Journal of Statistical Planning and Inference 166 (2015) 52–66

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A Bayesian nonparametric approach for the analysis of multiple categorical item responses

Andrew Waters [a], Kassandra Fronczyk [a], Michele Guindani [b,*], Richard G. Baraniuk [a], Marina Vannucci [a]

[a] Rice University, Houston, TX, USA

[b] UT MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 1 July 2014
Accepted 2 July 2014
Available online 25 July 2014

Keywords:
Bayesian nonparametrics
Cluster analysis
Multinomial probit model
Factor analysis
Learning analytics

ABSTRACT

We develop a modeling framework for joint factor and cluster analysis of datasets where multiple categorical response items are collected on a heterogeneous population of individuals. We introduce a latent factor multinomial probit model and employ prior constructions that allow inference on the number of factors as well as clustering of the subjects into homogeneous groups according to their relevant factors. Clustering, in particular, allows us to borrow strength across subjects, therefore helping in the estimation of the model parameters, particularly when the number of observations is small. We employ Markov chain Monte Carlo techniques and obtain tractable posterior inference for our objectives, including sampling of missing data. We demonstrate the effectiveness of our method on simulated data. We also analyze two real-world educational datasets and show that our method outperforms state-of-the-art methods. In the analysis of the real-world data, we uncover hidden relationships between the questions and the underlying educational concepts, while simultaneously partitioning the students into groups of similar educational mastery.

Published by Elsevier B.V.

1. Introduction

In this paper, we develop a Bayesian Nonparametric model for the joint factor and cluster analysis of datasets where multiple categorical response items are collected on a heterogeneous population of individuals. Similarly as in conventional Bayesian probit and multinomial regression models (Albert and Chib, 1993), we assume that each categorical response outcome is a surrogate for a continuous unobserved latent variable. A Bayesian factor model is then assumed on the latent variables. With respect to common factor analysis as well as multidimensional item response theory (Reckase, 2009) approaches, we allow the number of underlying factors to be inferred directly from the data. Our approach is similar to that of Rai and Daumé III (2008) and Knowles and Ghahramani (2011), who consider a nonparametric prior on the number of latent concepts based on the Indian Buffet Process (IBP) proposed by Griffiths and Ghahramani (2005). In addition, we employ a Dirichlet Process prior Ferguson (1973, 1974) to cluster subjects into groups characterized by similar factor structures. Clustering allows us to borrow strength across subjects, therefore helping in the estimation of the model parameters, particularly when the number of observations is small. We also discuss mechanisms for the imputation of missing data. We employ computationally efficient Markov chain Monte Carlo (MCMC) methods to provide tractable inference for the model parameters of interest.

* Corresponding author.
E-mail addresses: mguindani@mdanderson.org, michele.guindani@me.com (M. Guindani).

http://dx.doi.org/10.1016/j.jspi.2014.07.004
0378-3758/Published by Elsevier B.V.

Surveys and questionnaires with ordinal categorical responses are employed in many fields to gather relevant feedback information on individual attitudes toward a set of items. For example, in marketing, surveys are used to improve product delivery and pricing against competition. Here, we consider a specific application to personalized learning, which has recently emerged as an independent research topic within the field of education (Stamper et al., 2007; Li et al., 2011; Murray et al., 2004). Our model leverages the fact that knowledge in a given subject can typically be decomposed into a set of potential principles to learn, termed *concepts*. For personalized learning, in particular, statistical methods are widely employed to enhance student learning in a course, namely by assessing how well students understand educational concepts (*learning analytics*), and exploring the relationships between the test questions and the concepts (*content analytics*). Rigorous statistical methods for both learning and content analytics enable targeted feedback to learners, their instructors, and the content authors (Kulik, 1994).

Given the number of individuals typically surveyed and the number of topics assessed per individual, it is often of interest to reduce the dataset to an interpretable set of highly-informative variables. For example, in assessing tests or homework questions, a few skills or factors may play a role in understanding why certain learners succeed at some problems while failing at others. In turn, this information may be useful to predict future learner outcomes as well as diagnosing learner misconceptions. Traditionally, Item Response Theory (IRT) methods have been used to relate the individual responses to a set of latent traits, which summarize the non-observable characteristics of the person. However, many commonly used IRT approaches rely on the simplifying assumption that the relationship between each latent trait and the probabilities of correct response to a test item can be represented as a continuous mathematical function of a single or limited set of parameters (Reckase, 2009). For example, the popular Rasch model can be described as a two-parameter logistic model categorizing both users and items (Rasch, 1993). While this model works satisfactorily if the set of items is restricted to a limited domain, its performance suffers when items of mixed-type are introduced, such as test questions that span multiple academic disciplines.

The Bayesian modeling approach we propose allows increased flexibility with respect to current methods for analyzing educational data. In particular, we obtain joint estimation of (i) associations among questions and concepts, (ii) learner concept knowledge profiles, and (iii) underlying question difficulties. Current methods for analyzing educational data typically perform factor and cluster analyses separately, either to highlight different structures in the data or as part of two steps procedures. We show that performing factor analysis while clustering the population of interest into groups of individuals characterized by homogeneous patterns of underlying factors (i.e., groups of learners with comparable skill sets) improves the predictive performance of the model. Moreover, the assumption that all subjects are equally *reliable* (i.e., two students with the same concept mastery exhibit the same variability when answering questions) is commonly made in models for educational data. In contrast, by including a subject-specific precision parameter, we are able to obtain a more realistic representation of a student's ability and to improve the interpretability of the results. Another key aspect of our model is in its flexibility to infer the number of concepts from the data itself. This has been previously unexplored in the literature on educational data. Finally, missing values are readily handled within our Bayesian paradigm. This allows us, for instance, to impute whether a learner would answer an unattempted question correctly or not.

The remainder of the paper is organized as follows. Details regarding the fully Bayesian model and prior distributions are given in Section 2. Section 3 presents our MCMC method for posterior inference and analysis. Section 4 presents the applications, including a simulation study and results from experimental data. Section 5 provides some concluding remarks. The appendix contains technical details regarding our implementation.

2. Hierarchical Bayes model

In this section, we develop a modeling framework for joint factor and cluster analysis of datasets where multiple categorical response items are collected on a heterogeneous population of individuals. We start by introducing a latent factor multinomial probit model. Then, we discuss prior constructions that allow inference on the number of factors as well as the clustering of subjects into homogeneous groups of relevant factors. We also discuss prior distributions for the other model parameters and a mechanism for the imputation of missing data.

2.1. Latent factor probit model

Consider data from several subjects on a number of ordinal variables. For illustration, we investigate graded answers to a number of assessment items (questions) by a number of learners. A common approach to model such data is via a multinomial probit regression, where the probability of an observed outcome is modeled through the use of the normal cumulative distribution function. Let W_{ij} denote the response variable for subject (learner) $i = 1, \ldots, N$ on variable (question) $j = 1, \ldots, D$. For simplicity, we first examine the binary case, where W_{ij} can take values 0 or 1. We follow the data augmentation approach of McCullagh (1980) and Albert and Chib (1993), and assume that W_{ij} is a surrogate for a latent, continuous random variable, Y_{ij}, for individual i and item j, such that $W_{ij} = 1$ if $Y_{ij} > 0$, and 0 otherwise. Under the probit model, we assume that

$$P(W_{ij} = 1) = \Phi(Y_{ij}; 0, \psi_i^{-1}), \tag{1}$$

where $\Phi(\cdot)$ denotes the inverse probit link function, which maps a real value to a probability via the cumulative distribution function of the normal distribution, and ψ_i^{-1} is a subject-specific variance parameter.

Next, we assume that Y_{ij} is characterized as a linear combination of K underlying factors, i.e.

$$Y_{ij} = \tilde{\boldsymbol{\lambda}}_j^T \boldsymbol{\gamma}_i + \mu_j + \theta_i, \quad \forall i, j, \tag{2}$$

where $\boldsymbol{\gamma}_i$, $i \in \{1, \ldots, N\}$ denotes a K-dimensional column vector of latent factors, γ_{ki}, and $\tilde{\boldsymbol{\lambda}}_j$ is a vector of K real valued elements $\tilde{\lambda}_{jk}$, representing the factor loading of factor k with respect to item j. In the following, we assume that the elements $\tilde{\lambda}_{jk}$ are non-negative, so that larger values denote stronger involvement of the factor. This assumption holds, in particular, in our applications to educational data. In addition, we include a D-dimensional vector of means, $\boldsymbol{\mu}$, where each element, μ_j, represents the random effect for item j, and an N dimensional vector of random effects, $\boldsymbol{\theta}$, with each element θ_i representing the random effect for subject i. In matrix form, (2) can be summarized as

$$\boldsymbol{Y} = \tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma} + \boldsymbol{\mu}\mathbf{1}^T + \boldsymbol{\theta}\mathbf{1}, \tag{3}$$

with $\boldsymbol{Y}$ the $D \times N$ matrix of latent $Y_{i,j}$ and where $\boldsymbol{\Gamma}$ and $\tilde{\boldsymbol{\Lambda}}$ indicate the $K \times N$ matrix of latent factors and the $D \times K$ matrix of factor loadings, respectively, $\boldsymbol{\mu}$ is the D-dimensional vector matrix of random effects μ_j, and $\boldsymbol{\theta}$ is the N-dimensional vector matrix of random effects θ_i.

In the general setting, the latent factor probit regression model can handle ordered, polychotomous data. Here, the response, W_{ij}, takes one of C values, coded as $1, \ldots, C$. Then, we consider a latent variable Y_{ij} and posit that

$$W_{ij} = c \quad \text{if } Y_{ij} \in (\xi_{c-1}, \xi_c], \tag{4}$$

where $\{\xi_0, \ldots, \xi_C\}$ is an ordered set of real valued cutoff points, $-\infty = \xi_0 < \xi_1 < \xi_2 < \cdots < \xi_{C-1} < \xi_C = \infty$.

2.2. Infinite factor models via the Indian Buffet Process

The number of latent factors in (2) is generally not known a priori and selecting a reasonable value for this parameter is often difficult. To overcome this challenge, model selection methods such as cross-validation, BIC or DIC are often employed (e.g. Lee and Song, 2002; Lopes and West, 2004). In general, the number of latent factors, K, should be small relative to both the number of subjects, N, and the number of variables, D. Moreover, every factor may not affect every variable, i.e., $\tilde{\boldsymbol{\Lambda}}$ may not be fully populated. For such reasons, most approaches in the Bayesian parametric literature rely on mixture prior distributions that promote sparsity (West, 2003; Zhang et al., 2004; Carvalho et al., 2008; Henao and Winther, 2009).

An alternative approach is to employ nonparametric Bayesian models that automatically infer the number of factors K based solely on the available data, while enforcing sparsity through the use of variable selection priors. Here, we follow the approach of Knowles and Ghahramani (2011) for infinite factor models and break the latent features matrix into the product of a binary matrix $\boldsymbol{Z}$, indicating which concepts are present for each variable, and a matrix $\boldsymbol{\Lambda}$, capturing the effects of the associations between factors and variables. That is, we write $\tilde{\boldsymbol{\Lambda}} = \boldsymbol{Z} \odot \boldsymbol{\Lambda}$, where $\odot$ denotes the Hadamard (element-wise) matrix product. Assuming a truncated normal prior for the non-zero elements of $\boldsymbol{\Lambda}$, this product construction implies a mixture prior distribution of the type

$$\lambda_{jk} \sim Z_{jk}\, \mathrm{N}^+(0, \tau_k^{-1}) + (1 - Z_{jk})\, \delta_0,$$

where $\mathrm{N}^+(0, \tau_k^{-1})$ is a normal distribution with mean 0 and factor-specific precision τ_k truncated below at 0, and δ_0 is a point mass at 0.

As $\boldsymbol{Z}$ is unknown, it requires a prior distribution. We employ the Indian Buffet Process (IBP). The IBP is a stochastic process defining a probability distribution over sparse binary matrices with a finite number of rows (here, D) and an unbounded number of columns (Griffiths and Ghahramani, 2005; Ghahramani et al., 2007). This prior provides a means to learn the binary matrix without fixing the number of factors. Assume we have a finite number of columns, K. We say that feature k affects the jth row of $\boldsymbol{Y}$ if $Z_{jk} = 1$. Each dimension includes feature k independently with probability π_k, and can include multiple features. We place a Bernoulli distribution on each Z_{jk}

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \prod_{j=1}^{D} p(Z_{jk} \mid \pi_k) = \prod_{k=1}^{K} \pi_k^{m_k} (1 - \pi_k)^{D - m_k},$$

with $m_k = \sum_{j=1}^{D} z_{jk}$, the number of rows influenced by the kth factor. We then define a beta prior on π_k,

$$\pi_k \mid \alpha \overset{\text{i.i.d.}}{\sim} \text{Beta}\left(\frac{\alpha}{K}, 1\right). \tag{5}$$

Marginalizing over π_k and taking the limit for $K \to \infty$, we obtain

$$p(\boldsymbol{Z} \mid \alpha) = \frac{\alpha^{K_+} e^{-\alpha H_D}}{\prod_{h=1}^{2D-1} K_h!} \prod_{k=1}^{K_+} \frac{(D - m_k)!(m_k - 1)!}{D!},$$

where $H_D = \sum_{j=1}^{D} \frac{1}{j}$ is the Dth harmonic number, K_+ is the number of columns where $m_k > 0$, and K_h is the number of columns with pattern h.

An alternative representation of the IBP is characterized via a theoretical buffet with a possibly infinite number of dishes. The first customer chooses a number of dishes according to a Poisson(α). The ith subsequent customer samples previously sampled dishes with probability m_k/i, where m_k is the number of customers who have already sampled dish k. Then the customer considers new dishes according to a Poisson(α/i). Looking at the last customer, the probability $z_{ik} = 1$ given $\boldsymbol{z}_{-ik}$ is m_{-ik}/D, where $m_{-ik} = \sum_{s \neq i} z_{sk}$. Thus, the parameter α in (5) controls the number of features per dimension, as well as the total number of features.

2.3. Clustering subject-specific factors

Grouping subjects with similar latent factors can provide insights on the characteristics of the population. We assume that the latent factors arise from a mixture of normal densities. One possible strategy is to consider a finite mixture of L normals, where each component has a K-variate normal distribution with mean $\boldsymbol{\varphi}_l$ and covariance matrix $\boldsymbol{I}_K$:

$$p(\boldsymbol{\gamma}_i \mid \boldsymbol{\varphi}) = \sum_{l=1}^{L} \pi_l N_K(\gamma_i \mid \boldsymbol{\varphi}_l, \boldsymbol{I}_K),$$

and impose a conjugate Dirichlet prior on $\{\pi_l\}$. One major assumption for this model is that each vector of latent factors arises from one of the L mixture components, which has a distinct mean to capture the distribution of the factors assigned to it. However, the choice of the number of distinct components is not necessarily apparent. The possibly infinite dimensional models involving Dirichlet process (DP) priors (Ferguson, 1973) are the most widely used alternative to finite mixture models.

In our setting, we can regard the DP as a prior distribution specified on the space of all cumulative distribution functions (CDFs) on the real line. If a CDF G is a realization from a DP, we write $G \sim \mathrm{DP}(\beta G_0)$. Here, G_0 is a known base (or mean) distribution and β is a positive scalar which acts as a precision parameter that controls the variability of the random CDF G about G_0. By using the Pólya urn characterization of the DP (Blackwell and MacQueen, 1973), the $\boldsymbol{\gamma}$ are drawn as follows

$$\frac{\beta}{\beta + N} G_0 + \frac{1}{\beta + N} \sum_{\ell=1}^{L} n_\ell \, \delta_{\boldsymbol{\gamma}_\ell^*}(\cdot), \tag{6}$$

where $\boldsymbol{\gamma}_\ell^*$ denote the $\ell = 1, \ldots, L$ distinct values of $\boldsymbol{\gamma}$ and n_ℓ denotes the number of elements currently assigned to the ℓth cluster. Thus, with probability $\frac{\beta}{\beta+N}$, $\boldsymbol{\gamma}_i$ will be drawn from G_0, otherwise, with probability $\frac{n_\ell}{\beta+N}$, it will be set to one of the distinct values, $\boldsymbol{\gamma}_\ell^*$.

We define a Dirichlet process mixture of normal distributions to model the distribution of the γ_i, with a base distribution $G_0 = N(\boldsymbol{0}, \boldsymbol{I}_K)$ (Ferguson, 1983; MacEachern and Müller, 1998). This results in simultaneous inference on the latent factors as well as on the number of underlying groups within users. In the resulting clustering, each user assigned to a given cluster is characterized by a common distribution of the latent factors. The random user effect θ_i in (3) captures extra individual variation with respect to that explained by the cluster assignments. We place a normal prior on these subject random effects, with mean m_θ and variance v_θ.

2.4. Prior distributions for model parameters

We complete the specifications of the model by assuming computationally convenient prior distributions on the remaining parameters of interest. The model can then be fully summarized as follows:

$$\begin{aligned}
&P(W_{ij} = 1) = \Phi(Y_{ij}; 0, \psi_i^{-1}),\\
&Y_{ij} = \boldsymbol{\lambda}_j^T \boldsymbol{\gamma}_i + \mu_j + \theta_i\\
&\boldsymbol{\gamma}_i \mid G \sim G\\
&G \sim \mathrm{DP}(\beta, G_0),\\
&\psi_i \sim \mathrm{Gamma}(a_\psi, b_\psi),\\
&\mu_j \sim \mathrm{N}(m_\mu, v_\mu),\\
&\theta_i \sim \mathrm{N}(m_{\boldsymbol{\theta}}, v_{\boldsymbol{\theta}}),\\
&\lambda_{jk} | \tau_k, Z_{jk} \sim Z_{jk} \mathrm{N}(0, \tau_k^{-1}) + (1 - Z_{jk})\, \delta_0,\\
&\boldsymbol{Z} \sim \mathrm{IBP}(\alpha),\\
&\beta \sim \mathrm{Gamma}(a_\beta, b_\beta),\\
&\alpha \sim \mathrm{Gamma}(a_\alpha, b_\alpha),\\
&\tau_k \sim \mathrm{Gamma}(a_\tau, b_\tau),
\end{aligned} \tag{7}$$

for all $i = 1, \ldots, N$ and $j = 1, \ldots, D$. Here, $G_0 = \mathrm{N}_K(\mathbf{0}, \boldsymbol{I}_K)$ and $(a_\psi, b_\psi, m_{\boldsymbol{\mu}}, v_{\boldsymbol{\mu}}, m_{\boldsymbol{\theta}}, v_{\boldsymbol{\theta}}, a_\tau, b_\tau, a_\beta, b_\beta, a_\alpha, b_\alpha)$ are fixed hyperparameters.

2.5. Missing values

Survey data often contains missing entries. For example, in the evaluation of questionnaires in education, missing data may be due to the possibility that either students or teachers decide to skip some of the questions. Therefore, not every students' response to each question may be observed in the data. Hence, the set of observations, denoted by Ω_{obs}, is a proper subset of $\{1, \ldots, N\} \times \{1, \ldots, D\}$. As contended in Little and Rubin (1987) and Rubin (1996), ignoring the potential data can lead to biases. Instead, we take a Bayesian approach to handle the missing data and incorporate our uncertainty about the unobserved data. In essence, we treat the missing values as parameters and sample the probable responses for a given student's answers based on the observed data. In doing this, we avoid complex estimation algorithms since, conditioned on our estimated responses, $\boldsymbol{W}$ is now considered completely observed. This greatly simplifies the posterior sampling steps for the remaining parameters of interest.

3. Posterior inference

In this section we briefly describe the sampling algorithm for posterior inference, then discuss identifiability issues and ways to obtain posterior estimates of the parameters of interest.

3.1. Markov chain Monte Carlo algorithm

We employ a Markov chain Monte Carlo (MCMC) algorithm to obtain samples from the joint posterior distribution of the model parameters. We outline the algorithm below and report full details of the sampling procedure in the Appendix. At each iteration:

1. Sample any missing values in $\boldsymbol{W}$.
2. Update $\boldsymbol{Y}$ from the truncated normal full conditional.
3. Update $\boldsymbol{\mu}$ from the normal full conditional.
4. Update $\boldsymbol{\theta}$ from the normal full conditional.
5. For each j in $1, \ldots, D$
   - Update each (z_{jk}, λ_{jk}), $k = 1, \ldots, K$, marginally for z_{jk} then $\lambda_{jk} \mid z_{jk}$.
   - Propose the addition of k_j new factors with a Metropolis–Hastings step.
6. Update the current $\boldsymbol{\Gamma}$ to adapt to the current value of K.
7. For each i in $1, \ldots, N$
   - Sample each $\boldsymbol{\gamma}_i$ from either the base distribution, G_0, or assign it to a current cluster value.
8. Reshuffle the distinct cluster means for $\boldsymbol{\Gamma}$.
9. Propose new cutoff values $\boldsymbol{\xi}$ via a Metropolis–Hastings step, if applicable.
10. Update the precision parameters $\{\tau_k\}$ and $\{\psi_i\}$, the IBP parameter α, and the DP parameter β from their respective gamma full conditionals.

3.2. Identifiability

It is well known that both ordinal data and factor analysis models suffer from several identifiability issues (Johnson and Albert, 1999; Lopes and West, 2004). First, identifiability problems arise under certain scaling and shifting of the latent parameters. In our method, for example, one can shift the cutoff positions $\boldsymbol{\xi}$ by some constant while simultaneously shifting the intercept parameters $\boldsymbol{\mu}$ by the same constant without affecting the overall likelihood. Additionally, one can arbitrarily scale the factor loadings $\boldsymbol{\Lambda}$ while inversely scaling the factor scores $\boldsymbol{\Gamma}$ by the same amount. We follow Johnson and Albert (1999) and mitigate many of these difficulties by imposing proper priors on the latent factors as well constraining the first cutoff position ξ_1 to 0. We additionally constrain the first user precision ψ_1 to 1.

A more serious concern in many applied contexts is that factor analysis models are unidentifiable under any permutation of the latent factors (Lopes and West, 2004). Concretely, one can jointly permute the factors of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ without affecting the overall likelihood. This is commonly referred to as "label-switching". If not mitigated properly, the label switching problem can severely complicate posterior analysis. Here we recommend post-processing of the MCMC output, similarly to what was done in the mixture model literature (see Stephens, 2000). Let $\boldsymbol{\Lambda}^t$, $\boldsymbol{\Gamma}^t$, and $\boldsymbol{\mu}^t$ denote the tth samples from the MCMC. We first compute the posterior probability $p(\boldsymbol{W} | \boldsymbol{\Lambda}^t, \boldsymbol{\Gamma}^t, \boldsymbol{\mu}^t)$ and then select the iteration $t_{\max}$ that maximizes this probability. We then permute the factors $\boldsymbol{\Lambda}^t$, $\boldsymbol{\Gamma}^t$ obtained over all iterations $t \neq t_{\max}$ to best match $\boldsymbol{\Lambda}^{t_{\max}}$, $\boldsymbol{\Gamma}^{t_{\max}}$. Performing this step aligns the posterior samples to a common reference, enabling more meaningful posterior analysis, such as the computation of posterior means.

Table 1
Simulated data: Frobenius loss (with standard error) of the proposed IBP + DP model versus a simple IBP model for various data sizes $N_0 = N = D$, $L_0 = 3$ and binary responses. Results are averaged over 50 simulated datasets.

	$E_{\mathbf{Y}}$ (IBP + DP)	$E_{\mathbf{Y}}$ (IBP)
$N_0 = 50$	0.208 (0.0848)	0.472 (0.134)
$N_0 = 75$	0.136 (0.0437)	0.459 (0.139)
$N_0 = 100$	0.0995 (0.0403)	0.402 (0.104)
$N_0 = 200$	0.0598 (0.0239)	0.355 (0.113)

3.3. Posterior estimates

A posteriori, we are interested in the estimation of (i) the associations among questions and concepts, (ii) the learner concept knowledge profiles and (iii) the underlying question difficulties. These associations are captured in our model via the parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\mu}$, respectively.

At each iteration of the MCMC algorithm, the number of active features can change. We perform posterior inference by first estimating K and L via the posterior mode, say K^+ and L^+, and considering only those iterations where K and L are equal to K^+ and L^+, respectively. Then we obtain inference on the other model parameters based on the selected subset of MCMC iterations. For example, estimates for the means and quantiles of each cell of $\boldsymbol{\Gamma}$ are easily calculated. In addition, given the estimates of the posterior probability of inclusion (PPIs), for each cell of $\boldsymbol{Z}$, estimates of the factor loadings in $\boldsymbol{\Lambda}$ can be calculated by thresholding the PPIs and setting to zero the λ_{jk} that correspond to those PPIs smaller than a certain threshold, while estimating the others via the posterior mean.

4. Experiments

Here we assess the performance of our approach on simulated data as well as on real-world educational datasets.

4.1. Synthetic data

We first examine synthetic data generated under various settings. In each setting, we fix the number of latent variables K and the number of latent clusters L. Each entry of the support matrix $\boldsymbol{Z}$ is generated i.i.d. with $Z_{jk} \sim \text{Ber}(0.5)$. Each user is assigned to one of the L clusters uniformly at random. We then generate $\gamma_\ell \sim \text{N}_K(\mathbf{0}, \mathbf{I}_K)$ for each $\ell = 1, \ldots, L$. The remaining parameters are generated as in (7), with $a_\psi = 5$, $b_\psi = 5$, $m_{\boldsymbol{\mu}} = 0$, $v_{\boldsymbol{\mu}} = 0.5$, $m_{\boldsymbol{\theta}} = 0$, $v_{\boldsymbol{\theta}} = 0.5$, $a_\tau = 5$, and $b_\tau = 5$. After generating the synthetic data, we conduct model fitting and obtain posterior distributions for all model parameters using the MCMC sampling techniques described in Section 3.1. We consider broad priors for the specification of the parameters in the nonparametric priors. More specifically, we set $a_\alpha = 5$, $b_\alpha = 1$, $a_\beta = 5$, $b_\beta = 1$, which allows for adequate exploration of the posterior space. The fixed hyperparameters used in model fitting are identical to those used in the data generation. The posterior samples are analyzed as described in Section 3.2 and the relevant posterior estimates (e.g., posterior means) are computed as outlined in Section 3.3. In the following, for simplicity we refer to our method as the IBP + DP method.

We start by assessing the performance of our model relative to increasing data sizes. More specifically, we consider a binary response variable, i.e. we fix $C = 2$ in (4), and generate the data under the assumption of $K = 3$ factors and $L_0 = 3$ subject specific clusters. The sample size and number of items for the different settings are, respectively, $N_0 = N = D \in \{50, 75, 100, 200\}$.

We evaluate the performance of our model with respect to the true latent data $\boldsymbol{Y}$ using a normalized Frobenius loss metric, which is commonly employed in the factor analysis literature (Lan et al., submitted for publication; Hahn et al., 2012). This metric is defined as

$$E_{\boldsymbol{Y}} = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_F^2 / \|\boldsymbol{Y}\|_F^2. \tag{8}$$

Table 1 displays the mean normalized Frobenius loss (with standard error) over 50 simulated datasets. The results are compared against a version of our method that still estimates the number of latent factors non-parametrically from the data, but it does not allow clustering of the users. We refer to this method simply as IBP. The results show the advantage provided by our model if individual factors are truly clustered. The accuracy of both models improves with increasing sample sizes, as expected, but the improvement is more evident for the proposed IBP + DP than for the simpler IBP model. Table 2 shows similar results for ordinal responses with $C = 5$ outcome categories. We note that, in all trials, the posterior mode of L and K corresponds identically to the ground truth.

Next, we consider the problem of imputing missing data when only a subset of $\boldsymbol{W}$ is observed. For exploring the accuracy of the missing data sampling mechanism, we set $K = 3$, $L_0 = 3$ and $N = D = 100$ and consider the case of binary response data. We then remove a portion of the data, and obtain posterior MCMC estimates, imputing the missing values as described in 2.5. The subset of the observed data $\boldsymbol{W}$ retained in the different settings is selected by i.i.d. draws from a Bernoulli distribution with observation (success) probability p_{obs}, which is set at values, respectively, $\{0.5, 0.6, 0.7, 0.8, 0.9\}$.

Table 2
Simulated data: Frobenius loss (with standard error) of the proposed IBP + DP model versus a simple IBP model for various data sizes $N_0 = N = D$, $L_0 = 3$ and ordinal responses with $C = 5$. Results are averaged over 50 simulated datasets.

	$E_{\mathbf{Y}}$ (IBP + DP)	$E_{\mathbf{Y}}$ (IBP)
$N_0 = 50$	0.511 (0.931)	0.774 (0.998)
$N_0 = 75$	0.460 (0.874)	0.688 (0.883)
$N_0 = 100$	0.258 (0.483)	0.621 (0.514)
$N_0 = 200$	0.151 (0.151)	0.568 (0.562)

Table 3
Simulated data: imputation error $E_{\mathbf{W}}$ at various observation rates for both the IBP+DP and IBP methods. Results are averaged over 50 simulated datasets.

	$E_{\mathbf{W}}$ (IBP + DP)	$E_{\mathbf{W}}$ (IBP)
$p_{\text{obs}} = 0.5$	0.202 (0.020)	0.208 (0.020)
$p_{\text{obs}} = 0.6$	0.199 (0.021)	0.205 (0.020)
$p_{\text{obs}} = 0.7$	0.197 (0.021)	0.202 (0.021)
$p_{\text{obs}} = 0.8$	0.196 (0.021)	0.201 (0.021)
$p_{\text{obs}} = 0.9$	0.193 (0.022)	0.197 (0.023)

Table 4
Frobenius loss $E_{\mathbf{Y}}$ and standard error for the IBP + DP model as a function of data sizes $N_0 = N = D$ and ground truth cluster L_0. The results are computed over 50 randomized trials.

	$L_0 = 2$	$L_0 = 5$	$L_0 = 10$	$L_0 = N_0$
$N_0 = 50$	0.208 (0.0848)	0.28 (0.09)	0.324 (0.114)	0.375 (0.0972)
$N_0 = 75$	0.136 (0.0437)	0.167 (0.0468)	0.181 (0.0664)	0.239 (0.0469)
$N_0 = 100$	0.0995 (0.0403)	0.114 (0.0335)	0.127 (0.0372)	0.176 (0.0239)
$N_0 = 200$	0.0598 (0.0239)	0.0531 (0.0112)	0.0477 (0.008)	0.089 (0.0129)

We evaluate performance using the following imputation error metric:

$$E_{\mathbf{W}} = \frac{1}{|\Omega_{\text{obs}^c}|} \sum_{(i,j)\in\Omega_{\text{obs}^c}} |W_{ij} - W_{ij}^{\text{mode}}|,$$

where W_{ij}^{mode} is the posterior mode of the W_{ij}'s MCMC samples. Table 3 reports results averaged over 50 simulated datasets. The IBP + DP method outperforms the IBP method across all values of p_{obs}.

We further consider the performance of the IBP + DP as the number of underlying clusters of latent factors varies. We again consider binary responses and vary both the data sizes $N_0 = N = D \in \{50, 75, 100, 200\}$ and the number of clusters $L_0 \in \{3, 5, 10, N_0\}$ of the generated data. The case of $L_0 = N_0$ corresponds to the case where there are no clusters in the data. Table 4 displays results in terms of the Frobenius loss for the matrix of factors $\mathbf{\Gamma}$. Our method shows improved performance for increasing data sizes and for decreasing number of clusters. This is in accordance to expectations, since fewer clusters generally imply less diversity in the data which, in turn, enables better estimation of the underlying factors. However, our method, which seeks out structure in $\mathbf{\Gamma}$, shows good performances also when no such structure exists ($L_0 = N_0$). This is also to be expected given that the Bayesian Nonparametric prior can easily adapt to account for such situations.

In order to quantify the performance of the Bayesian Nonparametric clustering, we compute a measure of clustering misclassification rate for our method. Quantifying misclassification is difficult due to the label switching phenomenon, in which cluster labels can change over iterations. Further complicating the issue is that the number of ground truth clusters (say L_0) may be different than the number of clusters (say $\hat{L}$) revealed by the estimation method. In order to overcome those difficulties, we use the confusion matrix (Stehman, 1997), which provides a standard technique for dealing with label switching in misclassification tasks in the machine learning community. The confusion matrix computes the local misclassification error that would be incurred by associating each of the $\hat{L}$ post-estimation clusters with the L_0 ground truth clusters. By doing this, one can compute the optimal relabeling of clusters that minimizes the global misclassification rate E_{class} in a greedy fashion. For each simulated dataset, we compute the average value of E_{class} over all iterations of the MCMC taken post-burnin. We repeat this experiment over 50 randomized datasets and display our results in Table 5. Once again, we see that performance improves with increasing sample sizes and when decreasing the number of clusters.

4.2. Educational data

We now turn to real educational data for learning and content analytics. In each case, we examine the factors estimated by our method and what these factors reveal about the different response patterns observed in the data.

Table 5
Misclassification error E_{class} and corresponding standard errors for the IBP + DP method, for varying data sizes $N_0 = N = D$ over a range of ground truth cluster L_0. The results are computed over 50 randomized trials.

	$L_0 = 2$	$L_0 = 5$	$L_0 = 10$	$L_0 = N_0$
$N_0 = 50$	0.249 (0.166)	0.433 (0.156)	0.545 (0.111)	0.499 (0.0473)
$N_0 = 75$	0.160 (0.162)	0.321 (0.145)	0.460 (0.118)	0.491 (0.053)
$N_0 = 100$	0.086 (0.109)	0.225 (0.133)	0.344 (0.116)	0.487 (0.053)
$N_0 = 200$	0.051 (0.127)	0.071 (0.076)	0.150 (0.076)	0.408 (0.077)

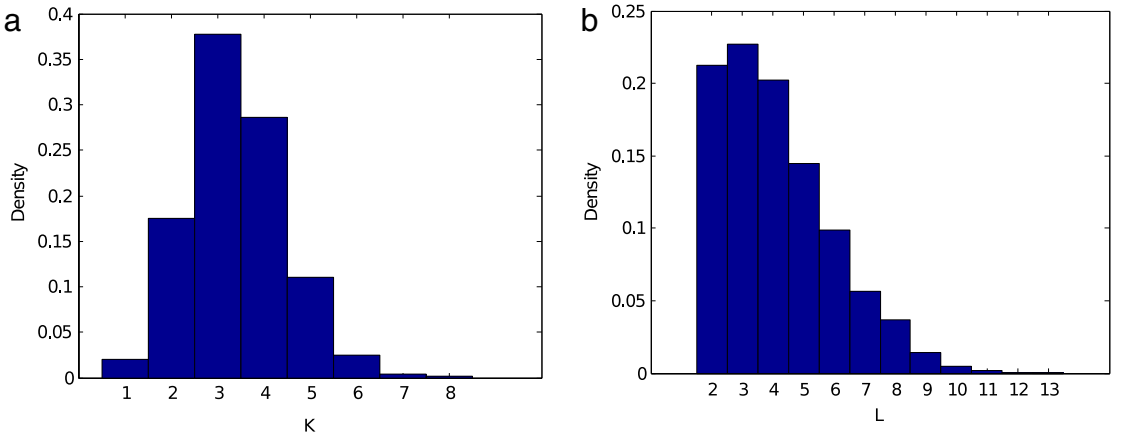

Fig. 1. Educational data: probability and statistics class results: (a) posterior distribution of the number of concepts K and (b) posterior distribution of the number of clusters L.

4.2.1. Probability and statistics course

We first consider a dataset consisting of an introductory course in probability and statistics taught at the Georgia Institute of Technology and administered by OpenStax Tutor (2014). This course consists of 89 questions answered by 42 students over the course of one semester. The questions have been collected from homeworks as well as from two mid-term examinations. We employ our method on this dataset and post-process our results as described in Section 3.3. We display histograms of K and L in Fig. 1. Our method explores many values both for the number of latent concepts and for the latent clusters. However, we find that choosing $K^+ = 4$ and $L^+ = 3$ is sufficient to capture salient features of this dataset.

We next examine the posterior mean of $\mathbf{\Lambda}$ and $\boldsymbol{\mu}$. First we show a heat map of the posterior mean of $\mathbf{\Lambda}$ in Fig. 2(a). Next we display the associations between questions and concepts as a bipartite graph in Fig. 2(b). In the bipartite graph, the concepts are visualized as (yellow) circles and the questions are displayed as (gray) boxes. The posterior mean of $\mathbf{\Lambda}$ connects questions to concepts, with the line thickness providing a visual summary of the amplitude of the respective $\mathbf{\Lambda}_{jk}$. The posterior mean of $\boldsymbol{\mu}$ for each question is displayed inside of each gray box. From the analysis of the bipartite graph, it is evident that many of the questions in this dataset do not appear to be related to any particular concept. Indeed, the probability that students answer successfully any of these questions appears to be modeled sufficiently well by considering only their latent ability, θ_i, and the intrinsic difficulty of the question. Such information is extremely useful for the examiners, as they would be able to determine if the questions are well-posed and adequately test the target concept, and, if needed, accordingly revisit the questionnaire.

Finally, we show a two-dimensional principal components projection of $\mathbf{\Gamma}$ in Fig. 3 for the posterior mode case of $L^+ = 3$. The components are rotated to maximize the correlation between the projected mastery vectors and the number of problems answered correctly. The main cluster consists of 33 learners with strong mastery of all subject material. The two remaining clusters consist of learners with varying degrees of mastery of the various course concepts. This clustering information is valuable to course instructors as it identifies groups in the class who struggle on similar portions of the material. A course instructor armed with this information could readily identify and specifically address the learning difficulties of the subpopulations of students who are struggling with different topics in the course.

4.2.2. University admissions test

We next consider a dataset for a 2013 timed University entrance examination first examined in Vats et al. (2014). This dataset consists of 1567 high school students answering 60 questions distributed evenly across four major subject areas: biology, chemistry, mathematics, and physics. Each of these subject areas cover a larger number of concepts (e.g., the mathematics portion includes concepts such as set theory, algebra, calculus, and combinatorics).

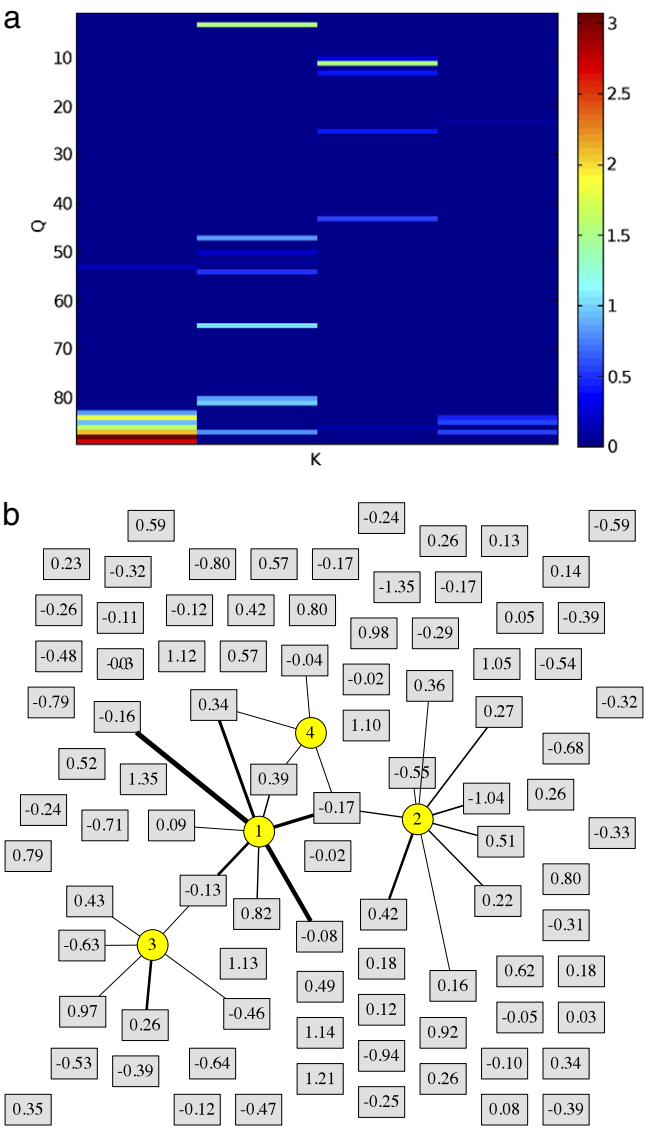

Fig. 2. Educational data: probability and statistics class results: (a) posterior mean of $\mathbf{\Lambda}$ and (b) bi-partite graph of the content. Yellow circles denote concepts and gray boxes denote questions. The numbers inside the boxes represent the posterior mean of the intrinsic difficulty μ_j for each question.

The exam is graded in a way similar to the American SAT test. For this exam, a student receives 3 points for correctly answering a question, loses 1 point for incorrectly answering a question, and receives 0 points if they choose not to respond. As expected, this grading procedure results in a number of students choosing not to respond to certain questions. For this dataset, 29% of the total 60×1567 question–answer pairs are unobserved.

We employ our IBP + DP model on this dataset to infer both the number of latent concepts as well as the number of latent student clusters from the data. We display posterior histograms of K and L in Fig. 4. Our method finds that $K = 19$ latent concepts and $L = 116$ clusters of students provide a sufficiently good fit to the data.

We further display a heat map of the inferred $\mathbf{\Lambda}$ in Fig. 5(a) as well as a bipartite graph that connects concepts to questions in Fig. 5(b). The inferred $\mathbf{\Lambda}$ shows significant agreement with the underlying exam questions. Concretely, the first 15 questions of the exam cover biology-related topics, and these questions are found by our model to generally share the same latent concept. Questions 29 and 30 concern the reactions of organic compounds and our method finds that they share multiple latent concepts. Mathematics and physics cover questions 31–45 and 46–60, respectively, and are also found by our method to share their own latent concept.

We display two-dimensional projection of the student clusters contained in $\mathbf{\Gamma}$ in Fig. 6. Examination of the raw data shows that these basis vectors correspond roughly to aptitude in biology and math/physics.

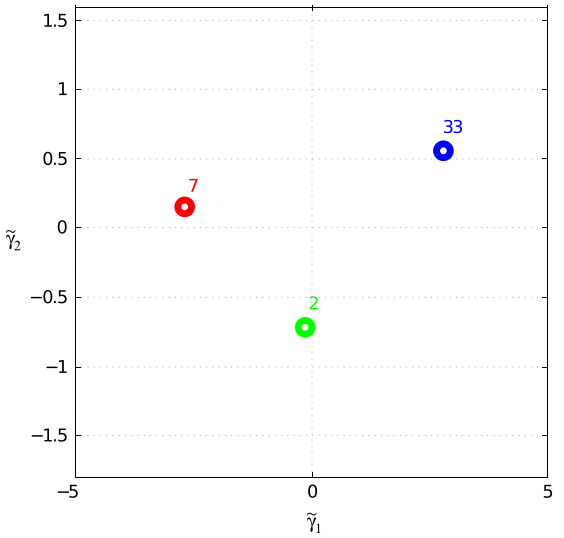

Fig. 3. Educational data: probability and statistics class results: Two-dimensional projection of the posterior mean of $\mathbf{\Gamma}$ taken over samples for which $L^+ = 3$ and $K^+ = 4$. Each circle corresponds with one cluster, with the adjacent numeral denoting the number of learners in the cluster.

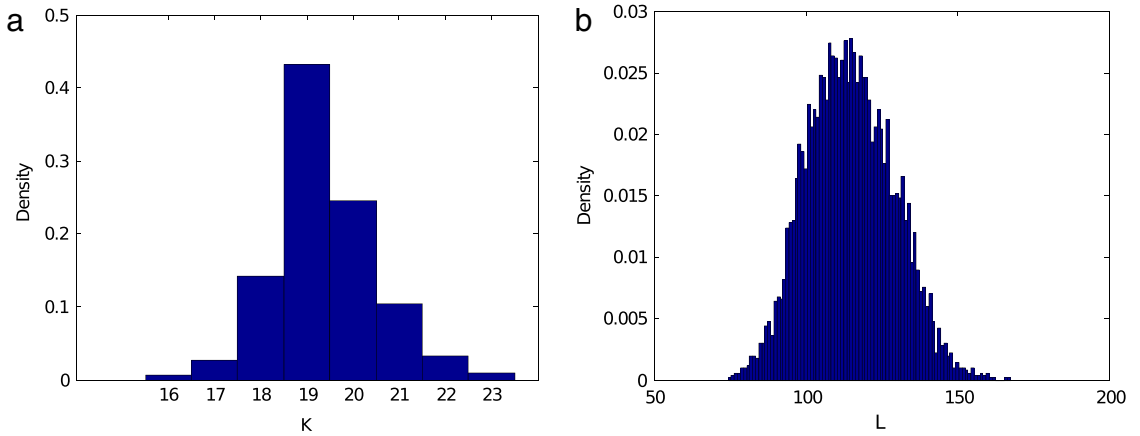

Fig. 4. IISER admissions test: (a) posterior distribution of the number of concepts K and (b) posterior distribution of clusters L.

Finally, we consider the role of missing value sampling for this dataset. Due to the scoring system and time constraint for this exam, students must strategize which questions they choose to answer. Assume that student i only has time to answer Q_i questions on this exam given the time constraint. The end-goal for each student is to choose the Q_i that they are most likely to answer correctly, while avoiding the questions that they feel they are likely to answer incorrectly.

It is well known from the cognitive psychology literature, however, that students are notoriously poor judges of their own concept mastery (Koriat and Levy-Sadot, 2001; Reder, 1987; Reder and Ritter, 1992). This cognitive bias will cause them to often use poor judgment when selecting which questions to answer. Each student, we might surmise, could potentially achieve a higher score on the exam if they did not suffer from this cognitive bias and instead chose to answer the actual questions for which they were most likely to succeed.

Therefore, we can use the missing value imputation abilities of our method to quantify how much of a performance improvement we could expect for each student if this cognitive bias were removed and students chose the optimal set of problems to answer. Let p_{ij} denote the unknown probability of success for student i on question j. Conditioned on the set of problems that student i chooses to answer, we can compute a student specific expected score S_i on the overall exam as follows

$$S^i = \mathbb{E}\left[\sum_{(i,j)\in\Omega_{\text{obs}}} 3\cdot\mathbb{I}(Y_{ij}=1) - 1\cdot\mathbb{I}(Y_{ij}=0)\right]$$
$$= \sum_{(i,j)\in\Omega_{\text{obs}}} 3\cdot\mathbb{E}[\mathbb{I}(Y_{ij}=1)] - 1\cdot\mathbb{E}[\mathbb{I}(Y_{ij}=0)]$$

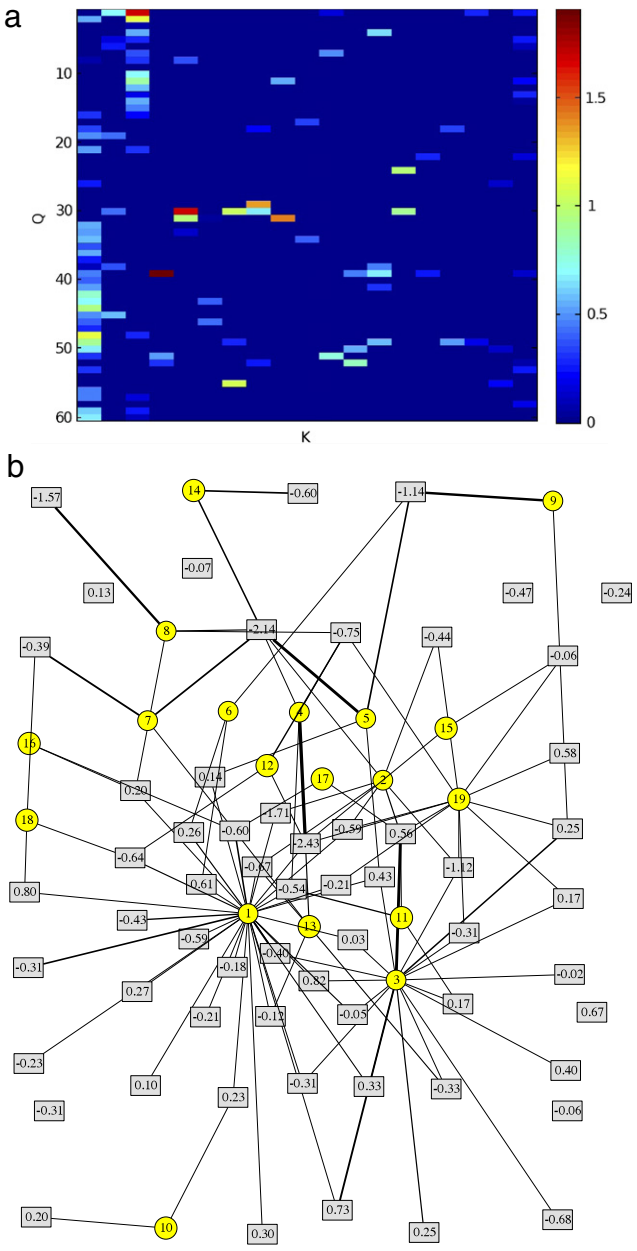

Fig. 5. IISER admissions test: (a) posterior mean of $\mathbf{\Lambda}$ and (b) Bi-partite graph of the content. Yellow circles denote concepts and gray boxes denote questions. The numbers inside the boxes represent the posterior mean of the intrinsic difficulty μ_j for each question.

$$= \sum_{(i,j)\in\Omega_{\text{obs}}} \left(3 \cdot p_{ij} - (1 - p_{ij})\right)$$

$$= \sum_{(i,j)\in\Omega_{\text{obs}}} \left(4 \cdot p_{ij} - 1\right). \tag{9}$$

Suppose now that each student, instead of choosing the Q_i questions they actually answered, could choose the Q_i questions with the largest values of p_{ij} (i.e., the questions for which the student is most likely to succeed). We can provide an estimate of the true p_{ij} by considering the posterior predictive mean of p_{ij} based on the MCMC estimates. Therefore, we can estimate the impact of the cognitive bias on each individual student's score by computing the expected score S^i in (9) for the Q_i questions with highest posterior predictive means p_{ij}'s and compare with the student's observed final score. Carrying out this procedure shows that, on average, without the cognitive bias, students would improve their test score by over 12 points, corresponding to an increased percentile ranking of 10%.

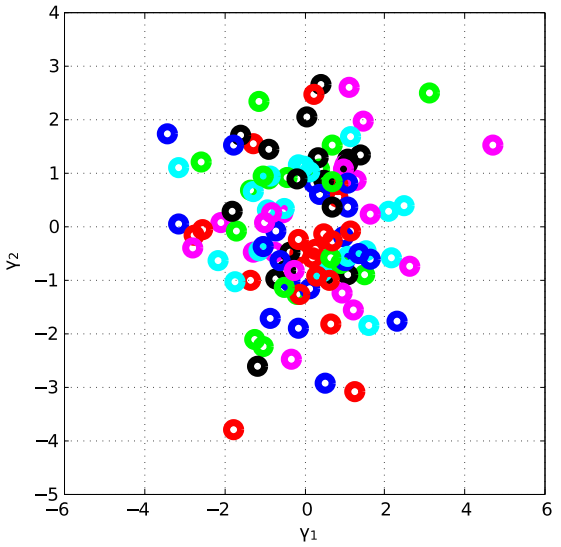

Fig. 6. IISER admissions test: two-dimensional projection of the posterior mean of $\mathbf{\Gamma}$ taken over samples for which $L^+ = 116$ and $K^+ = 19$. Each circle corresponds with one cluster.

5. Conclusions

We have proposed a Bayesian Nonparametric model for the joint factor and cluster analysis in datasets where multiple categorical response items are collected on a heterogeneous population. Our fully Bayesian method employs two nonparametric priors, for learning the number of latent variables K and for learning the number of subject defined clusters L from the data. By means of simulations, we have shown that the additional structure imposed by our model provides improved accuracy with respect to methods that do not take clustering into account. However, the flexibility of our non-parametric prior specifications ensures that good performance is retained even when the data are not truly clustered.

Automatically inferring clustering among the subject specific factors is important in several applications where users naturally belong to one of several subgroups. In the application on education, we have shown that such clustering allows us to identify groups of learners that could be qualified either as strong or poor performers, according to their patterns of responses in all or for subsets of the questions. Considerations of this type can ultimately lead instructors and schools to tailor their educational approaches to specific groups of students, and therefore lead to better educational outcomes. Other applications of such techniques could be easily found, e.g. for political voting (Eric et al., 2013), marketing and finance (Ando and Bai, 2014) and user recommender systems (Adomavicius and Tuzhilin, 2005; Resnick and Varian, 1997).

Future extensions include incorporating prior information to guide the selection of the relevant factors. For example, knowledge of the learning objectives of a course could potentially inform about the number and structure of the factors identified in the analysis. Furthermore, in many applications, surveys and questionnaires are repeatedly offered to the same group of subjects over time. Future work will explore dynamic joint factor and cluster analytic approaches to study how the association between items and subject specific factors varies longitudinally. For example, in education, a set of exams could be given at the beginning, middle and end of a semester to test a set of learning objectives. Then, the identification of groups of subjects showing substantial improvement in the mastery of the course concepts over time would provide an objective way to assess the efficacy of a teaching approach.

MCMC details

We provide details of the MCMC algorithm for our Bayesian infinite factor model. Given the observations, $\boldsymbol{W}$, we obtain inference for the parameters of interest using a combination of Gibbs sampling and Metropolis–Hastings updates.

1. **Update for $\boldsymbol{W}$**: We need to include possible missing values in $\boldsymbol{W}$. Let $\widetilde{W}_{ij}$ represent a missing answer for learner i at question j with a corresponding latent variable $\widetilde{Y}_{ij}$. Then, the likelihood can be split into observed and unobserved data,

$$p(\boldsymbol{Y} \mid \ldots) = \prod_{i,j \in \Omega_{obs}} \text{Bern}(W_{ij}; \Phi(Y_{ij}; 0, \psi_i^{-1})) \prod_{i,j \notin \Omega_{obs}} \text{Bern}(\widetilde{W}_{ij}; \Phi(\widetilde{Y}_{ij}; 0, \psi_i^{-1})).$$

The $\widetilde{Y}_{ij}$ are readily integrated out and, therefore, we sample the $\widetilde{W}_{ij}$ from a Bernoulli distribution with probability $\Phi(\boldsymbol{\lambda}_j \boldsymbol{\gamma}_i + \mu_j + \theta_i, \psi_i^{-1})$. Conditional on the sampled values, the rest of the updates are carried out assuming we have a fully observed $\boldsymbol{W}$.

2. **Update for $\boldsymbol{Y}$**: The latent variables, Y_{ij}, are updated from a truncated normal distribution with mean $\boldsymbol{\lambda}_j\boldsymbol{\gamma}_i + \mu_j + \theta_i$ and variance ψ_i^{-1}. This truncated normal distribution is truncated below by $\xi_{W_{ij}-1}$ and above by $\xi_{W_{ij}}$.
3. **Update for $\boldsymbol{\mu}$**: The full conditional for μ_j follows a normal distribution with mean $s^*\left(m_\mu/v_\mu + \sum_{i=1}^N \psi_i(\boldsymbol{Y}_i - \boldsymbol{\lambda}_j\boldsymbol{\gamma}_i - \theta_i)\right)$ and variance $s^* = (\sum_i \psi_i + v_\mu^{-1})^{-1}$.
4. **Update for $\boldsymbol{\theta}$**: The full conditional for θ_i follows a normal distribution with mean $s^*\left(m_\theta/v_\theta + \sum_{j=1}^D \psi_i(\boldsymbol{Y}_j - \boldsymbol{\lambda}_j\boldsymbol{\gamma}_i - \mu_j)\right)$ and variance $s^* = (D\psi_i + v_\theta^{-1})^{-1}$.
5. **Joint update for $(\boldsymbol{Z}, \boldsymbol{\Lambda})$**: The jkth element of the binary, IBP matrix, Z_{jk}, has a prior ratio of

$$\frac{\Pr(Z_{jk} = 1 \mid \ldots)}{\Pr(Z_{jk} = 0 \mid \ldots)} = \frac{m_{-jk}}{D - m_{-jk}}$$

where m_{-jk} counts the number of questions, excluding j, for which concept k is active. The likelihood given $z_{jk} = 1$ requires integrating over the truncated normal prior on λ_{jk}. Consequently, with τ_k is the precision of factor k and $\boldsymbol{E}_j = \boldsymbol{Y}_j - \mu_j\boldsymbol{1} - \boldsymbol{\theta}$, the ratio of likelihoods is given by

$$\frac{P(\boldsymbol{Y} \mid z_{jk} = 1, \ldots)}{P(\boldsymbol{Y} \mid z_{jk} = 0, \ldots)} = (\tau_k\sigma^*)^{1/2} \exp\left\{\frac{1}{2\sigma^*}\mu^{*2}\right\}\left(1 - \Phi(0; \mu^*, \sigma^*)\right),$$

where $\sigma^* = (\sum_i \psi_i * \gamma_{ki}^2 + \tau_k)^{-1}$, $\mu^* = \sigma^* \sum_i \psi_i\gamma_{ki}E_{ij}$, and $\widetilde{\boldsymbol{\Lambda}}$ is the $\boldsymbol{\Lambda}$ matrix with the jkth cell set to 0.

Multiplying the ratios of prior and likelihood gives the ratio of posterior probabilities to be used for sampling z_{jk}. Then, if $z_{jk} = 1$, we sample λ_{jk} from a truncated normal with mean μ^* and variance σ^*.

In order to add new concepts, we must sample the number of concepts active only for question j (call this k_j). We can integrate over the new elements of the mixing matrix, $\boldsymbol{\lambda}_{jk_j}$, or the new rows of the latent feature matrix, $\boldsymbol{\gamma}_{k_ji}$, but not both.

Given that $\boldsymbol{\gamma}_{k_j}$ will generally be of higher dimension than $\boldsymbol{\lambda}_{k_j}$, we choose to integrate over the $\boldsymbol{\gamma}_{k_j}$. In this case, the new elements of $\boldsymbol{\Lambda}$ are added to the proposal distribution, $J(k_j)$ is as follows

$$J(k_j) = \left\{(1-p)\text{Pois}\left(k_j; \frac{c\alpha}{D-1}\right) + p1_{k_j=1}\right\}\text{N}^+(\boldsymbol{\lambda}_{k_j}; \boldsymbol{0}, \boldsymbol{\tau}_{k_j}^{-1}).$$

Therefore, the proposal is accepted with probability $r = \min\{1, a_l a_p\}$. Here, $a_p = \text{Pois}(k_j; \frac{\alpha}{D-1})/\text{Pois}(k_j; \frac{c\alpha}{D-1})$ and $a_l = p(\boldsymbol{Y} \mid k_j, \boldsymbol{\lambda}_{k_j}, \ldots)/p(\boldsymbol{Y} \mid \ldots)$. The expression for a_l is given by

$$\prod_i |\boldsymbol{\Sigma}_i^*|^{-1/2} \exp\left\{\frac{1}{2}\sum_{i=1}^N \boldsymbol{m}_i'^* \Sigma_i^* \boldsymbol{m}_i^*\right\}$$

where $\boldsymbol{\Sigma}_i^* = [\boldsymbol{\lambda}_{k_j}\boldsymbol{\lambda}_{k_j}' * \psi_i + \boldsymbol{I}_{k_j}]$ and $\boldsymbol{m}_i^* = \boldsymbol{\Sigma}_i^{*-1}\boldsymbol{\lambda}_{k_j}\hat{e}_{ij}$ with $\hat{e}_{ij} = (Y_{ij} - \boldsymbol{\lambda}_j\boldsymbol{\gamma}_i - \mu_j)\psi_i$.

6. **Update $\boldsymbol{\Gamma}$**: The full conditional updates for $\boldsymbol{\gamma}_i$ are done via a multi-step process. First, we sample $\boldsymbol{\Gamma}$ based on newly sampled $\boldsymbol{\Lambda}$. This is done by first removing any rows of $\boldsymbol{\Gamma}$ that pertained to columns of $\boldsymbol{\Lambda}$ that were removed. Next, we divide $\boldsymbol{\Gamma}$ into a set $\boldsymbol{\Gamma}_{\text{old}}$ consisting of elements of $\boldsymbol{\Gamma}$ that were active previously and a set $\boldsymbol{\Gamma}_{\text{new}}$ consisting of the elements $\boldsymbol{\Gamma}$ pertaining to the newly added columns of $\boldsymbol{\Lambda}$ such that $\boldsymbol{\Gamma} = \{\boldsymbol{\Gamma}_{\text{old}}, \boldsymbol{\Gamma}_{\text{new}}\}$. We then sample $\boldsymbol{\Gamma}_{\text{new}}|\boldsymbol{Y}, \boldsymbol{\Gamma}_{\text{old}}, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\theta}$ from the full conditional for each set of i in the same cluster.

Next, we sample $\boldsymbol{\gamma}_i \mid \boldsymbol{\gamma}_{-i}, \cdot$, for $i = 1, \ldots, N$ under the DP prior. This is done sequentially for each $\boldsymbol{\gamma}_i$. We sample $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_\ell^*$ with probability proportional to $n_\ell \cdot P(Y_i \mid \boldsymbol{\gamma}_i = \boldsymbol{\gamma}_\ell^*, \cdot)$ and is drawn from the full conditional posterior $P(\boldsymbol{\gamma}_i \mid \cdot)$ with probability proportional to $\beta \cdot P(\boldsymbol{Y}_i)$, where $P(\boldsymbol{Y}_i)$ is the marginal likelihood of the ith column of $\boldsymbol{Y}$ defined by:

$$P(\boldsymbol{Y}_i) = \int_{\boldsymbol{\gamma}_i} P(\boldsymbol{Y} \mid \boldsymbol{\gamma}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\theta})\pi(\boldsymbol{\gamma}_i)d\boldsymbol{\gamma}_i,$$

which, given our choice of prior, can be computed in closed form.

Finally we perform a reshuffling step on $\boldsymbol{\Gamma}$ by drawing from the full conditional for each cluster.

7. **Update for $\boldsymbol{\xi}$**: Cutoff positions are sampled via a Metropolis–Hastings step. Concretely, we sample ξ_c for all $c = 2, \ldots, C-1$ using the following proposal distribution $\xi_c^0 \sim \text{N}^+(\xi_c, \sigma_{\text{MH}}^2, \xi_{c-1}, \xi_{c+1})$, where the cutoff values (ξ_{c-1}, ξ_{c+1}) enforce the ordering constraint on the cutoff positions.

The accept/reject ratio is given as follows:

$$R = \left(\prod_{i=1}^N\prod_{j=1}^D \frac{\Phi\left(\sqrt{\psi_i}(\xi_{W_{i,j}} - Y_{i,j})\right) - \Phi\left(\sqrt{\psi_i}(\xi_{W_{i,j}-1} - Y_{i,j})\right)}{\Phi\left(\sqrt{\psi_i}(\xi^0_{W_{i,j}} - Y_{i,j})\right) - \Phi\left(\sqrt{\psi_i}(\xi^0_{W_{i,j}-1} - Y_{i,j})\right)}\right)$$
$$\times \left(\prod_{c=2}^{C-1} \frac{\Phi((\xi_{c+1} - \xi_{i,c})/\sigma_{\text{MH}}) - \Phi((\xi^0_{c-1} - \xi_c)/\sigma_{\text{MH}})}{\Phi((\xi^0_{c+1} - \xi^0_c)/\sigma_{\text{MH}}) - \Phi((\xi_{c-1} - \xi^0_c)/\sigma_{\text{MH}})}\right),$$

where the first term corresponds the likelihood ratio while the second accounts for the non-symmetric transition probability of the proposal distribution. To make the final acceptance decision, we generate $U \sim \text{Unif}(0, 1)$ and accept if $U \leq R$.

8. **Update for** $\{\psi_i\}$, $\{\tau_k\}$, α, **and** β: The full conditionals for the ψ_i follow a gamma distribution with shape parameter $a_\psi + D/2$ and rate parameter $b_\psi + \sum_j (\boldsymbol{Y}_{ij} - \boldsymbol{\lambda}_j \boldsymbol{\gamma}_i - \mu_j - \theta_i)^2$.

   The concept precisions, τ_k, are given the same Gamma prior, and therefore have Gamma full conditionals with shape and rate parameters $a_\tau + \frac{m_k}{2}$ and $b_\tau + \frac{1}{2}\sum_j \lambda_{jk}^2$, where m_k is the number of questions for which concept k is active.

   The full conditional for the IBP parameter, α, given the conjugate Gamma prior, follows a Gamma distribution with shape parameter $K^+ + a_\alpha$ and rate parameter $b_\alpha + H_D$, where $H_D = \sum_{j=1}^{D} \frac{1}{j}$ is the Dth harmonic number.

   Finally the DP parameter β is sampled as described in Escobar and West (1995). Concretely, we define the variable $\pi = (a_\beta + L^+ - 1)/(a_\beta + L^+ - 1 + N \cdot (b_\beta - \log(x)))$, with $x \sim \text{Beta}(\beta + 1, N)$, using the previous sample of β. We then draw a uniform random variable $U \in [0, 1]$ and draw the new sample $\beta \sim \text{Gamma}(\beta + L^+, b_\beta - \log(x))$ for $U \leq \pi$ and draw $\beta \sim \text{Gamma}(\beta + L^+ - 1, b_\beta - \log(x))$ for $U > \pi$.

Acknowledgments

This work was supported by the National Science Foundation under Cyberlearning grant IIS-1124535, the Air Force Office of Scientific Research under grant FA9550-09-1-0432, and the Google Faculty Research Award program.

The authors would like to express their gratitude to the Chairman, JAC, IISER Pune, for sharing the educational data, as well as Divyanshu Vats for insightful discussion regarding this dataset.

References

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. 17 (6), 734–749.
Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Amer. Statist. Assoc. 88 (422), 669–679.
Ando, T., Bai, J., 2014. Asset pricing with a general multifactor structure, J. Finance Econom. (forthcoming).
Blackwell, D., MacQueen, J., 1973. Ferguson distributions via pólya urn schemes. Ann. Statist. 1, 353–355.
Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M., 2008. High-dimensional sparse factor modeling: applications in gene expression genomics. J. Amer. Statist. Assoc. 103 (484), 1439–1456.
Eric, W., Salazar, E., Dunson, D., Carin, L., 2013. Spatio-temporal modeling of legislation and votes. Bayesian Anal. 8 (1), 233–268.
Escobar, M., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–588.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1, 209–230.
Ferguson, T., 1974. Prior distributions on spaces of probability measures. Ann. Statist. 2, 615–629.
Ferguson, T., 1983. Bayesian density estimation by mixtures of normal distribution. In: Rizvi, M.H., Siegmund, D. (Eds.), Recent Advances In Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday. Academic Press, New York, pp. 287–302.
Ghahramani, Z., Griffiths, T., Sollich, P., 2007. Bayesian nonparametric latent feature models. In: Bayesian Statistics, vol 8. Oxford University Press, Oxford.
Griffiths, T., Ghahramani, Z., 2005. Infinite latent feature models and the indian buffet process. Technical Report, GCNU TR 2005-001.
Hahn, P.R., Carvalho, C.M., Scott, J.G., 2012. A sparse factor analytic probit model for congressional voting patterns. J. R. Stat. Soc. Ser. C Appl. Stat. 61 (4), 619–635.
Henao, R., Winther, O., 2009. Bayesian sparse factor models and dags inference and comparison. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems, vol. 22. pp. 736–744.
Johnson, V., Albert, J., 1999. Ordinal Data Modeling. Springer.
Knowles, D., Ghahramani, Z., 2011. Nonparametric bayesian sparse factor models with application to gene expression modeling. Ann. Appl. Stat. 5 (2B), 1534–1552.
Koriat, A., Levy-Sadot, R., 2001. The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing.. J. Exp. Psychol. Learn. Mem. Cognit. 27 (1), 34.
Kulik, J.A., 1994. Meta-analytic studies of findings on computer-based instruction. In: Baker, E., O'Neil, H.F., O'Neil, H.F. (Eds.), Technology Assessment in Education and Training. Routledge.
Lan, A.S., Waters, A.E., Studer, C., Baraniuk, R.G., Sparse factor analysis for learning and content analytics, J. Mach. Learn. Res. (submitted for publication).
Lee, S., Song, X., 2002. Bayesian selection on the number of factors in a factor analysis model. Behaviormetrika 29, 23–40.
Li, N., Cohen, W.W., Koedinger, K.R., 2011. A machine learning approach for automatic student model discovery, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, pp. 31–40.
Little, R., Rubin, D., 1987. Statistical Analysis with Missing Data. J. Wiley & Sons, New York.
Lopes, H., West, M., 2004. Bayesian model assessment in factor analysis. Statist. Sinica 14, 41–67.
West, M., 2003. Bayesian factor regression models in the "large p, small n" paradigm. In: Bayesian Statistics. Oxford University Press, pp. 723–732.
MacEachern, S.N., Müller, P., 1998. Estimating mixtures of Dirichlet process models. J. Comput. Graph. Statist. 7, 223–238.
McCullagh, P., 1980. Regression models for ordinal data. J. R. Stat. Soc. Ser. B 42, 109–127.
Murray, R.C., VanLehn, K., Mostow, J., 2004. Looking ahead to select tutorial actions: A decision-theoretic approach. Int. J. Artif. Intell. Educ. 14 (3–4), 235–278.
OpenStax Tutor 2014. OpenStax tutor. URL http://openstaxtutor.org/.
Rai, P., Daumé III, H., 2008. The infinite hierarchical factor regression model. In: Neural Information Processing Systems. MIT Press, Cambridge, MA.
Rasch, G., 1993. Probabilistic Models for Some Intelligence and Attainment Tests. MESA Press.
Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer Publishing Company Incorporated.
Reder, L.M., 1987. Strategy selection in question answering. Cogn. Psychol. 19 (1), 90–138.
Reder, L.M., Ritter, F.E., 1992. What determines initial feeling of knowing? Familiarity with question terms, not with the answer. J. Exp. Psychol. Learn. Mem. Cognit. 18 (3), 435.
Resnick, P., Varian, H.R., 1997. Recommender systems. Commun. ACM 40 (3), 56–58.
Rubin, D., 1996. Multiple imputation after 18+ years (with discussion). J. Amer. Statist. Assoc. 91, 473–489.
Stamper, J.C., Barnes, T., Croy, M., 2007. Extracting student models for intelligent tutoring systems. In: Proceedings of the National Conference on Artificial Intelligence, vol. 22. pp. 113–147.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. 1, 77–89.
Stephens, M., 2000. Dealing with label switching in mixture models. J. R. Stat. Soc. Ser. B 62 (4), 795–809.
Vats, D., Studer, C., Lan, A., Baraniuk, R., 2014. Test-size reduction via sparse factor analysis. J. Educ. Data Min. (under review).
Zhang, Z., Chan, K., Kwok, J., Yeung, D., 2004. Bayesian inference on principal component analysis using reversible jump Markov chain Monte Carlo. In: Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, California, USA. AAAI Press, pp. 372–377.