INTERPRETABLE AND ACCURATE SCALING IN LARGE-SCALE ASSESSMENT: A VARIABLE SELECTION APPROACH TO LATENT REGRESSION

Yunxiao Chen[1], Motonori Oka[1] and Matthias von Davier[2]

[1] Department of Statistics, London School of Economics and Political Science, (e-mail: y.chen186@lse.ac.uk, m.oka1@lse.ac.uk)

[2] Lynch School of Education and Human Development, Boston College, (e-mail: matthias.vondavier@bc.edu)

ABSTRACT: This paper concerns the construction of scaling models for large-scale assessments in education. A scaling model, which makes use of information from both responses to cognitive assessment and background survey items, produces plausible values for individual students. There are two major challenges when building a scaling model – (1) a large number of background variables and (2) many missing values in the background survey data. To tackle these challenges, we propose a variable selection approach to latent regression modelling. The proposed approach handles missing data by iterative imputation and controls variable selection error by a data-splitting procedure.

KEYWORDS: Latent regression, large-scale assessment, variable selection, missing data, imputation

1 Problem Setup

Consider data collected from N students, where data from different students are independent. For each student i, the data can be divided into two parts – (1) responses to cognitive items and (2) non-cognitive predictors. We use a random vector $\mathbf{Y}_i$ to denote student i's cognitive responses. Due to the matrix sampling design for cognitive items in international large-scale assessments (ILSAs), the length of $\mathbf{Y}_i$ can vary across students. More precisely, we use $\mathcal{B}_i$ to denote the set of cognitive items that student i is assigned. Then $\mathbf{Y}_i = \{Y_{ij} : j \in \mathcal{B}_i\}$. For simplicity, we assume all the items are binary, i.e., $Y_{ij} \in \{0,1\}$. In addition, consider p predictors collected via non-cognitive survey questions. Let $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{ip})^\top$ denote the complete predictor vector for student i. Often, there are missing values in $\mathbf{Z}_i$. Let $\mathcal{A}_i$ denote the set of observed predictors for student i, and let $\mathbf{Z}_i^{\text{obs}} = \{Z_{ij} : j \in \mathcal{A}_i\}$ and $\mathbf{Z}_i^{\text{mis}} = \{Z_{ij} : j \notin \mathcal{A}_i\}$. The predictors are of mixed types. Here, binary, categorical (ordinal/nominal), and continuous predictors are considered. Note that an ordinal variable will be treated as a nominal one here for simplicity. In what follows, we introduce a latent regression model, which can be decomposed into (1) a measurement model, (2) a structural model and (3) a predictor model.

Measurement model. We introduce a latent variable θ_i as the latent construct, which is measured by the cognitive items. The measurement model is an IRT model that specifies the conditional distribution of $\mathbf{Y}_i$ given θ_i. More specifically, this model assumes local independence, an assumption that is commonly adopted in IRT models (Embretson & Reise, 2000). That is, Y_{ij}, $j \in \mathcal{B}_i$, are conditionally independent given θ_i. For a dichotomous item j, the conditional distribution of Y_{ij} given θ_i is assumed to follow a two-parameter logistic model (2PL, Birnbaum, 1968). That is,

$$\mathbb{P}(Y_{ij} = 1|\theta_i) = \frac{\exp(a_j\theta_i + b_j)}{1+\exp(a_j\theta_i + b_j)}, \tag{1}$$

where a_j and b_j are two item-specific parameters. We assume that all the item parameters are known – they are fixed to be the pre-calibrated values.

Structural model. The structural model regresses the latent construct θ_i onto the complete-data predictors Z_{i1}, ..., Z_{ip}. A linear regression model is assumed for θ_i given Z_{i1}, ..., Z_{ip}. More specifically, for each variable j, we introduce a transformation $g_j(Z_j)$. When Z_j is an ordinal variable with categories $\{0,...,K_j\}$, the transformation function g_j creates K_j dummy variables, i.e., $g_j(Z_j) = (\mathbb{I}(\{Z_j = 1\}),...,\mathbb{I}(\{Z_j = K_j\}))^\top$. For continuous and binary variables, g_j is an identity link, i.e., $g_j(Z_j) = Z_j$. We assume $\theta_i|\mathbf{Z}_i \sim N(\beta_0 + \beta_1^\top g_1(Z_{i1}) + \cdots + \beta_p^\top g_p(Z_{ip}), \sigma^2)$, where β_0 is the intercept, β_1, ..., β_p are the slope parameters, and σ^2 is the residual variance. Note that β_j is a scalar when predictor j is continuous or binary and is a vector when the predictor is ordinal. Here, β_0, β_1, ..., β_p, and σ are unknown and will be estimated from the model. The main goal of our analysis is to find predictors for which $||\beta_j|| \neq 0$.

Predictor model. To handle missing values in Z_{ij}s, we impose a joint model for the predictors. Although different models may be imposed here, we assume a Second-Order Exponential (SOE) model, under which missing data imputation and parameter estimation can be carried out in a computationally efficient way. More precisely, we let $(\theta_i, \mathbf{Z}_i)$ be i.i.d., following an SOE model. Under this model, the conditional distribution of θ_i given $\mathbf{Z}_i$ is the linear regression model in the above structural model. The conditional distribution of Z_{ij} given $(\theta_i, Z_{i,-j})$ takes the following forms:

- A linear regression model (with normal residual), if variable j is continuous;
- A logistic regression model, if variable j is binary;
- A multinomial logistic regression model if variable j is categorical.

These conditional distributions will be used later for missing data imputation and parameter estimation. We remark that except for the parameters of the structural model, the rest of the parameters in the SOE can be viewed as nuisance parameters, as they are not of interest to us. The predictor model and these nuisance parameters are introduced to handle the missing values in the predictors.

2 Estimation and Variable Selection

The model introduced in the previous section implies a joint distribution of complete data, which further implies the distribution of observed data under the Missing At Random (MAR) assumption. We estimate the model and conduct variable selection based on this implied distribution for observed data. More specifically, we estimate the model parameters using an iterative imputation algorithm. According to Liu *et al.* , 2014, the estimate produced by this algorithm is asymptotically equivalent to a full Bayesian posterior-mean estimator based on the observed data likelihood. Thanks to the connection between the frequentist and Bayesian estimation provided by the Bernstein-von Mises Theorem (Van der Vaart, 2000, Chapter 10), this estimate also enjoys the desired frequentist properties, such as consistency and asymptotic normality.

Furthermore, we adopt a data-splitting method for controlled variable selection. More specifically, we combine the data-splitting method (Dai *et al.* , 2022) and the iterative imputation method to select the non-null predictors in the structural model of latent regression. Thanks to the properties of the iterative imputation method, this method has the theoretical guarantee to control the asymptotically false discovery rate for variable selection. The theoretical properties of the proposed method are confirmed by simulation results.

3 Discussions

Traditionally, a PCA-based latent regression model is used for the scaling of large-scale assessment data, in which the missing values are handled by a missing indicator approach, and the high dimensionality of the background variables and their missing indicators is reduced by Principal Component Analysis (PCA). However, this approach has three drawbacks: (1) the missing indicator approach does not perform well under certain data missingness patterns, (2) PCA may introduce spurious dependence between the achievement traits and background variables, and (3) the resulting model lacks interpretability due to the involvement of hard-to-interpret principal component scores. The proposed method does not suffer from these issues. It handles missing values more properly using iterative imputation. Furthermore, the FDR-controlled variable selection result is more interpretable and better characterises the relationship between the achievement traits and the background variables. Thus, this approach may be more suitable than the PCA-based approach in practice for scaling large-scale assessment data.

References

BIRNBAUM, ALLAN. 1968. Some latent trait models. *Pages 397–424 of:* LORD, F. M., & NOVICK, M. R. (eds), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

DAI, CHENGUANG, LIN, BUYU, XING, XIN, & LIU, JUN S. 2022. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 1–38.

EMBRETSON, SUSAN E, & REISE, STEVEN P. 2000. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

LIU, JINGCHEN, GELMAN, ANDREW, HILL, JENNIFER, SU, YU-SUNG, & KROPKO, JONATHAN. 2014. On the stationary distribution of iterative imputations. *Biometrika*, **101**, 155–173.

VAN DER VAART, AAD W. 2000. *Asymptotic statistics*. Cambridge, England: Cambridge University Press.