimensional approach. Qual. Quant. **49**, 805-826 (2015).
2. Gelman, A., Meng, X.L., Stern, H.S.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**, 733-807 (1996).
3. Hellinger, E.: Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik (in German) **136**, 210–271 (1909).
4. Levy, R., Svetina, D.: A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. Br. J. Math. Stat. Psychol. **64**, 208-232 (2011).
5. Levy, R., Mislevy, R.J., Sinharay, S.: Posterior predictive model checking for multidimensionality in item response theory. Appl. Psychol. Meas. **33**, 519-537 (2009).
6. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applies statistician. Ann. Stat. **12**, 1151-1172 (1984).
7. Sheng. Y., Wikle. C.: Bayesian IRT models incorporating general and specific abilities. Behaviormetrika **36**, 27-48 (2009).
8. Sinharay, S.: Bayesian item fit analysis for unidimensional item response theory models. Posterior predictive assessment of item response theory models. Br. J. Math. Stat. Psychol. **59**, 429-449 (2006).
9. Sinharay, S., Johnson, M.S., Stern, H.S.: Posterior predictive assessment of item response theory models. Appl. Psychol. Meas. **30**, 298-321 (2006).
10. van der Linden, W. J., Hambleton, R.K. Handbook of Modern Item Response Theory. Springer-Verlag, New York (1997).
11. Wu, H., Yuen, K.V., Leung, S.O.: A novel relative entropy-posterior predictive model checking approach with limited information statistics for latent trait models in sparse 2^k contingency tables. Comput. Stat. Data Anal. **79**, 261-276 (2014).