Qual Quant (2015) 49:917–926
DOI 10.1007/s11135-014-0059-y

The approach of power priors for ability estimation in IRT models

Mariagiulia Matteucci · Bernard P. Veldkamp

Published online: 25 July 2014

Abstract The aim of the paper is to propose the introduction of power prior distributions in the ability estimation of item response theory (IRT) models. In the literature, power priors have been proposed to integrate information coming from historical data with current data within Bayesian parameter estimation for generalized linear models. This approach allows to use a weighted posterior distribution based on the historical study as prior distribution for the parameters in the current study. Applications can be found especially in clinical trials and survival studies. Here, power priors are introduced within a Gibbs sampler scheme in the ability estimation step for a unidimensional IRT model. A Markov chain Monte Carlo algorithm is chosen for the high flexibility and possibility of extension to more complex models. The efficiency of the approach is demonstrated in terms of measurement precision by using data from the Hospital Anxiety and Depression Scale with a small sample.

Keywords Power priors · Item response theory models · Ability estimation · Gibbs sampler

1 Introduction

In Bayesian estimation, where posterior evidence is derived from the combination of the likelihood and a prior distribution, the elicitation of the prior remains a crucial issue. Although non-informative priors can be employed in absence of prior information (see Kass and Wasserman 1996), they can cause instability in the posterior estimates and convergence problems for the Gibbs sampler (Ibrahim and Chen 2000). In applied research, prior information on the

M. Matteucci (✉)
Department of Statistical Sciences, University of Bologna, via Belle Arti, 41, 40126 Bologna, Italy
e-mail: m.matteucci@unibo.it

B. P. Veldkamp
Research Center for Examination and Certification, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands
e-mail: b.p.veldkamp@gw.utwente.nl

parameters of interests is often available in the form of historical data, collateral information or predictions from subject matter experts. In order to combine historical data with observed data, power prior distributions were proposed for the parameter estimation of regression models (Ibrahim and Chen 2000) and more generally of generalized linear models (Chen et al. 2000). This approach allows to discount historical data by introducing an appropriate weight. Generally, the introduction of informative priors is strongly encouraged with small sample sizes, in case the available data provide only indirect information about the parameters of interests, and for preventing from inappropriate inferences, consistent with the likelihood (see Gelman 2002).

The availability of a small sample is also a recurring context in educational and psychological measurement, where item response theory (IRT) models (Lord and Novick 1968) are commonly used. Given the responses of a sample of subjects to categorical test items, IRT models express the probability of an item response as a function of the item psychometric properties and the latent abilities underlying the response process. Therefore, parameter estimation involves two phases: (a) calibration, where the item parameters are estimated, and (b) scoring, where the subjects are located on the latent trait scale. In both phases, the issue of measurement precision is very important, especially with small samples. In fact, a large number of respondents is needed for calibration and, conversely, as many items as possible are needed for an accurate ability estimation. As a rule of thumb, for a two-parameter unidimensional model, a calibration sample size of at least 500 subjects and 20 or more items is suggested (see de Ayala 2009, page 105). Obviously, in practice it is not always possible to rely on optimal testing conditions. For this reason, the introduction of collateral or historical information in model estimation assumes a particular importance and one possibility for improving the parameter estimation would be to include informative prior distributions.

Bayesian estimation of IRT models through Markov chain Monte Carlo (MCMC) has become very popular recently (see among others, Béguin and Glas 2001; Fox and Glas 2001; Sheng and Wikle 2009; Natesan et al. 2010) because it allows the simultaneous estimation of item parameters and individual abilities while overcoming multiple integration problems of marginal maximum likelihood (MML) estimation and model limitations of conditional maximum likelihood (CML) estimation (for a review of both methods in IRT, see van der Linden and Hambleton 1997). Under this approach, both item parameters and abilities are viewed as random variables, with a corresponding prior distribution. Therefore, it is easily possible to introduce informative priors, both empirical or not empirical, to increase measurement precision and compensate for a small sample size. In particular, the capability of taking uncertainty into account, i.e. to treat the item parameters not as fixed quantities, is fundamental (Veldkamp et al. 2013). Further advantages are the possibility of using Bayesian model comparison techniques, such as Bayes factors, Bayesian deviance and posterior predictive model checks (for a recent study see Azevedo et al. 2012). Despite the use of MCMC was limited in the past due to its computational intensity, the increasing availability of cheap computing power stimulated its present diffusion. The introduction of random item models, where the items are assumed to be random draws from a population with a specific distribution, is rather recent. This approach was found to be effective in measuring longitudinal ability change, explanation of item difficulties and studies on differential item functioning (see Briggs and Wilson 2007; De Boeck 2008).

In order to increase the measurement precision, the introduction of empirical prior information was proposed by capitalizing on the relationship between the parameters of interests and a set of auxiliary variables (e.g. van der Linden 1999; Matteucci et al. 2012a, b; Matteucci and Veldkamp 2013). However, historical data such as data on previous testing rounds or

previous test administrations, consisting in the item responses, the estimated item parameters and abilities of respondents, can also be introduced in the prior distributions.

For these reasons, we propose the introduction of historical data through the approach of power priors for ability estimation in the unidimensional two-parameter normal ogive (2PNO) IRT model. The power prior distribution is defined directly in the Gibbs sampler step performing the ability sampling. As a case study, we take into account response data collected through the administration of the Hospital Anxiety and Depression Scale (HADS) questionnaire to a sample of patients.

The paper is organized as follows. Section 2 gives a brief overview of the approach of power priors. In Sect. 3, MCMC estimation for IRT models is reviewed taking into account the 2PNO model, and the new proposal of introducing the power prior distribution in the ability sampling step is presented. Section 4 discusses the application of the approach to real data while concluding remarks end the paper.

2 Power priors

The underlying idea of power priors is to introduce information coming from historical data in the estimation of the model parameters based on current data. According to Ibrahim and Chen (2000), historical data are defined as data arising from previous similar studies, where the same response variable and covariates of the current study have been collected. In applied research, and especially in medical research, it is very common that similar studies are conducted repeatedly. For example, in clinical trials, data on similar or slightly modified treatments are often available. Also previous data collected in longitudinal studies can be classified as historical data, and could be used to compare and interpret the results based on current data.

Given the current data $D = (n, \mathbf{y}, \mathbf{X})$, where n, $\mathbf{y}$, and $\mathbf{X}$ are the sample size, the vector of dependent variables, and the matrix of covariates, respectively, the likelihood function for the current study can be expressed generally by $\mathcal{L}(\theta|D)$ with θ representing the parameters of interests. Analogously, let $D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0)$, be the historical data. Given the initial prior distribution $\pi_0(\theta|\cdot)$ for θ before the historical data are observed, the power prior distribution of θ for the current study can be defined as

$$\pi(\theta|D_0, a_0) \propto \mathcal{L}(\theta|D_0)^{a_0}\pi_0(\theta|c_0), \tag{1}$$

where $0 \leq a_0 \leq 1$ is a scalar parameter controlling for the influence of the historical data on the current data, and c_0 is the hyperparameter of the initial prior (Ibrahim and Chen 2000). According to this definition, the power prior distribution is the product of the likelihood function based on the historical data, raised to a power a_0, and the initial prior distribution for θ. The parameter a_0 is a relative precision parameter for the historical data which controls the heaviness of the tails for $\pi(\theta|D_0, a_0)$, in fact as a_0 becomes smaller, the tails become heavier. Two limit cases can be distinguished: $a_0 = 1$ where the prior of the current study corresponds to the posterior of the previous study, and $a_0 = 0$ where no historical data are incorporated.

By following an hierarchical approach, it is also possible to specify a proper prior distribution for the precision parameter a_0 in order to obtain the so called joint power prior distribution, as follows

$$\pi(\theta, a_0|D_0) \propto \mathcal{L}(\theta|D_0)^{a_0}\pi_0(\theta|c_0)\pi(a_0|\gamma_0), \tag{2}$$

where γ_0 is a vector of hyperparameters. There are several choices for the prior distribution for a_0, such as beta, truncated gamma and truncated normal distributions (Ibrahim and Chen 2000). Starting from the initial idea of Diaconis and Ylvisaker (1979), power priors have been developed for regression models (Ibrahim and Chen 2000), and more generally for generalized linear models (Chen et al. 2000).

3 Bayesian estimation via MCMC of IRT models

IRT models (see e.g. Lord and Novick 1968) allow the analysis of categorical response data through the assumption of one or more latent variables underlying the response process. Given a set of k categorial items submitted to a sample of n subjects, the model is designed to infer the psychometric characteristics of test items and the individual traits. Depending on the number of model parameters and the latent dimensionality, these models can be very complex making the classical estimation based on MML rather difficult due to multiple integration problems. An alternative approach is represented by simulation techniques such as MCMC, a general class of methods where the posterior distribution of interest is reproduced through the simulation of one or more sequences of correlated random variables. Basically, the advantages of MCMC are the simultaneous estimation of item parameters and candidate scores, the incorporation of the dependencies among variables and all sources of uncertainties and the possibility of using Bayesian model comparison techniques. A practical problem of MCMC is represented by the computational intensity of the method. However, thanks to the increasing availability of cheap computing power, this obstacle is going to be removed, even for large data sets.

In the IRT literature, within the MCMC methods, the Gibbs sampler (Geman and Geman 1984) was introduced for the estimation of the unidimensional 2PNO model by Albert (1992). The 2PNO model is chosen because of its feasibility within Bayesian estimation, in fact the probit link function allows to use normal-normal conjugate families. In the literature, the method was extended to, for example, the implementation of the Gibbs sampler for the estimation of multidimensional IRT models (Béguin and Glas 2001), multilevel IRT models (Fox and Glas 2001) and IRT models with general and specific latent traits (Sheng and Wikle 2009).

In the following, the general algorithm for the estimation of the 2PNO model is described and a modification of the Gibbs sampler step for ability estimation is proposed in order to introduce the power prior.

3.1 Gibbs sampler for the 2PNO model

Under the assumption of unidimensionality, i.e. the existence of a single latent ability underlying the response process, given a binary response variable Y_{ij} for individual i to item j, where $Y_{ij} = 1$ for a correct response and $Y_{ij} = 0$ for an incorrect response, with $i = 1, \ldots, n$ and $j = 1, \ldots, k$, the 2PNO model (Lord and Novick 1968) specifies the probability of a correct response as a monotonically increasing function of the trait, as follows

$$P\left(Y_{ij} = 1 | \theta_i, \alpha_j, \delta_j\right) = \Phi\left(\alpha_j \theta_i - \delta_j\right) = \int_{-\infty}^{\alpha_j \theta_i - \delta_j} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \tag{3}$$

where θ_i is the ability of person i, α_j and δ_j are the item parameters for item j, and Φ is the cumulative normal distribution function. In particular, the discrimination parameter α_j assesses the power of the item to differentiate respondents of different ability, while the

difficulty parameter δ_j represents the threshold or difficulty level of the item. Theoretically, discriminations could take all values in the real line. However, positive values are found in practice, ensuring that the probability of endorsing an item increases as ability increases. Difficulties can take values in the set of real numbers and a mean and median value around zero would ensure a balanced pool between easy and difficult items in the test.

In order to be used within the Gibbs sampler, the dichotomous response variables Y_{ij} should be modeled as indicators of the corresponding underlying variables Z_{ij}, which are defined to be conditionally independent and identically distributed as $Z_{ij} \sim N(\eta_{ij}; 1)$, where $\eta_{ij} = \alpha_j\theta_i - \delta_j$. Given a standard normal prior distribution for θ_i, i.e. θ_i i.i.d.$\sim N(0, 1)$, and an indicator function of positiveness for discrimination parameters as prior distribution for the item parameters $\boldsymbol{\xi}$, i.e. $P(\boldsymbol{\xi}) = \prod_{j=1}^{k} I(\alpha_j > 0)$, where $\boldsymbol{\xi}$ is the vector containing the k elements $\boldsymbol{\xi}_j = (\alpha_j, \delta_j)$, the joint posterior distribution (see Albert 1992) can be expressed as

$$\begin{aligned} P(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{Y}) &= P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y})P(\boldsymbol{\theta})P(\boldsymbol{\xi}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{k}\left\{\phi(Z_{ij}; \eta_{ij}, 1)\left[I(Z_{ij} > 0)I(y_{ij} = 1) + I(Z_{ij} \leq 0)I(y_{ij} = 0)\right]\right\} \\ &\quad \prod_{i=1}^{n}\phi(\theta_i; 0, 1)\prod_{j=1}^{k} I(\alpha_j > 0). \end{aligned} \tag{4}$$

Distribution (4) has an intractable form, so we should resort to the Gibbs sampler. The algorithm is applied in order to iteratively sample from the following conditional distributions, given $\mathbf{Y}$, until convergence:

1. $\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}$
2. $\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}$
3. $\boldsymbol{\xi}|\boldsymbol{\theta}, \mathbf{Z}$.

The first conditional distribution of the Z_{ij} is truncated normal, as follows

$$Z_{ij}|\boldsymbol{\theta}, \boldsymbol{\xi} \sim \begin{cases} N(\eta_{ij}, 1) \text{ with } Z_{ij} > 0 \text{ if } Y_{ij} = 1, \\ N(\eta_{ij}, 1) \text{ with } Z_{ij} \leq 0 \text{ if } Y_{ij} = 0. \end{cases} \tag{5}$$

The second conditional distribution is obtained by starting from the normal regression model $Z_{ij} + \delta_j = \alpha_j\theta_i + \epsilon_{ij}$, where ϵ_{ij} i.i.d. $\sim N(0, 1)$ and θ_i is viewed as regression coefficient. Therefore, the likelihood function of θ_i follows the normal distribution with mean equal to the least square estimate of θ_i, specifically $\hat{\theta}_i = \sum_{j=1}^{k}\alpha_j(Z_{ij}+\delta_j)/\sum_{j=1}^{k}\alpha_j^2$, and variance $v = 1/\sum_{j=1}^{k}\alpha_j^2$. According to the standard normal prior imposed on θ_i, the posterior distribution turns out to be normal and parameterized as

$$\theta_i|\mathbf{Z}, \boldsymbol{\xi} \sim N\left(\frac{\hat{\theta}_i/v}{1/v+1}, \frac{1}{1/v+1}\right). \tag{6}$$

The third conditional distribution can be derived by following the same approach and considering the normal regression model $\mathbf{Z}_j = \mathbf{X}\boldsymbol{\xi_j} + \boldsymbol{\epsilon_j}$, where $\mathbf{X} = [\boldsymbol{\theta} \ -\mathbf{1}]$, $\boldsymbol{\epsilon_j}$ is a random sample from $N(0; 1)$, $\boldsymbol{\xi_j}$ are viewed as regression coefficients. Given a multivariate normal likelihood $\mathbf{Z}_j|\boldsymbol{\xi_j} \sim N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j; (\mathbf{X}'\mathbf{X})^{-1})$, a vague prior $P(\boldsymbol{\xi}) = \prod_{j=1}^{k} I(\alpha_j > 0)$ turns out with the following posterior distribution for the item parameters

$$\boldsymbol{\xi_j}|\mathbf{Z}, \boldsymbol{\theta} \sim N((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}_j); (\mathbf{X}'\mathbf{X})^{-1})I(\alpha_j > 0). \tag{7}$$

Alternatively, by using a normal prior distribution for both discrimination and threshold parameters $\alpha_j \sim N(0; \sigma_\alpha^2)$, $\delta_j \sim N(0; \sigma_\delta^2)$ (Béguin and Glas 2001), the corresponding posterior distribution becomes

$$\boldsymbol{\xi_j}|\mathbf{Z}, \boldsymbol{\theta} \sim N\left(\left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma_0}^{-1}\right)^{-1}(\mathbf{X}'\mathbf{Z}_j); \left(\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma_0}^{-1}\right)^{-1}\right), \tag{8}$$

where $\boldsymbol{\Sigma_0}$ is the variance-covariance matrix for the item parameters.

Starting from a set of starting values, the Gibbs sampler proceeds with the iterative sampling from the conditional distributions until convergence. The choice of starting values is not crucial in MCMC but reasonable initial points may speed convergence.

3.2 Ability estimation for the 2PNO model with power priors

The introduction of a power prior distribution for the ability requires a modification of the second conditional distribution within the Gibbs sampler described in Sect. 3.1. First of all, recall that the conditional distribution for θ_i derives from the regression model $Z_{ij} + \delta_j = \alpha_j\theta_i + \epsilon_{ij}$, where θ_i is treated as regression coefficient. Our historical data consist of $D_0 = (k_0, Z_{ij0} + \delta_{j0}, \alpha_{j0})$, where k_0 is the number of items in the historical study, $Z_{ij0} + \delta_{j0}$ and α_{j0} are the observed values for the dependent variable and the regressor, respectively.

Given the scalar precision parameter a_0, a general formulation for the power prior is $\pi(\theta_i|D_0, a_0) \propto L(\theta_i|D_0)^{a_0}\pi_0(\theta_i)$. The likelihood is again normal and specifically: $Z_{ij} + \delta_j|\theta_i, D_0 \sim N(\hat{\theta}_{i0}, a_0^{-1}v_0)$, where $\hat{\theta}_{i0} = \sum_{j=1}^{k}\alpha_{j0}(Z_{ij0} + \delta_{j0})/\sum_{j=1}^{k}\alpha_{j0}^2$ and $v_0 = 1/\sum_{j=1}^{k}\alpha_{j0}^2$. Assuming a standard normal as initial prior $\pi_0(\theta_i)$, i.e. $\theta_i \sim N(0, 1)$, it can be easily shown that, from its combination with the normal likelihood, the following power prior distribution for ability can be derived

$$\theta_i|D_0, a_0 \sim N\left(\frac{a_0\hat{\theta}_{i0}/v_0}{a_0/v_0 + 1}; \frac{1}{a_0/v_0 + 1}\right). \tag{9}$$

The last step consists in deriving the corresponding posterior distribution for ability, to be used as the conditional distribution for θ_i in the Gibbs sampler step. Taking into account once again the normal–normal Bayesian model, combining the power prior (9) with the normal likelihood for ability described in Sect. 3.1, the following posterior distribution can be derived after simple arithmetic manipulations

$$\theta_i|\mathbf{Z}, \boldsymbol{\xi}, D_0, a_0 \sim N\left(\left(\frac{\hat{\theta}_i}{v} + \frac{a_0\hat{\theta}_{i0}}{v_0}\right)\left(\frac{1}{v} + \frac{a_0}{v_0} + 1\right)^{-1}; \left(\frac{1}{v} + \frac{a_0}{v_0} + 1\right)^{-1}\right). \tag{10}$$

4 Case study

An empirical application is presented in order to show the effectiveness of including a power prior distribution in increasing the measurement precision of the ability estimates when the sample size is small.

Despite IRT applications are mainly in the field of educational and psychological measurement, there is an increasing interest in using these models in the medical field. In fact, data used in this study come from the administration of the well-known Hospital Anxiety and Depression Scale (HADS) questionnaire to a sample of Dutch hospital patients. The questionnaire investigates the anxiety and depression state by using a set of 14 items, which are originally scored on an ordinal scale from 0 to 3, denoting an increasing level of psychological disease. Here, the item responses were reverse scored, when needed, and dichotomized

Table 1 Item parameter estimates for the baseline measurement

Item	$\hat{\alpha}_j$	SD_α	$\hat{\delta}_j$	SD_δ
1	1.25	0.26	−0.70	0.16
2	1.04	0.19	0.04	0.12
3	0.78	0.16	0.12	0.11
4	1.57	0.28	0.90	0.17
5	1.35	0.26	−0.31	0.14
6	2.05	0.48	0.61	0.18
7	1.17	0.23	−0.72	0.15
8	0.67	0.19	−1.26	0.16
9	0.90	0.17	0.33	0.12
10	0.71	0.15	0.31	0.11
11	1.62	0.32	−0.64	0.16
12	1.32	0.25	−0.14	0.13
13	1.31	0.28	0.72	0.16
14	0.81	0.18	0.63	0.13

SD Standard deviation

so that 0 means "lack of disease" and 1 means "presence of disease". Data consist of the item responses at a baseline measurement and four subsequent repeated measurements. The idea is to use the data at the third measurement as historical data to estimate the "ability" θ of patients for the current and last fourth measurement. In this context, ability should be interpreted as the anxiety and depression state, so that patients with an high ability estimates are associated to an high level of disease.

The sample of patients at the baseline measurement consists of 58.2 % of males and 34.6 % of smokers. The average age is about 72, with the youngest patient being 52 and the oldest one being 85 years old. At baseline measurement, the average rate of negative responses meaning the presence of disease was 47.67 %. This rate is only slightly reduced at measurement 3 (47.62 %) while at measurement 4 a relevant improvement may be observed in the patients' state, in fact the rate becomes 46.17 %. By computing the raw score for all the measurements, a median value of 7 can be observed for the baseline and the third measurement, while a median value of 6.5 is recorded for the fourth measurement, meaning again that the psychological conditions of patients improved over time.

The 2PNO item parameters were estimated in the baseline measurement on a sample of $n = 154$ patients. The parameter estimates are reported in Table 1.

The results show that the α parameters are all largely positive ensuring a good capability of all items to differentiate among patients with different levels of anxiety and depression. Moreover, the threshold parameters δ are rather balanced in terms of positive and negative values ensuring that the questionnaire contains items with different levels of criticity, i.e. items which are more or less likely to be answered positively by patients with different levels of anxiety and depression. The estimated item parameters in the baseline measurement are then treated as fixed over test administrations.

The focus is instead on investigating the measurement precision of the ability estimates under different approaches. For this reason, we used the Gibbs sampler under the 2PNO model to estimate the abilities of $n = 123$ patients that responded to the HADS questionnaire at measurement 4. We introduced historical data on measurement 3 in the form of the power prior distribution (9) for ability, by specifying different precision parameters: $a_0 = 0.0$, $a_0 = 0.2$, $a_0 = 0.5$, and $a_0 = 0.9$. In particular, a_0 equal to zero means that historical data are not

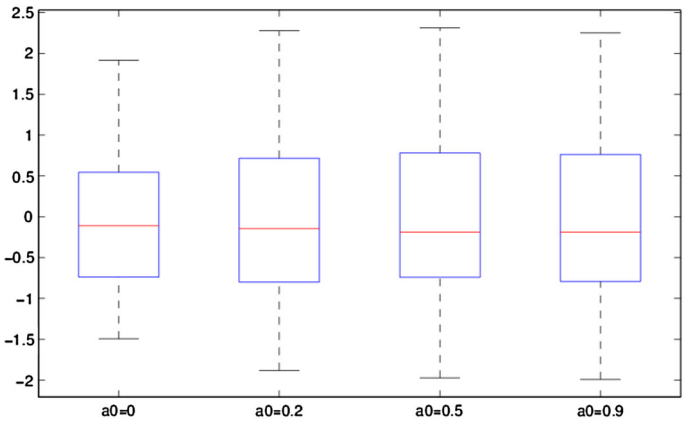

Fig. 1 Boxplots of ability estimates under different choices for a_0

Table 2 Summary statistics for the standard deviation (SD) and the 95 % credibility interval (CI) width of the ability estimates under different choices for a_0

	$a_0 = 0$	$a_0 = 0.2$	$a_0 = 0.5$	$a_0 = 0.9$
SD				
Mean	0.35	0.29	0.23	0.19
Median	0.32	0.28	0.23	0.19
Min	0.26	0.23	0.20	0.18
Max	0.57	0.40	0.29	0.22
CI width				
Mean	1.35	1.12	0.91	0.76
Median	1.24	1.09	0.90	0.75
Min	1.01	0.92	0.80	0.68
Max	2.21	1.55	1.11	0.86

introduced in the ability estimation for the current study while, on the contrary, a_0 equal to 0.9 means an assignment of a very large prior weight to historical information. The Gibbs sampler reached the convergence within all approaches after 5,000 iterations (500 burn-in). All analyses were conducted by using a Matlab program written by the Authors. The output consists on the entire posterior distribution of the ability for each patient. The estimated abilities are derived as the expected mean from the posterior distribution. Box-plots showing the distribution of the estimated abilities for all patients are reported in Fig. 1. The results clearly show that, when a prior weight is given to historical data, the range is higher meaning that it is possible to distinguish patients with different levels of the latent trait deeply.

For each ability parameter, the expected a posteriori estimate, the standard deviation, and the 95 % credibility interval were identified. In Table 2, some summary statistics on the measurement precision (standard deviation and width of the credibility interval) are reported to summarize the results.

As can be easily noticed, there is a relevant improvement in the precision of the ability estimates as the prior weight is increased. This is demonstrated by lower standard deviation and CI width on average.

Lastly, Fig. 2 reports the kernel density for the ability of a respondent with an intermediate ability which is estimated within all approaches in $\hat{\theta} = 0.03$. An increase in the prior weight a_0 is represented by an increase in the thickness of the density line. The plot clearly shows

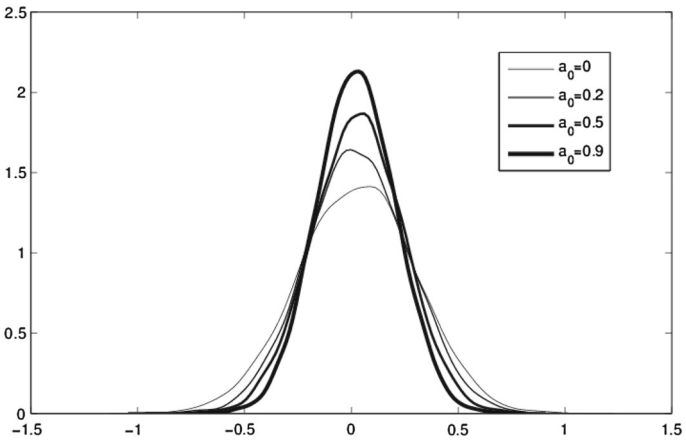

Fig. 2 Kernel density for the ability of a respondent with $\hat{\theta} = 0.03$ under different choices for a_0

that the heaviness of the tails is reduced as a_0 becomes larger and, consequently, there is a noticeable improvement in the measurement precision. The large influence of the historical data is also enhanced by the presence of a small sample ($n = 123$).

5 Discussion

In this work, the introduction of a power prior distribution based on historical data is proposed for the ability estimation based on the current data in the unidimensional 2PNO IRT model. In particular, the power prior is introduced in the ability sampling step for the Gibbs sampler, within a fully Bayesian approach. The effectiveness of the method is demonstrated in a case study by using data coming from a HADS study, where the latent trait of interest is the anxiety and the depression state of hospital patients. A small sample was taken into account in order to show the potential of the power prior to improve the measurement precision when there is lacking information from the current study. The inclusion of power priors allowed to integrate coherently data coming from the current study and historical data, by adopting different weights. An increase in the importance given to the prior weight resulted in an improvement in the measurement precision of the anxiety and depression state of patients. However, we should note that the effectiveness of the approach relies on the quality of historical data.

The paper provided several insights from which draw on some further research. First of all, a sensitivity study should be conducted on different choices for the relative precision parameter a_0. In particular, elicitation of the parameter a_0 can be driven by meta-analytic arguments by using the connections between power priors and hierarchical models as underlined in Chen and Ibrahim (2006). An alternative approach would be to specify a joint power prior distribution for the parameter of interest and the power parameter (Ibrahim and Chen 2000). However, some difficulties have been highlighted in the literature with reference to the specification of a prior distribution for the a_0 parameter (see Neuenschwander et al. 2009). Also, more informative initial priors could be combined in the definition of the power prior. Another relevant issue would be to study the integration of multiple sources of information in the power prior and the possibility of combining multiple historical data which can be easily available in longitudinal studies. Finally, from the IRT model point of view, it would be very

important to derive an analogous approach to introduce power priors in the estimation of item parameters.

Acknowledgments This research has been partially funded by the Italian Ministry of Education with the FIRB ("Futuro in ricerca") 2012 project on "Mixture and latent variable models for causal-inference and analysis of socio-economic data". We would like to thank Prof. Job van der Palen from the University of Twente for providing us the HADS data.

References

Albert, J.H.: Bayesian estimation of normal ogive item response curves using Gibbs sampling. J. Educ. Stat. **17**, 251–269 (1992)

Azevedo, C.L.N., Andrade, D.F., Fox, J.-P.: A Bayesian generalized multiple group IRT model with model-fit assessment tools. Comput. Stat. Data Anal. **56**, 4399–4412 (2012)

Béguin, A.A., Glas, C.A.W.: MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika **66**, 541–562 (2001)

Briggs, D.C., Wilson, M.: Generalizability in item response modeling. J. Educ. Meas. **44**, 131–155 (2007)

Chen, M.H., Ibrahim, J.G.: The relationship between the power prior and hierarchical models. Bayesian Anal. **1**(3), 551–574 (2006)

Chen, M.H., Ibrahim, J.G., Shao, Q.M.: Power prior distributions for generalized linear models. J. Stat. Plan. Inference **84**, 121–137 (2000)

de Ayala, R.J.: The Theory and Practice of Item Response Theory. Guilford Press, New York (2009)

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

Diaconis, P., Ylvisaker, D.: Conjugate priors for exponential families. Ann. Stat. **7**, 269–281 (1979)

Fox, J.-P., Glas, C.A.W.: Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika **66**, 271–288 (2001)

Gelman, A.: Prior distribution. In: El-Shaarawi, A.H., Piegorsch, W.W. (eds.) Encyclopedia of Environmetrics, vol. 3, pp. 1634–1637. Wiley, New York (2002)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

Ibrahim, J.G., Chen, M.H.: Power prior distributions for regression models. Stat. Sci. **15**, 46–60 (2000)

Kass, R.E., Wasserman, L.: The selection of prior distributions by formal rules. J. Am. Stat. Assoc. **91**, 1343–1370 (1996)

Lord, F.M., Novick, M.R.: Statistical theories of mental test scores. Addison-Wesley, Reading, MA (1968)

Matteucci, M., Veldkamp, B.P.: Computer adaptive testing with empirical prior information: a Gibbs sampler approach for ability estimation. Stat. Methods Appl. **22**(2), 243–267 (2013)

Matteucci, M., Mignani, S., Veldkamp, B.P.: Prior distributions for item parameters in IRT models. Commun. Stat. Theory Methods **41**(16–17), 2944–2958 (2012a)

Matteucci, M., Mignani, S., Veldkamp, B.P.: The use of predicted values for item parameters in item response theory models: an application in intelligence tests. J. Appl. Stat. **39**(12), 2665–2683 (2012b)

Natesan, P., Limbers, C., Varni, J.W.: Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. Educ. Psychol. Meas. **70**(3), 420–439 (2010)

Neuenschwander, B., Branson, M., Spiegelhalter, D.J.: A note on the power prior. Stat. Med. **28**, 3562–3566 (2009)

Sheng, Y., Wikle, C.K.: Bayesian IRT models incorporating general and specific abilities. Behaviormetrika **36**, 27–48 (2009)

van der Linden, W.J.: Empirical initialization of the trait estimation in adaptive testing. Appl. Psychol. Meas. **23**, 21–29 (1999)

van der Linden, W.J., Hambleton, R.K.: Handbook of Modern Item Response Theory. Springer, New York (1997)

Veldkamp, B.P., Matteucci, M., de Jong, M.: Uncertainties in the item parameter estimates and robust automated test assembly. Appl. Psychol. Meas. **37**(2), 123–139 (2013)