A note on large-scale logistic prediction

Using an approximate graphical model to deal with collinearity and missing data

Marsman, M.; Waldorp, L.; Maris, G.

DOI
10.1007/s41237-017-0024-x

Publication date
2017

Document Version
Final published version

Published in
Behaviormetrika

Behaviormetrika (2017) 44:513–534
DOI 10.1007/s41237-017-0024-x

NOTE

A note on large-scale logistic prediction: using an approximate graphical model to deal with collinearity and missing data

Maarten Marsman[1] · **Lourens Waldorp**[2] · **Gunter Maris**[3]

Received: 30 September 2016 / Accepted: 1 June 2017 / Published online: 14 June 2017

Abstract Large-scale prediction problems are often plagued by correlated predictor variables and missing observations. We consider prediction settings in which logistic regression models are used and propose a novel approach to make accurate predictions even when predictor variables are highly correlated and only partly observed. Our approach comprises three steps: first, to overcome the collinearity issue, we propose to model the joint distribution of the outcome variable and the predictor variables using the Ising network model. Second, to render the application of Ising networks feasible, we use a latent variable representation to apply a low-rank approximation to the network's connectivity matrix. Finally, we propose an approximation to the latent variable distribution that is used in the representation to handle missing observations. We demonstrate our approach with numerical illustrations.

Keywords Logistic regression · Ising model · IRT model

Communicated by: Brandon Malone.

✉ Maarten Marsman
m.marsman@uva.nl

Lourens Waldorp
l.waldorp@uva.nl

Gunter Maris
g.k.j.maris@uva.nl

[1] Psychological Methods, University of Amsterdam, Nieuwe Prinsengracht 129-B, 1018 VZ Amsterdam, The Netherlands

[2] Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands

[3] Psychological Methods and Cito, Psychometric Research Centre, University of Amsterdam, Amsterdam, The Netherlands

1 Introduction

Most large-scale or big data applications involve conditional models that utilize covariates to make predictions about a variable of interest. For instance, Google needs to predict which links to websites will be most advantageous based on millions of previous clicks, and Netflix needs to predict movie preferences based on millions of previous viewings and rankings. In these applications the interest is not in explaining why the connections between websites or movies exist, but in predicting which website will be most often requested or which movie will be preferred by an individual user. We will focus on the prediction problem where both the outcome variable and the covariates are binary, and the logistic regression model is an appropriate statistical model.

As is the case with all regression models, we observe that the logistic regression model is developed for situations where the covariates are independent and completely observed. However, a different situation is usually observed in large-scale applications, where covariates are typically correlated. As a consequence of the correlations between covariates, i.e., collinearity, the obtained set of coefficients is no longer unique. This can be seen in, for instance, the coordinate descent algorithm (Hastie et al. 2015), where each covariate is treated separately. For two equivalent covariates, any solution with a linear combination of the two (normalized) coefficients is correct, even when regularization is applied. This is certainly an issue for the identification of relevant covariates, i.e., variable selection. In a particular sample, one of the collinear covariates will have a slightly larger coefficient and, therefore, ends up in the solution, while the other does not. But in another sample it could be the other way around. This means that variable selection with collinear covariates is unreliable. We will illustrate that collinearity is not a problem for prediction.

Another issue is that in most large-scale applications covariates are only partially observed (e.g., Rubin 1976; Rousseeuw 2016). For estimation and variable selection, it is then pertinent to know in which way the data came to be missing. For instance, data could be missing completely at random, which means that there is no connection between the missing observations and the data generating process. But data could also be missing precisely because of the data-generating process. For instance, a response to the question"do you drink more on average than others in your circle of friends" will be missing if a negative response is observed to the question "do you take alcohol". In such cases, conditional on taking alcohol, the two covariates will be correlated, and so the missing observations cannot be ignored. We illustrate that predictions based on partially observed data can still be accurate, even when the process that generates missing data cannot be ignored in a statistical sense.

Our goal in this paper was to introduce a novel approach to make accurate predictions with the logistic regression model when covariates are highly correlated and only partly observed. Our approach comprises three steps: First, we propose to model the joint distribution of the outcome variable and the predictor variables with the Ising model. In the Ising model the correlations between the observed variables are explicitly modeled, which overcomes the collinearity issue. Second, we use recent results that relate Ising networks to latent variable models to render the application of

Ising models computationally tractable. Specifically, we use a low-rank approximation to the network's connectivity matrix, which is opportune when variables are highly correlated (i.e., collinearity). Finally, we propose to approximate the latent variable distribution in the representation of the Ising model, which results in a model-based approximation to the full Ising model that is able to handle missing observations. Numerical illustrations are used to demonstrate different features of our approach.

2 Step I: the Ising model to overcome collinearity

For prediction, we are interested in the conditional distribution $\mathbb{P}(x_i \mid x_{\setminus i})$, where x_i is an element of $x \in \{-1, +1\}^{\mathrm{p}}$ and $x_{\setminus i}$ is the vector x excluding the ith element. Even though we are only interested in the predictive distribution $\mathbb{P}(x_i \mid x_{\setminus i})$, our observations are the realizations of a multivariate random variable x. The multivariate distribution $\mathbb{P}(x)$ that is consistent with the logistic regression model is the Ising network model (Lenz 1920; Ising 1925),

$$\mathbb{P}(x) = \frac{\exp\left(x^{\mathsf{T}}\Sigma x + x^{\mathsf{T}}\mu\right)}{\sum_x \exp(x^{\mathsf{T}}\Sigma x + x^{\mathsf{T}}\mu)},$$

where $\Sigma \in \mathbb{R}^{p\times p}$ is a symmetric matrix of pairwise interaction parameters σ_{ij}, and $\mu \in \mathbb{R}^p$ a vector of main effects. Observe that the correlations between elements in x are no problem for the Ising model as their interactions σ_{ij} are explicitly modeled. That is, there is no collinearity issue when estimating the full Ising model $\mathbb{P}(x)$. From the joint distribution we can then obtain the correct full-conditional $\mathbb{P}(x_i \mid x_{\setminus i})$, and it is easily seen that the full-conditional distribution that is obtained from the Ising model is a logistic regression model:

$$\mathbb{P}(x_i \mid x_{\setminus i}) = \frac{\exp\left(x_i\left[\mu_i + 2\sum_{j\neq i}\sigma_{ij}x_j\right]\right)}{\sum_{x_i}\exp\left(x_i\left[\mu_i + 2\sum_{j\neq i}\sigma_{ij}x_j\right]\right)}.$$

Hence, we can overcome the collinearity issues with logistic regression by estimating the full Ising model.

However, estimating an Ising model proves to be far more complex than estimating a logistic regression model. A first problem is the number of parameters that needs to be estimated for the Ising model. Whereas the number of parameters is linear in the number of covariates p for the logistic regression model, it is quadratic in p for the Ising model. A second problem is that the density of the Ising model is computationally intractable, except for small or heavily constrained networks. This computational burden is due entirely to the model's normalizing constant:

$$Z = \sum_x \exp\left(x^{\mathsf{T}}\Sigma x + x^{\mathsf{T}}\mu\right),$$

which is the sum over all 2^p possible realizations of x. Thus, even though we have resolved the collinearity issue with the Ising model, we have also increased the

number of parameters with an order of magnitude and need to deal with estimating a model that is computationally intractable.

3 Step II: low-rank approximations for computational tractability

A latent variable representation of the Ising model, in combination with a low-rank approximation to the full-connectivity matrix Σ, is the two crucial ingredients to render large-scale applications of the Ising model entirely tractable. The latent variable representation of the Ising model was introduced by Kac (1968). Specifically, Kac showed that every eigenvector of the connectivity matrix Σ generates a latent variable, such that the manifest random variables x are independent given the full set of latent variables η. Since the diagonal elements from the connectivity matrix Σ are not identifiable from the data, we decompose it as

$$\Sigma + cI_p = Q(\Lambda + cI_p)Q^{\mathsf{T}} = UU^{\mathsf{T}},$$

where Λ is a diagonal matrix consisting of the eigenvalues of the original connectivity matrix, and the translation by c serves to ensure that all eigenvalues are positive, i.e., ensuring that UU^{T} is positive (semi-)definite and at the same time preserve the off-diagonal elements of Σ. The latent variable representation of Kac then follows immediately from a clever use of the Gaussian identity:

$$\exp\left(x^{\mathsf{T}}UU^{\mathsf{T}}x + x^{\mathsf{T}}\mu\right) = \int_{\mathbb{R}^p} \frac{1}{\sqrt{\pi}^p} \exp\left(2x^{\mathsf{T}}U\eta + x^{\mathsf{T}}\mu - \eta^{\mathsf{T}}\eta\right) \mathrm{d}\,\eta$$

This latent variable representation has been further developed by Emch and Knops (1970) and has been independently rediscovered many times (e.g., Olkin and Tate 1961; Besag 1974; McCullagh 1994; Anderson and Vermunt 2000).

It was recently shown (Marsman et al. 2015; Epskamp et al. 2017) that the associated conditional distribution $\mathbb{P}(x \mid \eta)$ is the multidimensional item response theory (MIRT) model (Reckase 2009)

$$\mathbb{P}(x_i \mid \eta) = \frac{\exp\left(x_i\left[2u_i^{\mathsf{T}}\eta + \mu_i\right]\right)}{\sum_{x_i}\exp\left(x_i\left[2u_i^{\mathsf{T}}\eta + \mu_i\right]\right)},$$

with u_i being the ith column of U^{T}. MIRT models are frequently used in psychological and educational measurement (Ackerman et al. 2003; Ackerman 1996), where the observed variables correspond to item responses on some test or questionnaire, and the latent variables relate to the trait or abilities being assessed (Borsboom and Molenaar 2015). Importantly, this representation inspired a full-data-information estimation procedure that avoids having to compute the Ising model's intractable normalizing constant (Marsman et al. 2015).

A second crucial ingredient is the low-rank approach that Marsman et al. (2015) proposed to approximate the full connectivity matrix, such that the number of parameters becomes linear in p. Their low-rank approach makes use of the Eckart

and Young Theorem (1936), which states that in a least squares sense the best rank-r approximation to the full connectivity matrix Σ is one in which all but the r largest eigenvalues are equated to zero. Low-rank approximations have become increasingly popular in prediction problems since their crucial role in winning the Netflix price competition (Koren et al. 2009; Bell and Koren 2007; Bell et al. 2010) and it has been part of Google's system ever since the very first implementation of the pageRank algorithm (Page et al. 1999; Brin and Page 2012). Most important for our present endeavors, however, is that a low-rank approximation to the full connectivity matrix is expected to make accurate predictions when predictor variables are highly correlated.

4 Step III: an approximate latent variable distribution for missing data

An important feature of IRT models is that they are closed under marginalization. That is, because the manifest variables x_i are independent given the latent variable η, we find that the marginal,

$$\sum_{x_i \in \{\pm 1\}} \prod_{j=1}^{p} \mathbb{P}(x_j \mid \eta) = \prod_{j \neq i} \mathbb{P}(x_j \mid \eta), \tag{1}$$

is again an IRT model. That the IRT model is closed under marginalization makes it a valuable tool for applications with data that are subject to missing data (Eggen 2004), as one can simply marginalize over the missing observations. In contrast, the Ising model is not closed under marginalization. That is, in general we find that

$$\mathbb{P}(x_{\setminus i}) = \sum_{x_i \in \{\pm 1\}} \mathbb{P}(x_i, \mathbf{x}_{\setminus \mathrm{i}})$$

is itself not an Ising model. Unfortunately, the marginalization property of the IRT model does not transfer to the latent variable expression of the Ising model, which is entirely due to the latent variable distribution $f(\eta)$. To train the Ising model in the face of incomplete data, we, therefore, either have to omit any incomplete cases from the analysis, use imputation techniques to artificially complete the observed data (Rubin 1987), or make use of approximate models that allow for missing data. We take the latter approach and show that we can make reliable predictions using a low-rank approximate model, even in the presence of correlated and missing data.

A key difference between regular applications of IRT models and the latent variable representation of the Ising model is in the prior (or population) distribution of the latent variables that are used. The distribution of latent variables is typically assumed to be multivariate normal, but in the representation of the Ising model it is the mixture:

$$f(\eta) = \frac{1}{\sqrt{\pi^p}\, Z} \prod_i \left[\sum_{x_i} \exp\left(x_i [2 u_i^{\mathsf{T}} \eta + \mu_i]\right) \right] \exp(-\eta^{\mathsf{T}} \eta) = \sum_x \mathbb{P}(x) f(\eta \mid x),$$

where to each of 2^p possible realizations of x we have a multivariate normal posterior distribution with mean $U^{\mathsf{T}} x$ and variance $2 I_p$. Even though this latent variable

model is computationally intractable, it often takes a simple form in each of its dimensions. Specifically, it closely resembles either a single normal distribution with a mean of zero or a mixture of two normal distributions with their respective means placed symmetrically about zero.

That the latent variable distribution tends to have either a single mode or has two modes can be seen by inspecting the derivative of $\log(f(\eta))$ with respect to η (for ease of presentation assuming a single dimension);

$$\frac{\mathrm{d}}{\mathrm{d}\,\eta}\log f(\eta) = 2\sum_{i=1}^{p} u_i \tanh(2u_i\eta + \mu_i) - 2\eta = 2h(\eta) - 2\eta,$$

which shows that the latent variable distribution has modes (and minima) at the fixed points $h(\eta) = \eta$. A plot of $h(\eta)$ against η is shown in Fig. 1, together with the line $\eta = \eta$, which rotates with respect to $h(\eta)$ as a function of the variables in the model. Note that the line can cross the curve $h(\eta)$ either once (at zero) such that there is a single mode, or three times (as depicted here) such that there are two modes (symmetric about zero) and a local minimum at zero. The latent variable distribution $f(\eta)$ can thus be closely approximated by a small mixture of normal distributions.

There are two important consequences of replacing the latent variable distribution of the Ising model with some other latent variable distribution, say $g(\eta)$. A first consequence is that the marginal distribution of the observed variables

$$\mathbb{P}(x) = \int_{\mathbb{R}^p} \mathbb{P}(x \mid \eta) g(\eta) \mathrm{d}\eta,$$

is, in general, not analytically available and requires numeric procedures to compute. A second consequence follows from the indeterminancy of the parameters U, μ (and η) in the MIRT model. By replacing the latent variable distribution $f(\eta)$ with a distribution $g(\eta)$, the parameters are placed on a different scale. This

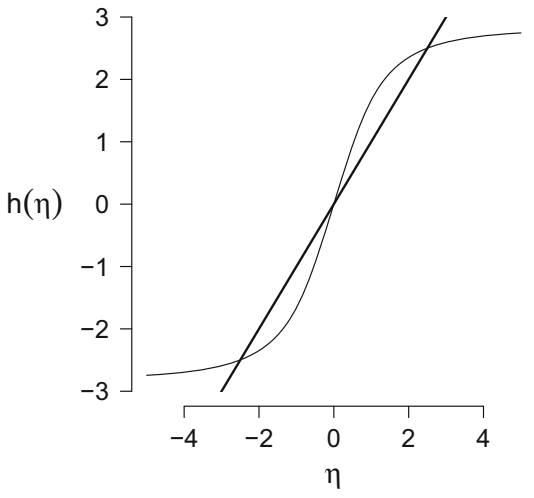

Fig. 1 A plot of η versus $h(\eta)$ illustrating that $f(\eta)$ can have either one or two modes. A local minimum and two local maxima are found on the intersection with the *straight solid line*

indeterminancy does not affect the marginal distribution of observable variables (i.e., predictions), although it does affect parameter recovery.

The validity of our approximate model rests on how well $f(\eta)$ is approximated by $g(\eta)$. To assess the validity of this approach, we can make use of recent advances in plausible value methodology and explicitly consider whether or not the true latent variable distribution is equal (or similar) to $g(\eta)$. Plausible values are draws from the posterior distribution of the latent variables η (Mislevy 1991; von Davier et al. 2009) and are commonly used in large-scale educational surveys to accommodate researchers in the field that are not able to estimate the complex IRT models used for these surveys. Recently, it was shown that the marginal distribution of plausible values is a consistent estimator of the true latent variable distribution $f(\eta)$ (Marsman et al. 2016), meaning that one can assess the validity of using a single multivariate normal distribution by inspecting the (marginal) distribution of plausible values.

5 Numerical illustrations

Below we demonstrate the different aspects of our theory using three broad illustrations. The first illustration aims to showcase that the IRT model is able to make accurate predictions with correlated variables. The second illustration aims to showcase that the IRT model is able to accurately predict both observed and missing observations when the missing data mechanism is ignorable and the data are completely missing at random (MCAR; see Appendix A). We end this illustration with a comparison with logistic regression. The third illustration is used to demonstrate that there is a limit to the IRT model's capacity to accurately predict observed and missing data points when the missing data mechanism is nonignorable (data not missing at random; NMAR). We consider several situations in which we vary the effect of the missing data mechanism on the observed correlations and mix settings with MCAR and MNAR in the training and testing phase. The main results from our illustrations are shown in Fig. 2. We first discuss the methods and models that are used.

5.1 Generating correlated binary data

To generate correlated binary variables we use the Ising model, with a rank ten connectivity matrix Σ that is based on the following eigenvalues:

$$\lambda_\tau = \frac{1}{\tau} \times [1.00, 0.80, 0.65, 0.30, 0.25, 0.20, 0.16, 0.11, 0.06, 0.01].$$

The value of τ modifies the strengths of pairwise correlations of variables in the network. Figure 3 illustrates that the more extreme correlations (i.e., ± 1) occur for smaller values of τ. In the three illustrations we will use the values $\tau = 0.5$ and $\tau = 1.0$. The Ising model's parameters Q and μ will be sampled uniformly between $-0.1/p$ and $0.1 / p$ for the p-variable networks (Q is made orthogonal), where scaling by p^{-1} ensured similar dynamics for the different sized networks.

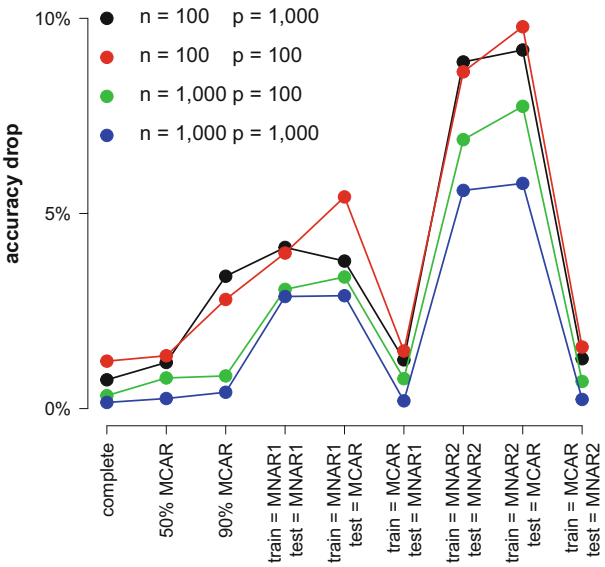

Fig. 2 The main results for applications of a two-dimensional IRT model to data generated from a $\tau = 1.0$ network. The *y-axis* shows the drop in accuracy for the test set predictions as compared to accuracy of the true model. That is, $c_{\text{true}} - c_{\text{test}}$ if the complete data are used, or $c_{\text{true}} - c_{\text{test}}^{(o)}$ when there are missing data. Predictions from the true model always make use of the complete data. The *x-axis* shows the different situations; collinearity (c.f. Table 1), ignorable missingness (denoted MCAR; c.f. Table 2), nonignorable missingness with moderate effects (denoted MNAR1; c.f. Table 5) and nonignorable missingness with severe effects (denoted MNAR2; c.f. Table 6)

Table 1 Prediction accuracy for the two-dimensional IRT model applied to correlated data

n	p	c_{true}	c_{train}	c_{test}
$\tau = 1.0$ network data				
100	1000	81	81	80
100	100	80	80	79
1000	100	81	81	81
1000	1000	82	82	82
$\tau = 0.5$ network data				
100	1000	91	91	90
100	100	91	91	91
1000	100	92	92	92
1000	1000	91	90	89

Data were generated from the Ising model using a Gibbs sampler (Geman and Geman 1984) applied to the joint distribution $f(x,\eta)$ of the latent variables η and the data x. In each iteration of the Gibbs sampler, we sample from the full-conditional posterior distribution $f(\eta \mid x)$ of the latent variables, and the full-conditional distribution $\mathbb{P}(x \mid \eta)$ of the data. Both full-conditional distributions are easy to sample from; the posterior distribution of the latent variable $f(\eta \mid x)$ is a multivariate

Table 2 Prediction accuracy for the two-dimensional IRT model applied to $\tau = 1.0$ network data with ignorable missing observations

n	p	c_{true}	$c^{(o)}_{\text{train}}$	$c^{(m)}_{\text{train}}$	$c^{(o)}_{\text{test}}$	$c^{(m)}_{\text{test}}$
50% missing observations						
100	1000	81	82	81	80	80
100	100	81	82	79	79	79
1000	100	81	80	80	80	80
1000	1000	81	82	81	81	81
90% missing observations						
100	1000	82	86	79	78	77
100	100	81	88	71	78	71
1000	100	80	81	76	79	76
1000	1000	81	82	81	81	80

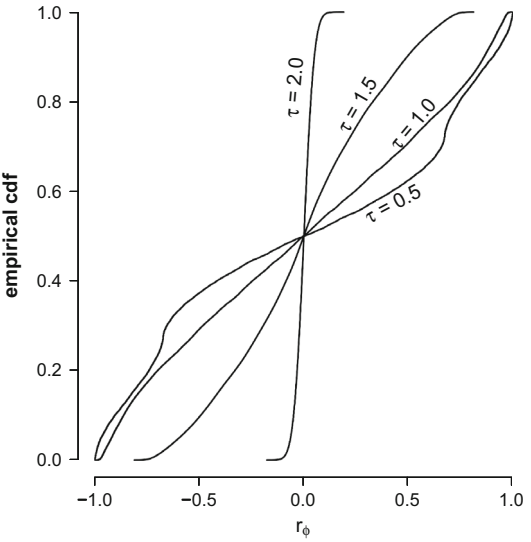

Fig. 3 Distribution of 499,500 correlations—Pearson's-ϕ—in a $p = 1000$ variable network based on $n = 1000$ observations for $\tau \in \{0.5,1.0,1.5,2.0\}$. With the value $\tau = 2.0$ most of the observed correlations are close to zero, and with the value $\tau = 0.5$ most of the observed correlations are near the extremes

normal distribution with mean vector $U^{\mathsf{T}}x$ (where $U^{\mathsf{T}} = Q^{\mathsf{T}}\Lambda^{\frac{1}{2}}$ as before) and a variance-covariance matrix $2I_{10}$, and $\mathbb{P}(x \mid \eta)$ is a ten-dimensional IRT model.

5.2 Estimating the MIRT model

In each simulation we train the two-dimensional IRT model

$$\mathbb{P}(x_i \mid \eta,\, u_i,\, \mu_i) = \frac{\exp(x_i[\mu_i + 2u_{i1}\eta_1 + 2u_{i2}\eta_2])}{\sum_{x_i}\exp(x_i[\mu_i + 2u_{i1}\eta_1 + 2u_{i2}\eta_2])}.$$

We assume a simple bivariate normal distribution for the latent variables; $\eta \sim \mathcal{N}(0,2\mathrm{I}_2)$. We take a Bayesian approach to estimate the IRT model, which requires us to formulate prior distributions for U and μ. We will use logistic prior distributions with location 0 and scale 1 for both U and μ.

A Gibbs sampler is used to produce samples from the joint multivariate posterior distribution of the parameters and latent variables, i.e., $g(U, \mu, \eta \mid x)$. Whereas the full-conditional distribution of the latent variables in the Ising model representation is a completely tractable normal posterior distribution, i.e., $\eta \mid x, U, \mu \sim \mathcal{N}(U^{\mathsf{T}}x, 2I_2)$, this is not the case when we replace the associated latent variable distribution. Specifically, when we utilize a normal prior distribution for the latent variables, we observe that the full-conditional posterior distribution is the intractable:

$$g(\eta \mid x,U,\mu) \propto \prod_{i=1}^{p} \frac{\exp(x_i[\mu_i + 2u_{i1}\eta_1 + 2u_{i2}\eta_2])}{\sum_{x_i} \exp(x_i[\mu_i + 2u_{i1}\eta_1 + 2u_{i2}\eta_2])} \times g(\eta).$$

Similarly, we find that the full-conditional distributions of the "item" parameters U and μ are also intractable:

$$g(u_{ij} \mid x,\eta,\mu) \propto \prod_{v=1}^{n} \frac{\exp(x_{iv}[\mu_i + 2u_{i1}\eta_{1v} + 2u_{i2}\eta_{2v}])}{\sum_{x_i} \exp(x_i[\mu_i + 2u_{i1}\eta_{1v} + 2u_{i2}\eta_{2v}])} \times g(u_{ij})$$

$$g(\mu_i \mid x,\eta,U) \propto \prod_{v=1}^{n} \frac{\exp(x_{iv}[\mu_i + 2u_{i1}\eta_{1v} + 2u_{i2}\eta_{2v}])}{\sum_{x_i} \exp(x_i[\mu_i + 2u_{i1}\eta_{1v} + 2u_{i2}\eta_{2v}])} \times g(\mu_i)$$

where $g(u_{ij})$ and $g(\mu_i)$ are the logistic prior distributions, and v indexes the n observations. The problem of sampling from these full-conditional distributions has been addressed in several places (e.g., Patz and Junker 1999a, b; Maris and Maris 2002). We use a Metropolis approach (Metropolis et al. 1953; Hastings 1970; Tierney 1994) that was specifically designed to handle full-conditional distributions of this form (Marsman et al. 2015, 2017).

5.3 Calculating prediction accuracy

The prediction accuracy can be calculated by 0–1 loss or Bayes risk. Computationally this has the advantage of being easy to compute, but it is also tied to convex alternatives like logistic loss that are asymptotically equivalent to 0–1 loss (see, e.g., Bartlett et al. 2006). In our simulations we use 0–1 prediction accuracy, which is defined as

$$c = c(x_i, x_i^*) = \frac{1}{n}\sum_{v=1}^{n} \mathbb{1}\{x_{vi} = x_{vi}^*\},$$

where $\mathbb{1}$ is the indicator function and we predict x_{vi} using x_{vi}^*. Observe that $c(x_i, x_i^*)$ is the ratio of correct predictions (true positives and true negatives) out of the n predictions that are made.

Since each of the p variables could be used as a dependent variable in Logistic regression—there are p full-conditionals $\mathbb{P}(x_i \mid x_{\setminus i})$—we calculate the prediction accuracy for each of the p variables and then average them. We furthermore repeat each procedure five times and average the results.

5.4 The two stages of our prediction procedure

Our prediction procedure consists of two stages, a training stage and a testing stage.

5.4.1 The training stage comprises six steps

(1) Generate training data x_{train} from the Ising model.
(2) Generate new predictions x^* from the Ising model and compute $c_{\text{true}} = c(x_{\text{train}},\mathrm{x}^*)$.
(3) Split the data into an observed part $x_{\text{train}}^{(O)}$ and a 0%, 50% or 90% missing part $x_{\text{train}}^{(M)}$. The missing part $x_{\text{train}}^{(M)}$ will only be used to evaluate predictions.
(4) Use the Gibbs sampler to estimate the IRT parameters $\theta_{\text{train}} = \{U,\mu\}$ and the latent variables η_{train} using the observed training data $x_{\text{train}}^{(O)}$.
(5) Generate new predictions from the IRT model on the observed part,

$$x_*^{(O)} \sim \mathbb{P}(x \mid \eta_{\text{train}},\theta_{\text{train}}),$$

and compute $c_{\text{train}}^{(O)} = c(x_{\text{train}}^{(O)},\mathrm{x}_*^{(\mathrm{O})})$.
(6) Generate new predictions from the IRT model on the missing part,

$$x_*^{(M)} \sim \mathbb{P}(x \mid \eta_{\text{train}},\theta_{\text{train}}),$$

and compute $c_{\text{train}}^{(M)} = c(x_{\text{train}}^{(M)},\mathrm{x}_*^{(\mathrm{M})})$.

This ends the *training stage*. Observe that the missing data were not used to estimate the parameters θ_{train} and η_{train}.

5.4.2 The testing stage comprises five steps

(7) Generate testing data x_{test} from the Ising model.
(8) Split the data into an observed part $x_{\text{test}}^{(O)}$ and a 0%, 50% or 90% missing part $x_{\text{test}}^{(M)}$. The missing part will only be used to evaluate predictions.
(9) Use the Gibbs sampler to estimate the latent variables η_{test} using the *observed* testing data $x_{\text{test}}^{(O)}$ and the IRT parameters θ_{train} obtained from the training stage, e.g., step (4).
(10) Generate new predictions from the IRT model on the observed part,

$$x_*^{(O)} \sim \mathbb{P}(x \mid \eta_{\text{test}},\theta_{\text{train}}),$$

and compute $c_{\text{test}}^{(O)} = c(x_{\text{test}}^{(O)},\mathrm{x}_*^{(\mathrm{O})})$.
(11) Generate new predictions from the IRT model on the missing part,

$$x_*^{(M)} \sim \mathbb{P}(x \mid \eta_{\text{test}},\theta_{\text{train}}),$$

and compute $c_{\text{test}}^{(M)} = c(x_{\text{test}}^{(M)},\mathrm{x}_*^{(\mathrm{M})})$.

This ends the *testing stage*. Observe that *testing* data were not used to estimate the IRT parameters θ_{train} and that only the latent variables η_{test} were estimated on the *observed* part from the *testing* data.

5.5 Illustration I: Collinearity

Table 1 reveals the prediction accuracy of a two-dimensional IRT model applied to data generated from a $\tau = 1.0$ network, and data generated from a $\tau = 0.5$ network. Evidently, the two-dimensional IRT model provides accurate predictions that cross-validate well. As expected, the prediction accuracy is an increasing function of the observed sample correlations (c.f. Fig. 3). Furthermore, Table 1 shows that the approach is largely insensitive to variations in n and p, ensuring that the procedure scales when more observations n and/or more variables p become available. This scalability is important for situations where the number of observations n becomes too large, and one has to use a selection of the available observations, with the estimated IRT model cross-validating well in such applications.

The prediction accuracy of the IRT model is similar to the prediction accuracy of the true model. This indicates a good fit of the IRT model to the network data, which is somewhat surprising since we expect a bimodal or mixture distribution for the latent variables in a network of such correlated variables x_i. This is indeed the case. Figure 4 shows the marginal distribution of the latent variables in the first dimension, i.e., the marginal distributions of plausible values. It is clear that the marginal distribution of plausible values is bimodal and diverges from the unimodal population distribution that we have used. The fact that our predictions are still on par irrespective of the fit of the latent variable distribution $f(\eta)$ shows that the prediction procedure is robust against misspecification of the latent variable model.

5.6 Illustration II: Ignorable missing data

Table 2 reveals the prediction accuracy of a two-dimensional IRT model applied to data generated from a $\tau = 1.0$ network, with either 50 or 90% of the data missing completely at random (MCAR; see Appendix A). Similarly, in Table 3 we report the prediction accuracy of a two-dimensional IRT model applied to data generated from a $\tau = 0.5$ network, with either 50 or 90% of the data MCAR. We report both the accuracy in predicting the observed data $c^{(O)}_{\text{train}} = c(x^{(O)}_{train}, \mathrm{x}^{(\mathrm{O})}_{*})$, and the accuracy in predicting the missing data $c^{(M)}_{\text{train}} = c(x^{(M)}_{train}, \mathrm{x}^{(\mathrm{M})}_{*})$. Since the true model cannot be used with missing observations, we evaluate the predictions from the true model using the completely observed test data: $c_{\text{true}} = c(x_{\text{train}}, \mathrm{x}_{*})$.

The results that are reported in Tables 2 and 3 indicate that the IRT model efficiently operates when large portions of the data are MCAR. This is particularly evident when we compare the results in Tables 2 and 3 that are based on incomplete data with the results in Table 1 that are based on complete data. The prediction accuracy of the IRT model also compares favorably to the accuracy that is obtained from the true model that is also based on the complete data. Note also that the IRT

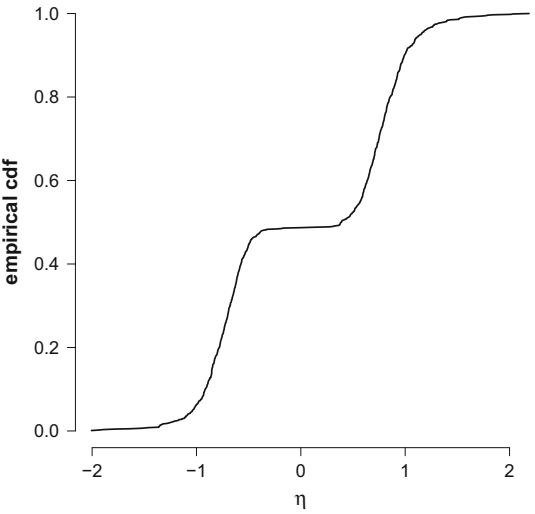

Fig. 4 The marginal distribution of the $n = 1000$ plausible values based on $p = 100$ variables generated from a $\tau = 0.5$ network

Table 3 Prediction accuracy for the two-dimensional IRT model applied to $\tau = 0.5$ network data with ignorable missing observations

n	p	c_{true}	$c^{(o)}_{\text{train}}$	$c^{(m)}_{\text{train}}$	$c^{(o)}_{\text{test}}$	$c^{(m)}_{\text{test}}$
50% missing observations						
100	1000	91	91	90	90	90
100	100	92	92	91	91	91
1000	100	92	92	91	92	91
1000	1000	92	91	90	90	90
90% missing observations						
100	1000	91	91	85	85	84
100	100	92	93	83	89	84
1000	100	91	90	86	90	86
1000	1000	91	91	90	91	90

model is entirely capable of making accurate predictions about the missing data, even with only 10% observations left to train the model.

5.6.1 Logistic regression

It is instructive to compare the MIRT model's performance with that of logistic regression. The ideal situation for this comparison would have more observations than variables ($n > p$) so that there is no need for regularization in estimating the logistic regression model. We, therefore, use $n = 1000$ observations on $p = 100$ variables.

The results that are reported in Tables 2 and 3 are based on data with missing observations. To allow a meaningful comparison between logistic regression and MIRT when some observations are missing, we use multiple imputation to complete the datasets. Unfortunately, we now encounter a serious complication. To impute

the missing observations, we need to formulate an *a priori* distribution for the missing observations (c.f. Ibrahim et al. 2005). The most straightforward solution is to specify a joint distribution for a $p+1$ dimensional vector of binary variables x. That is, we assume that the variables are dependent on each other and inform about each other's missing values. Even though this is a straightforward strategy, it will boil down to an *a priori* distribution for the missing observations that is at least as complex as the computationally intractable Ising model.

To overcome this complication, we assume that the variables are *a priori* independent. Specifically, we assume for each missing observation that the prior probability that its value is $+1$ is equal to some number π. We use the value $\pi = 0.5$ since we have no *a priori* preference for a particular value of the missing observation. Denote the logistic regression model as

$$\mathbb{P}(y \mid x) = \frac{\exp\left(y\left[\alpha + x^{\mathsf{T}}\beta\right]\right)}{\sum_y \exp(y\left[\alpha + x^{\mathsf{T}}\beta\right])},$$

with $y \in \{-1, +1\}$ the dependent variable and $x \in \{-1, +1\}^{p-1}$ a vector of covariates. The posterior distribution for a missing observation x_i is easily computed:

$$\mathbb{P}(x_i \mid y, x_{\setminus i}) = \frac{\mathbb{P}(y \mid x_{\setminus i}, x_i)\mathbb{P}(x_i)}{\sum_{x_i} \mathbb{P}(y \mid x_{\setminus i}, x_i)\mathbb{P}(x_i)}.$$

Observe that the distribution $\mathbb{P}(x_i \mid y, x_{\setminus i})$ favors values of x_i that minimize $|(2y-1) - \mathbb{E}(y)| = |(2y-1) - \mathbb{P}(y \mid x)|$.

We use the Gibbs sampler to estimate the logistic regression model and use logistic prior distributions with location 0 and scale 1 for the model's parameters α and β. Our imputation strategy expands the Gibbs sequence by two distinct steps. In the first step, missing values for the dependent variable are drawn from the predictive distribution $\mathbb{P}(y \mid x)$ (i.e., the logistic regression model). In the second step, missing values for each of the p covariates are drawn from their respective posterior distributions $\mathbb{P}(x_i \mid y, x_{\setminus i})$. After these two steps the data are complete and we can simulate the model's parameters α and β from their full-conditional posterior distributions as if all data had been observed.

We generate 300 datasets from the $\tau = 1.0$ network and 300 datasets from the $\tau = 0.5$ network and randomly remove 0, 50 or 90 of the observations. The prediction accuracy for logistic regression applied to these datasets are reported in Table 4 and reveals three important results. The first result is that the MIRT model performs better on the completely observed data than the logistic regression model. This is likely due to collinearity, as the relative performance of logistic regression deteriorates with increasing correlations. For instance, when compared to the true model's prediction accuracy we observe an 8% accuracy drop for the $\tau = 1.0$ network data and a 15% accuracy drop for the $\tau = 0.5$ network data.

The second important result is that the logistic regression model's prediction accuracy on observed data is much improved when missing observations are introduced. In fact, logistic regression outperforms both the MIRT model and the

Table 4 Prediction accuracy for the logistic regression model applied to $n = 1000$ observations of $p = 100$ correlated variables (one dependent and $p - 1$ covariates)

	$\tau = 1.0$ network data			$\tau = 0.5$ network data		
	Observed	Missing	%	Observed	Missing	%
c_{true}	80	–	0	91	–	0
c_{train}	72	–	0	76	–	0
c_{test}	72	–	0	76	–	0
c_{train}	92	59	50	92	55	50
c_{test}	94	59	50	93	55	50
c_{train}	97	62	90	97	65	90
c_{test}	99	61	90	99	64	90

The missing observations are missing completely at random

true generating model on the remaining—observed—data. (This improvement was only seen in the dependent variable, not the covariates.) This striking increase in accuracy is due to the way that we impute the missing values. The imputation distribution $\mathbb{P}(x_i \mid y, x_{\setminus i})$ tends to favor values that make the observed outcomes y more likely and minimize $|(2y - 1) - \mathbb{P}(y \mid x)|$. As a result, prediction accuracy increases when more observations are missing and the model over-fits the remaining observed data.

The final important result that we observe from Table 4 is the poor prediction accuracy on the missing observations. Compared to the MIRT model the accuracy of predicting missing values drops approximately 25–35%. This striking difference between the accuracy on the missing data and on the observed data is a clear illustration of the poor cross-validation that follows from over-fitting on the observed data points. This is particularly problematic when one aims to predict non-observed data points, e.g., classification of future preferences: whereas one believes to be doing quite a good job based on predictions of the observed data, one unknowingly is doing a very poor job in predicting non-observed data.

5.7 Illustration III: Nonignorable missing data

From the results that are reported in Tables 2 and 3 we have learned that the two-dimensional IRT model provides accurate predictions when applied to correlated data where some of the observations are MCAR. Since there is no additional difficulty for the case when the data are MAR instead of MCAR, we consider here the situation where the IRT model is applied to data where the missing data mechanism is nonignorable, i.e., not missing at random (NMAR; see Appendix A). We compare situations where either the training data, the testing data, or both the training data and the testing data have 50% data NMAR or 50% data MCAR.

We use several mechanisms to produce nonignorable missing data patterns, with the missing data mechanism explicitly depending on the missing observations and/or the parameters of the observed data model. In Appendix B we describe two procedures, one of which produces missing data patterns that have a moderate effect on observed correlations and parameter estimates—moderate nonignorability—and one which produces missing data patterns that have a severe effect on observed

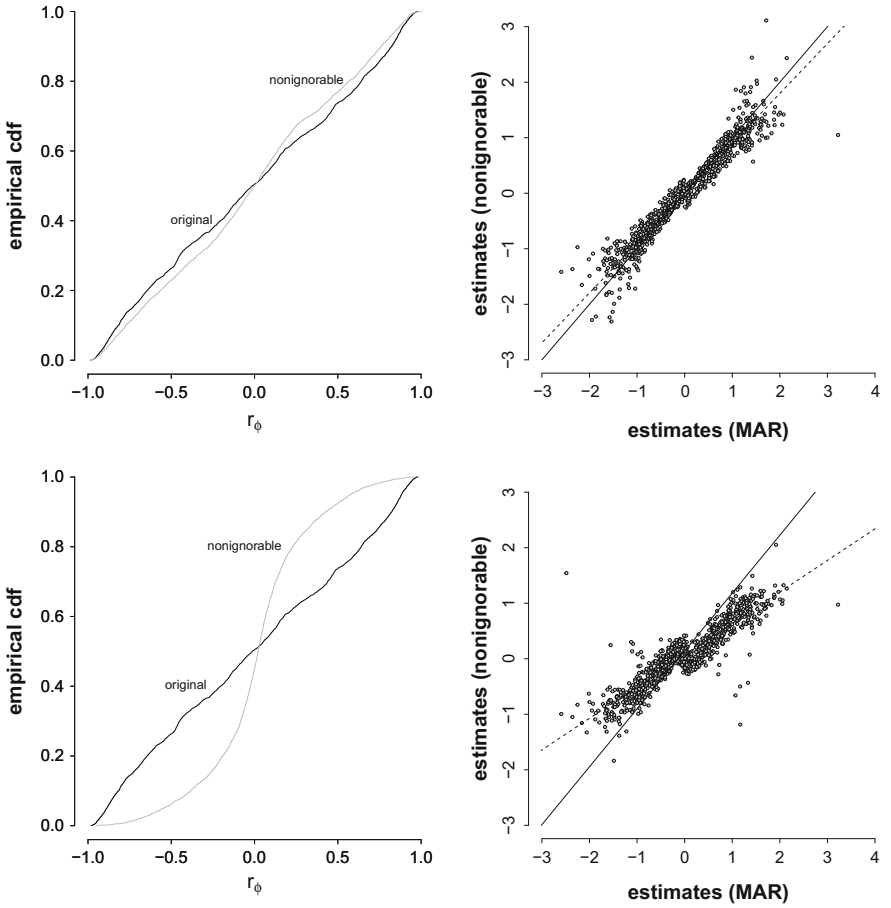

Fig. 5 The result of nonignorable missing data on observed correlations and parameter estimates. The *two panels* on the *left* show the distribution of the observed correlations from data generated from a $\tau = 1.0$ network (*black line*), together with the observed correlations based on data subject to moderate nonignorability (*top-left panel*; *gray line*) and data subject to severe nonignorability (*bottom-left panel*; *gray line*). The *two panels* on the *right* show the estimates of u_1, the first column of U, based on data that are MCAR against the estimates based on data subject to moderate nonignorability (*top-right panel*) and data subject to severe nonignorability (*bottom-right panel*)

correlations and parameter estimates—severe nonignorability. The effect of the two described missing data mechanisms on the observed correlations is shown in Fig. 5a for moderate nonignorability, and Fig. 5c for severe nonignorability. In Fig. 5b, d we plot the corresponding estimates of u_1, the first column of U, for the two situations described by Fig. 5a, c, respectively, against the estimates that are obtained from data where the missing observations are MCAR. Clearly, ignoring the missing data mechanism produces bias to the estimates of u_1, especially for the severe nonignorability case.

Table 5 Prediction accuracy of the two-dimensional IRT model applied to $\tau = 1.0$ network data with moderate nonignorable missingness

x_{train}	x_{test}	n	p	c_{true}	$c^{(o)}_{\text{train}}$	$c^{(m)}_{\text{train}}$	$c^{(o)}_{\text{test}}$	$c^{(m)}_{\text{test}}$
NMAR	NMAR	100	1000	81	81	80	77	79
NMAR	NMAR	100	100	81	81	79	77	78
NMAR	NMAR	1000	100	81	79	79	78	79
NMAR	NMAR	1000	1000	82	79	80	79	80
NMAR	MCAR	100	1000	81	80	79	77	78
NMAR	MCAR	100	100	80	80	77	75	76
NMAR	MCAR	1000	100	80	77	77	76	77
NMAR	MCAR	1000	1000	82	80	80	79	80
MCAR	NMAR	100	1000	81	82	81	80	80
MCAR	NMAR	100	100	80	81	78	78	78
MCAR	NMAR	1000	100	81	80	80	80	80
MCAR	NMAR	1000	1000	81	81	81	81	81

Table 5 reveals the prediction accuracy of the two-dimensional IRT model applied to data generated from a $\tau = 1.0$ network, with three distinct situations in which either the training data, the test data, or both data sets have missing data patterns that have a moderate effect on observed correlations,i.e., moderate nonignorability. When we compare our predictions with the predictions that are made with the true model $\mathbb{P}(x_i \mid x_{\setminus i})$ based on completely observed data, we see that the prediction accuracy drops approximately: 3–4% when some of the training data are NMAR and some of the testing data are NMAR, 4% when some training data are NMAR but testing data only have data MCAR, and 1% when the training data has data MCAR but the testing data have data NMAR. The 1% drop in accuracy that is found when the training data have some observations MCAR is within the range found for the complete data in Table 1 and data with observations MCAR in Table 2. This suggests that training the prediction model on data for which the missing data are missing at random does not threaten prediction accuracy. The 3–4% drop in accuracy that is found when training the model on data with nonignorable missing data patterns is slightly higher than the range found for the complete data and data with observations MCAR. Predictions about the missing testing data are, on average, more accurate than that of the observed testing data, whereas the opposite is found for the training data.

Table 6 reveals the prediction accuracy in the same situations as in Table 5, but where the nonignorable missing data mechanism has a severe effect on the observed correlations. When compared to predictions that are made by the true model $\mathbb{P}(x_i \mid x_{\setminus i})$ based on complete data, we observe that the prediction accuracy drops approximately: 7–8% when some of the training data are NMAR and some of the testing data are NMAR, 8–9% when some training data are NMAR but testing data only have data MCAR, and 1% when the training data have data MCAR but the testing data have data NMAR. Thus, in comparison to the results reported in Table 5 for the missing data having a moderate effect on observed correlations, we have that the drop in prediction accuracy is roughly doubled when training the model on data

Table 6 Prediction accuracy of the two-dimensional IRT model applied to $\tau = 1.0$ network data with severe nonignorable missingness

x_{train}	x_{test}	n	p	c_{true}	$c^{(o)}_{\text{train}}$	$c^{(m)}_{\text{train}}$	$c^{(o)}_{\text{test}}$	$c^{(m)}_{\text{test}}$
NMAR	NMAR	100	1000	81	82	65	72	73
NMAR	NMAR	100	100	82	83	61	73	71
NMAR	NMAR	1000	100	82	81	63	75	75
NMAR	NMAR	1000	1000	81	81	64	76	75
NMAR	MCAR	100	1000	81	82	64	72	73
NMAR	MCAR	100	100	81	82	59	71	70
NMAR	MCAR	1000	100	79	79	56	71	69
NMAR	MCAR	1000	1000	81	81	64	75	74
MCAR	NMAR	100	1000	81	82	81	80	80
MCAR	NMAR	100	100	81	83	79	79	79
MCAR	NMAR	1000	100	80	79	79	79	78
MCAR	NMAR	1000	1000	81	81	81	81	81

subject to nonignorable missing data patterns. When the training data have data NMAR, we see that predicting the missing testing data is about as difficult as predicting the observed testing data. However, predicting the missing training data is clearly more difficult than predicting the observed training data.

6 Discussion

We have illustrated that the combination of the Ising model and its latent variable approximation can be used to overcome collinearity and missing data issues in prediction applications of the logistic regression model. The prediction accuracy of the latent variable model on the observed data compares favorably to the true model used in Illustrations I–III (e.g., Fig. 2). The latent variable model was also able to accurately predict the non-observed data points in Illustration II (e.g., Tables 2, 3) and compares favorably to our illustration of the logistic regression model (e.g., Table 4). The model has its limits, which was clearly demonstrated in Illustration III using nonignorable missing data mechanisms. The prediction accuracy deteriorates when missing data mechanisms have a strong effect on the data that is used to train the model (e.g., Tables 5, 6). However, even with the poor quality of the data that were used to train the model in Illustration III, the prediction accuracy of the latent variable model compares favorably to our application of the logistic regression model (e.g., compare Tables 4 and 6). Therefore, we believe that our approach and the associated latent variable model are superior to the logistic regression model in prediction settings with correlated covariates and/or missing observations.

We have considered a specific prediction setting with only binary random variables for our approach to overcome collinearity and missing data. Observe, however, that the two primary ideas that form our approach are entirely general. The first idea is to model the joint distribution of dependent and independent variables when the variables are correlated, e.g., Ising networks models, Gaussian graphical models (Lauritzen 1996), or mixtures thereof (Olkin and Tate 1961; Lauritzen and

Wermuth 1989). A specific feature of these particular models is that they specifically model the correlations between variables and thus overcome collinearity issues in conditional regression models. However, the application of these models will also increase the number of parameters that need to be estimated. Our second idea is to approximate the full graphical model with a low-rank latent variable model, e.g., an IRT model, a factor model or mixtures thereof. This has two important benefits: it reduces the number of parameters that need to be estimated and introduces an elegant way of handling missing data. We believe that these two general ideas will inspire new avenues of future research and furthermore offer practical solutions to issues that are widespread in large-scale applications.

The latent variables η in this paper are used to summarize the observed data in order to make predictions. However, the latent variable, in combination with the model's parameters $\theta = \{U,\mu\}$, can also be used to inform about the structure of the prediction problem. For instance, in regular applications of IRT models to educational tests (e.g., an end of primary school test), the model informs about the dimensionality of the test (e.g., separates a mathematics and language dimension) and informs how different aspects of the test covary across the ability spectrum. In much the same way we may study the latent variable model in prediction settings, which might improve the model and its predictions. For example, in psychological data we nearly always observe a pattern of positive correlations (in contrast to Fig. 3), which means that the entries in the first eigenvector tend to have the same sign. Without much loss of accuracy we can then replace the p unknown values in u_1 with a single (positive) number. This would significantly reduce the number of parameters that we need to estimate, and we obtain sparse models that can make accurate predictions.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A: Types of missing data

In many applications, some of the observations are missing (denoted x_{mis}) and models have to be trained on an incomplete dataset (denoted x_{obs}). The missing observations are referred to as missing at random (MAR) whenever the missing data indicator M does not depend on the missing observations:

$$\mathbb{P}_{\text{MAR}}(M = m \mid x_{\text{obs}}, x_{\text{mis}}, \phi) = \mathbb{P}(M = m \mid x_{\text{obs}}, \phi),$$

but may depend on the observed datapoints x_{obs}. The missing data are referred to as missing completely at random (MCAR) whenever the missing data mechanism does not depend on any data, missing or observed:

$$\mathbb{P}_{\text{MCAR}}(M = m \mid x_{\text{obs}}, x_{\text{mis}}, \phi) = \mathbb{P}(M = m \mid \phi).$$

Whenever the parameters ϕ that govern the missing data mechanism are distinct from the parameters that govern the observed data (here U, μ and η), and the missing data mechanism is MAR (or MCAR), the missing data mechanism can be safely ignored without introducing bias to the estimates of U, μ and η (Rubin 1976; Little and Rubin 1987; Heitjan 1994). In a Bayesian framework distinctness refers to the parameters ϕ and U, μ, and η being *a priori* independent:

$$f(\phi, U, \mu,\eta) = \mathrm{f}(\phi)\mathrm{f}(\mathrm{U}, \mu,\eta).$$

Whenever either of these two conditions (MAR and distinctness) are not satisfied, the missing data are nonignorable and failure to correctly model the missing data mechanism can bias the estimates of the observed data parameters (Rubin 1976; Little and Rubin 1987; Heitjan 1994).

Appendix B: Generate nonignorable missing data

Moderate nonignorability

Moderately nonignorable missing data were created as follows: First, for each row v, $v = 1, \ldots, n$, of the generated dataset x, a latent variable $\gamma_v = x_v^{\mathsf{T}} u_1 + \epsilon$ was created, where $\epsilon \sim \mathcal{N}(0, 5)$ and where $x_v^{\mathsf{T}} U_1$ is the posterior expectation of η_{1v}; the correlation between the γ_v and η_{1v} was approximately 0.6. We then cycle through the consecutive column-pairs of the generated data matrix (i.e., (1, 2), (2, 3),...,$(p-1, p)$, $(p, 1)$) and take one of three actions (with equal probability):

1. If the correlation of observations in columns i and j, with $(i,j) \in \{(1,2), (2, 3),\ldots,(p-1,p), (p,1)\}$, was positive, we removed the observations in column i, rows v, with probability $(1 + e^{-x_{vi}x_{vj}})^{-1}$, which is approximately 0.7 for observations $x_{vi} = x_{vj} = \pm 1$. If the correlation was negative we removed observations with probability $(1 + e^{x_{vi}x_{vj}})^{-1}$ to the same effect.
2. We randomly generated a value $y \in \{0,1,2\}$ and removed observations in column i, rows v, with probability

$$P(M_{vi} = 1) = \begin{cases} (1 + e^{-\gamma_v})^{-1} & \text{if } y = 0 \\ (1 + e^{\gamma_v})^{-1} & \text{if } y = 1 \\ (1 + e^{|\gamma_v|}\})^{-1} & \text{if } y = 2 \end{cases}$$

   favoring exclusion for large positive, large negative and moderate values of γ_v, respectively.
3. Randomly remove 50% of the observations in column i.

The above procedure created approximately 35–40% missing observations, and observations were either randomly omitted or placed back to make it an exact 50%.

Severe nonignorability

Severely nonignorable missing data were created as follows: First, we cycled through the consecutive column-pairs of the generated matrix (i.e., (1, 2), (2, 3),...,$(p-1,p)$, $(p, 1)$). When the correlation between columns i and j was positive we either removed all (1, 1) or $(-1,-1)$ observations, and when the correlation was negative we removed either all $(1,-1)$ or $(-1, 1)$ observations. We then cycled through the n rows of the generated matrix. For each row, we randomly generated a value $y \in \{0,1,2,3\}$. We removed 50% of the $+1$ observations when y equaled 0, 50% of the -1 observations when y equaled 1, removed 25% of the $+1$ and 25% of the -1 responses if y equaled 2, and did nothing when y equaled 3. This created approximately 50% missing observations, and observations were either randomly omitted or placed back to make it an exact 50%.

References

Ackerman T (1996) Developments in multidimensional item response theory. Appl Psychol Measure 20:309–310

Ackerman T, Gierl M, Walker C (2003) Using multidimensional item response theory to evaluate educational and psychological tests. Educ Measure Issues Pract 22:37–51

Anderson C, Vermunt J (2000) Log-multiplicative association models as latent variable models for nominal and/or ordinal data. Sociol Methodol 30:81–121

Bartlett P, Jordan M, McAuliffe J (2006) Convexity, classification, and risk bounds. J Am Stat Assoc 101:138–156

Bell R, Koren Y (2007) Lessons from the netflix prize challenge. ACM SIGKDD Explor Newslett 9:75–79

Bell R, Koren Y, Volinsky C (2010) All together now: a perspective on the netflix prize. Chance 23:24–29

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B (Methodol) 36:192–236

Borsboom D, Molenaar D (2015) Psychometrics. In: Wright J (ed) International Encyclopedia of the Social and Behavioral Sciences, vol 19, 2nd edn, pp 418–422

Brin S, Page L (2012) Reprint of: the anatomy of a large-scale hypertextual web search engine. Comput Netw 56:3825–3833

Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1:211–218

Eggen TJHM (2004) Contributions to the theory and practice of computerized adaptive testing. Ph.D. thesis, University of Twente, Enschede, The Netherlands

Emch G, Knops H (1970) Pure thermodynamical phases as extremal KMS states. J Math Phys 11:3008–3018

Epskamp S, Maris G, Waldorp L, Borsboom D (2017) Network psychometrics. In: Irwing P, Hughes D, Booth T (eds) Handbook of psychometrics. Wiley, New York (**in press**)

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

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Heitjan D (1994) Ignorability in general incomplete-data models. Biometrika 81:701–708

Ibrahim J, Chen M, Lipsitz S, Herring A (2005) Missing-data methods for generalized linear models: a comparative review. J Am Stat Assoc 100:332–346

Ising E (1925) Beitrag zur theorie des ferromagnetismus. Zeitschrift für Physik 31:253–258
Kac M (1968) Mathematical mechanisms of phase transitions. In: Chretien M, Gross E, Deser S (eds) Statistical physics: phase transitions and superfluidity, vol 1. Brandeis University Summer Institute in Theoretical Physics., pp 241–305. Gordon and Breach Science Publishers, New York
Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42:30–37
Lauritzen S (1996) Graphical models. Oxford University Press
Lauritzen S, Wermuth N (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17(1):31–57
Lenz W (1920) Beiträge zum verständnis der magnetischen eigenschaften in festen körpern. Physikalische Zeitschrift 21:613–615
Little R, Rubin D (1987) Statistical analysis with missing data. Wiley, New York
Maris G, Maris E (2002) A MCMC-method for models with continuous latent responses. Psychometrika 67:335–350
Marsman M, Maris G, Bechger T, Glas C (2015) Bayesian inference for low-rank Ising networks. Sci Rep 5(9050):1–7
Marsman M, Maris G, Bechger T, Glas C (2016) What can we learn from Plausible values? Psychometrika. doi:10.1007/s11336-016-9497-x
Marsman M, Maris G, Bechger T, Glas C (2017) Turning simulation into estimation: generalized exchange algorithms for exponential family models. PLoS One 12(e0169787):1–15
McCullagh P (1994) Exponential mixtures and quadratic exponential families. Biometrika 81:721–729
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Mislevy R (1991) Randomization-based inference about latent variables from complex samples. Psychometrika 56:177–196
Olkin I, Tate R (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465
Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web
Patz R, Junker B (1999) Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. J Educ Behav Stat 24:342–366
Patz R, Junker B (1999) A straightforward approach to Markov chain Monte Carlo methods for item response models. J Educ Behav Stat 24:146–178
Reckase M (2009) Multidimensional item response theory. Springer
Rousseeuw P, van den Bossche W (2016) Detecting deviating data cells. arXiv preprint arXiv:1601.07251
Rubin D (1976) Inference and missing data. Biometrika 63:581–592
Rubin D (1987) Multiple imputation for nonresponse in surveys. Wiley, New-York
Tierney L (1994) Markov chains for exploring posterior distributions. Ann Stat 22:1701–1762
von Davier M, Gonzalez E, Mislevy R (2009) What are plausible values and why are they useful? In: von Davier M, Hastedt D (eds) IERI monograph series: issues and methodologies in large scale assessments, vol 2. IEA-ETS Research Institute