Improve diagnosis with DINA model: the power of denoising

Peng Xu, Michel Desmarais

Abstract

Latent variable models are commonly used to model student performance in educational data mining. It is generally difficult to compare different models since there lacks of ground truth for student profile. In this paper, we use the prediction competence as the criterion to evaluate model performance. Such way is also the most commonly used approach in machine learning community. FA(Factor Analysis), PCA(Principle Component Analysis), IRT(Item Response Theory) and MM(Mixture of Models) are compared in the research.

1 Introduction

In educational diagnosis task, we generally assume that the data we observe is noised. In a typical setting like an school exam, the students are asked to answer a series of questions and we are supposed to conduct diagnosis of student mastery of skills based on the response data. The questions are usually called items in the community and an item-skill mapping matrix called Q-matrix are usually given to help diagnose the mastery of each skill. There are two types of errors in this setting. First is the experiment or observation error. For example, students might slip an item due to some negligence, or guess an item correct even without mastery the required skills. Second is the model error. For example the pre-given Q-matrix only assumes 3 skills are involved, but in fact there should be 4 skills, then the effects of the 4-th skill will be incorporated in the form of noise.

It is natural to think that use some denoising techniques as a preprocessing step would improve the result of some planned analysis. For example, kalman filtering in signal processing used this idea. It considers a series of measurements over time,

Denoising is a common technique used in signal processing and image processing. In signal processing, Educational measurement is a typical task in educational data mining. Generally, it tries to infer the latent abilities

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
r_1	0	0	0	0	0	1	0	0	0
r_2	1	0	0	1	0	1	0	0	0
r_3	1	1	1	1	1	1	1	1	1
r_4	0	1	0	0	1	1	0	1	1
...	...	...	...	...	...	...	...	...	...

Figure 1: Response Matrix

of students, such as their mastery of mathematical skills or verbal skills, from a bunch of test results. This is a type of knowledge discovery, and from a machine learning perspective, it is a standard task of unsupervised learning, since we do not have a labelled dataset for us to train. That is, to predict one item using the values of other items. Latent variable Models(LVM)(Bartholomew, Knott, & Moustaki, 2011) are commonly used in this field thus in this paper, we conduct a simple comparison among different latent variable models. More precisely, all these latent variable models are discussed in a probabilistic framework.

In the previous work from (Desmarais, 2011), the author compared several models with IRT in a classification task. In this paper, the task is a bit different, and the detail is in section 4.

2 Latent Variable Models

In educational data mining, the mostly seen datasets are respondent-item datasets. A typical respondent-item dataset looks like below,

Usually these items are correlated, but we can suppose that they are caused by some common latent variables, thus inducing conditional independence. For example, in the matrix of Figure 1, if we suppose there are 2 latent variables for the 9 items, a directed graph (Figure **??**) could be drawn to show the conditional independence.

Therefore we expressed the latent variable models as graph models. By applying different restrictions to the observed variables V and hidden variables H, we will obtain different latent variable models, as discussed below.

2.1 FA model

First, let us consider the case that both the latent variables and observed variables are continuous. We set the prior for the latent variable to be

Gaussian. That is,

$$p(h_i) = \mathcal{N}(h_i|\mu_0, \Sigma_0)$$

Let the likelihood to be Gaussian too, and the mean of observed variable to be a linear combination of latent variables, then we have the likelihood as below

$$p(v_i|h_i, \theta) = \mathcal{N}(Wh_i + \mu, \Psi)$$

When Ψ is diagonal, this model is the classical Factor Analysis model. W is the loading matrix, Ψ are the random errors for each observed sample(Murphy, 2012). In history, FA was first proposed to tackle the problem of cognitive diagnosis in the field of psychometrics, serving for the task exactly the same as in this paper.

2.2 PCA model

In the FA model, If we constrain $\Psi = \sigma^2 I$ and W to be orthogonal, then as $\sigma^2 \to 0$, this model reduces to principal components analysis(PCA)(Murphy, 2012). Besides this probabilistic view, PCA is more commonly seen as a deterministic method and a frequently used tool for dimension reduction. It has a direct link to SVD technique and making the parameter computation easier.

2.3 IRT model

If we require the observed variable to be binary, with hidden variable to maintain continuous, and using logit function as the link function, then we have

$$p(v_i|h_i, \theta) = \prod_{r=1}^{R} Ber(v_{ir}|sigm(W_r^T h_i + W_{0r}))$$

This model carries the name of Categorical PCA(Murphy, 2012) in machine learning community and is also famous in psychometrics with the name item response theory(IRT).

2.4 MM model

A simpler way to use the latent variable model is to consider there is only one latent variable, which denotes the category of respondents. That is to say, all the respondents are categorised into several classes. For each class, it corresponds to a model. In our case, we suppose that for each class, it is a Bernoulli model. And for simplicity, a hypothesis is made that all the

conditional probabilities are independent. That is, for a latent category k, we have

$$P(v_i|z_i = k) = \prod_{j=1}^{d} P(v_{ij}|z_i = k)$$

$$P(v_{ij}|z_i = k) = Ber(v_{ij}|\theta_{kj})$$

3 Datasets

All datasets are from R package 'CDM'(Robitzsch, Kiefer, George, & Ünlü, 2017).

sim.dina: Artificial Dataset.

fraction1: A fraction subtraction data set with 536 students and 15 items. The Q-matrix was defined in (De La Torre, 2009).

fraction2: Another fraction subtraction data set with 536 students and 11 items. The Q-matrix was defined in (De La Torre, 2009).

subtraction: (Tatsuoka, 1984) fraction subtraction data set, is comprised of responses to 20 fraction subtraction test items from 536 middle school students.

ecpe: From (Templin & Hoffman, 2013) tutorial of specifying cognitive diagnostic models in Mplus.

Name	# Observations	# Items
sim.dina	400	9
ecpe	2922	28
fraction1	536	15
fraction2	536	11
subtraction	536	20

Table 1: Datasets

4 Methodology

To determine the scores or states of latent variables is an typical unsupervised task, since we can never know the real value of them. Therefore, in order to compare the performance among different models, we instead scrutinize their predictability. The procedure we use is a 10-folds cross validation. However, we do not divide datasets by students, but by cells. That

is, the dataset fed to the model looks like below, this is to guarantee that our models can learn parameters for every student and item:

$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ ... \end{array} \begin{array}{c} \begin{array}{ccccccccc} i_1 & i_2 & i_3 & i_4 & i_5 & i_6 & i_7 & i_8 & i_9 \end{array} \\ \left[\begin{array}{ccccccccc} 0 & 0 & 0 & 0 & X & 1 & 0 & 0 & 0 \\ 1 & X & 0 & 1 & 0 & 1 & X & 0 & 0 \\ 1 & 1 & X & 1 & X & 1 & 1 & 1 & 1 \\ X & 1 & 0 & 0 & 1 & 1 & X & 1 & X \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \end{array}\right] \end{array}$$

where X denotes the missing value and will be predicted.

4.1 Criterion

We use RMSE as the measure of performance, which is the square root of the L_2 risk, also can be seen as an estimate of the standard deviation. It is defined as below:

$$RMSE(Model) = \sqrt{E(\hat{r_{ij}} - r_{ij})^2}$$

where r_{ij} is the real value of the missing cell at i, j-th position while $\hat{r_{ij}}$ is the predicted value given by the model.

We set a benchmark for all the models which is to use mean value as the prediction. That is, $\hat{r_{ij}} = \mu + b_u + b_i$, where μ is the global average, b_u is the student bias and b_i is the item bias.

4.2 Determine the number of latent variables

To determine the number of latent variables used in a model is a difficult task. Generally speaking, there is a trade-off between increasing the model complexity and increasing the model reliability(DiBello, Roussos, & Stout, 2006). This is a version of the famous bias-variance trade-off(Friedman, Hastie, & Tibshirani, 2001). In machine learning community, when dataset is small, cross-validation is commonly used to determine the number. However, it is very computationally expensive for large datasets.

In fact, from a Bayesian view, we should pick the model with the largest marginal likelihood, $K^* = argmax_k p(D|K)$, where K is the latent number and D is the dataset(Murphy, 2012). However, this likelihood is also difficult to calculate. Simple approximations, such as BIC can be used(Fraley & Raftery, 2002).The idea for BIC is to add a penalty term to log likelihood. It carries the form below:

$$BIC = k * npar - 2log(L)$$

where $npar$ is the number of parameters,$log(L)$ is the log-likelihood, and k is a coefficient to decide the degree of penalty. If $k = 2$, it is the AIC. if $k = ln(N)$, it is the BIC.

For FA and PCA, we have another option, i.e. the Horn's Parallel Analysis (PA) (Horn, 1965)for determining, which is the most recommended method(Hayton, Allen, & Scarpello, 2004).

We show the model selection by cross-validation for all models.

4.3 Implementation

All experiments were conducted in R. The core functions used are listed below:

FA: function 'factanal'
PCA: function 'svd'
IRT: function 'mirt' from 'mirt' package
MM: function written by the author

5 Results

We show the result of precsion, recall and F-score for different denoising methods in table 2.

Denoising Methods	Precision	Recall	F-score
Non-denoised	0.84	0.84	0.84
IRP	0.90	0.81	0.85
knn	0.83(-)	0.82(-)	0.82
rasch	0.66	0.69	0.67
DINA	0.85(-)	0.84(-)	0.84(-)

Table 2: Results of N=100, k=3, l=10, slip=guess=0.2

From the RMSE results, we can see that MM outperforms all other models.

We also have the results for model selection by cross validation in table 4.

For illustration purpose, we show the relation between RMSE and latent number k for models on subtraction dataset.

Denoising Methods	Precision	Recall	F-score
Non-denoised	0.84	0.84	0.84
IRP	0.91	0.82(-)	0.86
knn	0.83(-)	0.81	0.82
rasch	0.66	0.71	0.68
DINA	0.85(-)	0.84(-)	0.84(-)

Table 3: Results of N=1000, k=3, l=10, slip=guess=0.2

Datasets	FA	PCA	IRT	MM
sim.dina	5	1	1	5
fraction1	1	3	3	6
fraction2	3	4	1	6
subtraction	1	5	1	8
ecpe	1	1	-	-

Table 4: Results of Model selection by cross validation. For FA, PCA and IRT, the number denotes the latent dimensions. For MM, it is the number of classes represented by the single latent variable.

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

De La Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

Desmarais, M. C. (2011). Performance comparison of item-to-item skills models with the irt single latent trait model. In *International conference on user modeling, adaptation, and personalization* (pp. 75–86).

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, *26*, 979–1030.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611–631.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, *7*(2), 191–205.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2017). Cdm: Cognitive diagnosis modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=CDM` (R package version 5.4-0)

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50.