Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **187**, 723–747
https://doi.org/10.1093/jrsssa/qnad137
Advance access publication 12 December 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Variable selection in latent variable models via knockoffs: an application to international large-scale assessment in education

Zilong Xie[1], Yunxiao Chen[2], Matthias von Davier[3] and Haolei Weng[4]

[1]School of Mathematical Sciences, Fudan University, Shanghai, People's Republic of China
[2]Department of Statistics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK
[3]Lynch School of Education, Boston College, Chestnut Hill, MA, USA
[4]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

Address for correspondence: Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: Y.Chen186@lse.ac.uk

Abstract

International large-scale assessments (ILSAs) play an important role in educational research and policy making. They collect valuable data on education quality and performance development across many education systems, giving countries the opportunity to share techniques, organisational structures, and policies that have proven efficient and successful. To gain insights from ILSA data, we identify non-cognitive variables associated with students' academic performance. This problem has three analytical challenges: (a) academic performance is measured by cognitive items under a matrix sampling design; (b) there are many missing values in the non-cognitive variables; and (c) multiple comparisons due to a large number of non-cognitive variables. We consider an application to the Programme for International Student Assessment, aiming to identify non-cognitive variables associated with students' performance in science. We formulate it as a variable selection problem under a general latent variable model framework and further propose a knockoff method that conducts variable selection with a controlled error rate for false selections.

Keywords: international large-scale assessment, latent variables, missing data, Model-X knockoffs, variable selection

1 Introduction

International large-scale assessments (ILSAs), including the Programme for International Student Assessment (PISA), Programme for the International Assessment of Adult Competencies (PIAAC), Progress in International Reading Literacy Study (PIRLS), and Trends in International Mathematics and Science Study (TIMSS), play an important role in educational research and policy making. They collect valuable data on education quality and performance development across many education systems in the world, giving countries the opportunity to share techniques, organisational structures, and policies that have proven efficient and successful (Singer et al., 2018; von Davier et al., 2012).

PISA is a worldwide study by the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations intended to evaluate educational systems by measuring 15-year-old school students' scholastic performance in the subjects of mathematics, science, and reading, as well as a large number of non-cognitive variables, such as students' socioeconomic status, family background, and learning experiences. Students' scholastic performance is measured by response data from cognitive items that measure ability/proficiency in each of the three subjects, and non-cognitive variables are collected through non-cognitive questionnaires for students, school principals,

Received: August 16, 2022. **Revised:** August 1, 2023. **Accepted:** November 12, 2023

teachers, and parents. In this study, we focus on the knowledge domain of science in PISA 2015, where science was the assessment focus in this survey. Given the importance of science education (Gov.UK, 2015; National Research Council, 2012), it is of particular interest for educators and policymakers to understand what non-cognitive variables (e.g. socioeconomic status, family background, learning experiences) are significantly associated with student's knowledge of science. Naturally, one would consider a regression model with students' performance in science as the response variable and the non-cognitive variables as predictors and identify the predictors with non-zero regression coefficients. Seemly straightforward, constructing such a regression model and then selecting the non-null variables is nontrivial due to three challenges brought by the complexity of the current problem. First, students' performance in science is not directly observed but instead measured by a set of test items. The measurement is further complicated by a matrix sampling design adopted by PISA (Gonzalez & Rutkowski, 2010). That is, each student is administered a small subset of available cognitive items in order to cover an extensive content domain while not overburdening students and schools in terms of their time and administration costs. Consequently, one cannot simply calculate a total score as a surrogate for student science performance. We note that OECD provides plausible values, which are obtained using a multiple imputation procedure (von Davier et al., 2009), as a summary of each student's overall performance in each subject domain. However, it is not suitable to use a plausible value as the response variable when performing the current variable selection task. This is because the multiple imputation procedure for producing the plausible values involves the predictors through a principal component analysis step (Chapter 9, OECD, 2016b), due to which all the predictors are associated with the plausible values and thus, performing variable selection is not sensible. Second, students' non-cognitive variables are collected via survey questions, which contain many missing values. In fact, in the US sample considered in the current study, around 6% of the entries are missing, and the proportion of sample points that are fully observed is less than 26%. Consequently, it is virtually impossible to conduct the regression analysis without a proper treatment of the missing values. Finally, PISA collects a large number of non-cognitive variables. In the current study of PISA 2015 data, we have 62 predictors, even though careful pre-processing is performed that substantially reduces the number of variables. Due to the multiple comparison issues, it is a challenge to control for a reasonable error metric when conducting variable selection.

We tackle these challenges through several methodological contributions. We introduce a latent construct for science knowledge and use an Item Response Theory (IRT) model (Chen et al., 2023) to measure this latent construct based on students' responses to science items. The relationship between the latent construct and non-cognitive variables is further modelled through a structural model that regresses the latent construct onto the non-cognitive variables. This structural equation model is often known as the latent regression IRT model, or simply the latent regression model (Mislevy, 1984; von Davier & Sinharay, 2010). When there are many missing values in the non-cognitive variables, estimating the latent regression model is a challenge. To tackle this problem, we propose to model the predictors using a Gaussian copula model (Fan et al., 2017; Han & Pan, 2012), which allows the predictors to be of mixed types (e.g. continuous, binary, ordinal). Thanks to the Gaussian copula model, we can estimate the latent regression model with a likelihood-based estimator. In dealing with multiple comparisons, we consider the knockoff framework for controlled variable selection (Barber & Candès, 2015; Candès et al., 2018). More specifically, we adapt the derandomised knockoffs method (Ren et al., 2023) to the current latent regression model with missing values. This approach allows us to control the per family error rate (PFER), i.e. the expected number of false positives among the detections. We choose the derandomised knockoff method instead of the Model-X knockoff method because the latter is a randomised procedure that may suffer from a high Monte Carlo error. The derandomised knockoff method leverages the Model-X knockoff method by aggregating the results from multiple knockoff realisations. To our best knowledge, this is the first time that missing data are considered in a knockoff approach with theoretical guarantees.

In real-world applications, especially in social sciences, missing data are commonly encountered. In addition, many variables of interest, such as individuals' attitudes, personality traits, and abilities, are latent constructs that are not directly observable. They are often defined by multiple indicators and play the role of a response variable or predictors in a model (Chapter 4, Skrondal & Rabe-Hesketh, 2004). For example, the latent construct for students' science knowledge is such a variable, and it serves as the response variable in the latent regression analysis of PISA data. While we focus on data from an education survey and a tailored latent regression

model, we describe the proposed knockoff method under a general latent variable model framework so that the proposed method can be applied to variable selection problems involving missing values, latent constructs or both. Model selection of latent variable models is usually performed based on information criteria, such as the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). These methods suffer from several issues under the current complex data setting. First, when the latent regression model involves many predictors, an information criterion needs to be combined with a search method, such as a stepwise selection method or a Lasso-type regularised estimation method. The search method is used to retain a smaller number of candidate models from the original model space that is exponentially large so that the information criterion can be computed. Even so, this approach may be computationally infeasible when there exist many latent variables or missing values, as a search method needs to optimise many marginal likelihoods that involve high-dimensional integrals. For regularised estimation methods, the optimisation additionally involves non-smooth regularisation terms and thus can be computationally even more time-consuming. Moreover, stepwise selection methods are greedy algorithms that lack a theoretical guarantee for identifying the true model. Second, the computational burden with the information-criteria-based methods mostly comes from handling missing data and latent variables. One may naturally wonder whether we can use a two-step procedure that first handles the missing data and latent variables using an off-the-shell missing data handling methods, such as imputation methods (Little & Rubin, 2019; Van Buuren, 2018) and missing indicator methods (Cohen & Cohen, 1975; Dardanoni et al., 2015, 2011), and then applies an information criterion to the imputed or augmented data. Unfortunately, such a procedure lacks theoretical justification and is often practically infeasible or inaccurate. For example, the missing indicator method cannot be performed when some predictors are latent variables. In addition, a small simulation study in Section F.2 of the online supplementary material shows that the BIC performs poorly when calculated based on imputed data. Finally, the proposed method is more flexible, as it allows the users to choose the threshold for the PFER, allowing a trade-off between type I and type II error rates. This is an advantage that model selection based on an information criterion does not offer.

The remainder of the article is structured as follows. Section 2 provides the background on the central substantive question—how students' knowledge of science is associated with their non-cognitive variables—and a description of the PISA 2015 data. In Section 3, we introduce the latent regression IRT model for studying the relationship between a latent construct of science knowledge and non-cognitive variables and a Gaussian copula model for handling missing predictors, which are of mixed types. Section 4 proposes knockoff methods for controlled variable selection under the latent regression IRT model with missing data. The proposed method is evaluated via a simulation study in Section 5 and then applied to data from PISA 2015 in Section 6. Finally, we discuss the implications of our results and possible directions for future research in Section 7. Proof of theoretical results, details of computation, additional simulation studies, and further information about the PISA data are given in the online supplementary material.

2 Background and overview of PISA 2015 data

2.1 Academic achievement and non-cognitive predictors

The term 'non-cognitive' typically refers to a broad range of personal attributes, skills, and characteristics representing one's attitudinal, behavioural, emotional, motivational, and other psychosocial dispositions. It is often used as a catch-all phrase encompassing variables that are potentially important for academic achievement but not measured by typical achievement or cognitive tests (Farkas, 2003). Social science researchers have devoted considerable research effort towards identifying non-cognitive predictors of students' academic achievement (e.g. Duckworth & Yeager, 2015; Lee & Stankov, 2018; Richardson et al., 2012).

Science has changed our lives and is vital to the future prosperity of society. Thus, science education plays an important role in the modern education system (Gov.UK, 2015; National Research Council, 2012). Identifying the predictors of science education helps educators, policymakers, and other stakeholders understand the psychosocial factors behind science education, which may lead to better policies and practices of science education. PISA, which collects both students' science

achievement and non-cognitive variables, provides a great opportunity for identifying the key non-cognitive predictors of science achievement.

2.2 PISA 2015 data

PISA is conducted in a 3-year cycle, with each cycle focusing on one of the three subjects, i.e. mathematics, science, and reading. PISA 2015 is the most recent cycle that focused on science. It collected data from 72 participating countries and economics. Computer-based tests were used, with assessments lasting a total of 2 h for each student. Following a matrix sampling design, different students took different combinations of test items on science, reading, mathematics, and collaborative problem-solving. Test items involved a mixture of multiple-choice and constructive-response questions. See OECD (2016b) for the summary of the design and results of PISA 2015.

This study considers a subset of the PISA 2015 dataset. Specifically, to avoid modelling country heterogeneity, we considered data from a single country, the U.S. After some data pre-processing which excluded observations with poor-quality data, the sample size is 5,685. PISA 2015 contained 184 items in the science domain that were dichotomously or polytomously scored. Due to the matrix sampling design of PISA, on average, each student was only assigned 16.25% of the items.

In addition, we consider non-cognitive variables collected by the student survey, which provides information about the students themselves, their homes, and their school and learning experiences. We constructed 62 variables as candidates in variable selection. These variables include 11 raw responses to questionnaire items [e.g. GENDER (gender), LANGAH (language at home)), 34 indices that OECD constructed (e.g. CULTPO (cultural possession), HEDRES (home educational resources)], and 17 composites that we constructed based on students' responses to questionnaire items [e.g. OUT.GAM (play games out of school), OUT.REA (reading out of school)]. We decided to include these constructed variables rather than the corresponding raw responses for better substantive interpretations. For some ordinal variables, certain adjacent categories were merged due to sample size considerations. Details of these 62 candidate variables are given in Section 6 and the online supplementary material. Unlike the cognitive items, students were supposed to answer all the items in the student survey. However, there are still many missing responses in the student survey data. Among the candidate variables, 20 variables have more than 5% of their data missing, and the variable DUECEC (duration in early childhood education and care) has the largest missing rate, 37.17%.

3 Model framework

In this section, we describe a general latent variable model framework, which includes the latent regression model for analysing PISA data as a special case. The model is defined through (a) a structural model, (b) a measurement model, and (c) a data missingness mechanism.

3.1 Structural model

We consider data collected from N observations. For each observation, there is a response variable θ_i and predictors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$, where θ_i and some or all entries of $\mathbf{Z}_i$ can be latent constructs measured by observed indicators. We allow the variables in $\mathbf{Z}_i$ to be binary, ordinal, continuous, or a mixture of them. Without loss of generality, we assume that $Z_{i1},\ldots, Z_{ip_1}$ are continuous latent constructs, with $p_1 = 0$ when all entries of $\mathbf{Z}_i$ are observable. In the PISA application, each observation is a student, θ_i represents the student's latent construct on science knowledge, and $\mathbf{Z}_i$ contains observable non-cognitive predictors.

The structural model defines the joint distribution of $(\theta_i, \mathbf{Z}_i)$ through two steps—(a) the conditional distribution of θ_i given $\mathbf{Z}_i$ and (b) the marginal distribution of $\mathbf{Z}_i$. A linear regression model is assumed for θ_i given $Z_{i1},\ldots, Z_{ip}$. More specifically, for each variable j, we introduce a transformation $g_j(Z_j)$. When Z_j is an ordinal variable with categories $\{0, \ldots, K_j\}$, the transformation function g_j creates K_j dummy variables, i.e. $g_j(Z_j) = (\mathbb{I}(\{Z_j \geq 1\}), \ldots, \mathbb{I}(\{Z_j \geq K_j\}))^\top$. For continuous and binary variables, g_j is an identity link, i.e. $g_j(Z_j) = Z_j$. We assume

$$\theta_i|\mathbf{Z}_i \sim N(\beta_0 + \boldsymbol{\beta}_1^\top g_1(Z_{i1}) + \cdots + \boldsymbol{\beta}_p^\top g_p(Z_{ip}), \sigma^2), \quad (1)$$

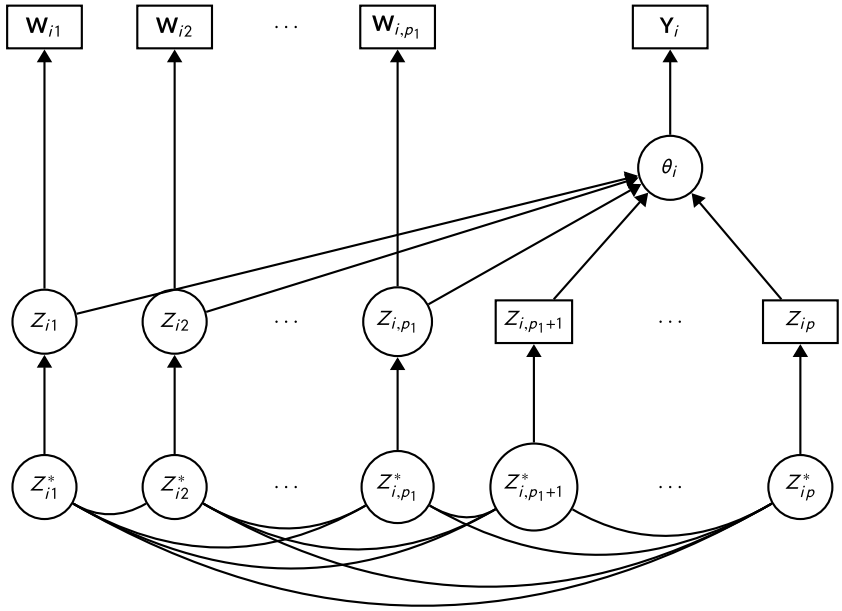

Figure 1. Path diagram for the general latent variable model framework. Variables within a circle represent unobserved or latent variables, while those within a rectangle represent observed variables. The measurement models are represented by the directed edges from θ_i to $\mathbf{Y}_i$ and those from Z_{ik} to $\mathbf{W}_{ik}$, $k = 1, 2, \ldots, p_1$. The structural model is represented by the directed edges from Z_{ij}s to θ_i. The predictor model is represented by the directed edges from Z^*_{ij} to Z_{ij} and the undirected edges between Z^*_{ij}s.

where β_0 is the intercept, $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$ are the slope parameters, and σ^2 is the residual variance. Note that $\boldsymbol{\beta}_j$ is a scalar when predictor j is continuous or binary and is a vector when the predictor is ordinal. Here, β_0, $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$, and σ^2 are unknown and will be estimated from the model. The main goal of our analysis is to find predictors for which $\|\boldsymbol{\beta}_j\| \neq 0$. The structural model is depicted in Figure 1.

Since $\mathbf{Z}_i$ may contain variables of mixed types, one cannot simply adopt a Gaussian assumption. Here, we consider a Gaussian copula model. This model introduces underlying random variables $\mathbf{Z}^*_i = (Z^*_{i1}, \ldots, Z^*_{ip})^\top$, for which $\mathbf{Z}^*_1, \ldots, \mathbf{Z}^*_N$ are independent and identically distributed, following a p-variate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. We assume that the normal distribution is non-degenerate, i.e. $\text{rank}(\boldsymbol{\Sigma}) = p$. Each underlying variable Z^*_{ij} is assumed to marginally follow a standard normal distribution, i.e. the diagonal entries of $\boldsymbol{\Sigma}$ are 1. Each predictor Z_{ij} is assumed to be a transformation of its underlying variable Z^*_{ij}, denoted by $Z_{ij} = F_j(Z^*_{ij})$. For a continuous predictor j, let $F_j(Z^*_{ij}) = c_j + d_j Z^*_{ij}$, where c_j and d_j are unknown parameters. For the latent constructs, we let $Z_{ij} = Z^*_{ij}$ as their location and scale need to be fixed for identification, i.e. $c_j = 0$ and $d_j = 1, j = 1, \ldots, p_1$. For a binary or ordinal predictor j, let $F_j(Z^*_{ij}) = k$ if $Z^*_{ij} \in (c_{jk}, c_{j,k+1}]$, $k = 0, \ldots, K_j$, where $c_{j1}, \ldots, c_{jK_j}$ are unknown parameters, and $c_{j0} = -\infty$ and $c_{j,K_j+1} = \infty$. Note that $K_j = 1$ for a binary variable and $K_j > 1$ for an ordinal variable. The predictor model is also illustrated in Figure 1.

We note that the above model specifies a joint distribution for $Z_{i1}, \ldots, Z_{ip}$. More specifically, let $\mathcal{D} \subset \{1, \ldots, p\}$ be the set of dichotomous and polytomous predictors. We use Ξ as generic notation for the unknown parameters in the Gaussian copula model, including $\boldsymbol{\Sigma}$ and the parameters in the transformations between Z^*_{ij} and Z_{ij}. We further use $\phi(\cdot \mid \boldsymbol{\Sigma})$ to denote the density function of the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Then, the density function of $\mathbf{Z}_i$ takes the form

$$f(\mathbf{z}|\Xi) = \int \ldots \int \left\{ \left(\prod_{j \in \mathcal{D}} \mathrm{d}z^*_j \right) \times \left[\phi(\mathbf{z}^*|\boldsymbol{\Sigma}) \times \left(\prod_{j \notin \mathcal{D}} d_j^{-1} \right) \times \left(\prod_{j \in \mathcal{D}} \mathbb{I}(z^*_j \in (c_{j,z_j-1}, c_{j,z_j}]) \right) \right] \Bigg|_{z^*_j = \frac{z_j - c_j}{d_j}, j \notin \mathcal{D}} \right\}. \tag{2}$$

3.2 Measurement model

We note that θ_i is a latent construct that is not directly observable. In addition, some variables in $\mathbf{Z}_i$ may also be latent constructs. The latent constructs are defined by observable data through a measurement model. We now specify this measurement model based on complete data (i.e. no data are missing). The treatment of missing values is left to Section 3.3. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$ be the indicators for θ_i. And let $\mathbf{W}_{ij} = (W_{ij1}, \ldots, W_{ijl_j})^\top$ be the indicators for $Z_{ij}, j = 1, \ldots, p_1$. The measurement model defines the conditional distribution of $(\mathbf{Y}_i^\top, \mathbf{W}_{i1}^\top, \ldots, \mathbf{W}_{ip_1}^\top)^\top$ given $(\theta_i, Z_{i1}, \ldots, Z_{ip_1})^\top$. This conditional distribution is specified by assuming (a) $\mathbf{Y}_i$ is conditionally independent of all the other variables given θ_i, (b) $\mathbf{W}_{ij}$ is conditionally independent of all the other variables given Z_{ij}, for all $j = 1, \ldots, p$, and (c) the conditional model of $\mathbf{Y}_i$ given θ_i and those of $\mathbf{W}_{ij}$ given Z_{ij}. These conditional models are visualised in Figure 1. We now elaborate on these conditional models.

Conditional model of $\mathbf{Y}_i$ given θ_i. If θ_i is observable, then we just let $\mathbf{Y}_i = \theta_i$, and in this case, the conditional model of $\mathbf{Y}_i$ given θ_i is degenerate. Otherwise, when θ_i is a latent construct, a unidimensional linear factor model or IRT model (Chapter 3, Skrondal & Rabe-Hesketh, 2004) can be used for this conditional distribution, depending on the variable types in $\mathbf{Y}_i$. We assume that this measurement model satisfies the standard identifiability conditions. In the application to PISA data, students' science performance in science, θ_i, is measured by cognitive items. In this application, $\mathbf{Y}_i$ contains students' responses to cognitive items, where the responses are either binary (correct/incorrect) or ordinal. In what follows, we describe the measurement model used in the scaling of PISA 2015 data (Chapter 9, OECD, 2016b). This model will be used in our simulation studies and application to PISA data.

More specifically, this model assumes local independence, an assumption that is commonly adopted in IRT models (Embretson & Reise, 2000). That is, $Y_{ij}, j = 1, \ldots, J$, 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)

$$\mathbb{P}(Y_{ij} = 1|\theta_i) = \frac{\exp(a_j\theta_i + b_j)}{1 + \exp(a_j\theta_i + b_j)}, \tag{3}$$

where a_j and b_j are two item-specific parameters. For a polytomous item j with $K_j + 1$ categories, Y_{ij} given θ_i is assumed to follow a generalised partial credit model (GPCM, Muraki, 1992), for which

$$\mathbb{P}(Y_{ij} = k|\theta_i) = \frac{\exp\left[\sum_{r=1}^{k}(a_j\theta_i + b_{jr})\right]}{1 + \sum_{k'=1}^{K_j}\exp\left[\sum_{r=1}^{k'}(a_j\theta_i + b_{jr})\right]}, \quad k = 1, \ldots, K_j, \tag{4}$$

where $a_j, b_{j1}, b_{j2}, \ldots, b_{j,K_j}$ are item-specific parameters. In OECD's analysis of PISA data, the item-specific parameters are first calibrated based on item response data from all the countries and then treated as known when inferring the proficiency level of students or the proficiency distributions of countries (Chapter 9, OECD, 2016b). We follow this routine when analysing PISA data. Specifically, the item-specific parameters are fixed to the values used by OECD for scaling PISA 2015 data[1].

In the rest, we denote the conditional probability density/mass function of $\mathbf{Y}_i$ given θ_i at $\mathbf{Y}_i = \mathbf{y}_i$ as $h(\mathbf{y}_i|\theta_i; \Delta)$, where Δ denotes the unknown parameters in this conditional model. In the PISA application, $h(\mathbf{y}_i|\theta_i; \Delta) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij}|\theta_i)$, where $P(Y_{ij} = y_{ij}|\theta_i)$ follows (3) or (4) depending on whether item j is dichotomous or polytomous. As all the item parameters are pre-calibrated in this application, Δ becomes an empty vector and will not be estimated. In situations where the item parameters are unknown, Δ can be estimated from data; see Section F.1 in the online supplementary material for a simulation study under this setting.

Conditional model of $\mathbf{W}_{ij}$ given Z_{ij}. When Z_{ij} is a latent construct, a unidimensional linear factor model or IRT model can be used for this conditional distribution, depending on the variable types

[1] The item parameters can be found from: https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final-AnnexA.pdf

in $\mathbf{W}_{ij}$. We assume that these measurement models satisfy the standard identifiability conditions. In the rest, we denote the conditional probability density/mass function of $\mathbf{W}_{ij}$ given Z_{ij} at $\mathbf{W}_{ij} = \mathbf{w}_{ij}$ as

$$q_j(\mathbf{w}_{ij}|Z_{ij}; \Lambda_j), \quad j = 1, \ldots, p_1, \tag{5}$$

where Λ_j denotes the unknown parameters in this conditional model if they exist.

3.3 Data missingness and statistical inference

In PISA data, as well as many other multivariate data in the social and behavioural sciences, there are often a substantial proportion of missing values. Here, we impose assumptions for data missingness. First, we assume that entries of $\mathbf{Y}_i$ are missing completely at random. This assumption is sensible in the PISA application, where the missing responses to cognitive items are due to the matrix sampling design of PISA. Second, we assume that W_{i1},..., W_{ip_1} do not have missing values. This assumption is made for simplicity and can be easily relaxed. Finally, we assume that the missing data in $(Z_{i,p_1+1}, \ldots, Z_{ip})^\top$ are missing at random (MAR), which is a quite strong but commonly adopted assumption in missing data analysis (Little & Rubin, 2019; Van Buuren, 2018).

More specifically, let $\mathbf{w}_i$ be the realisation of $\mathbf{W}_i = (\mathbf{W}_{i1}^\top, \ldots, \mathbf{W}_{ip_1}^\top)^\top$, $i = 1, \ldots, N$, and recall that Ξ denotes the unknown parameters of the Gaussian copula model and Λ_1, ..., Λ_{p_1} denote the parameters in the measurement models for Z_{ij}, $j = 1$. Let $\mathcal{B}_i$ be an index set containing all the indicators j such that Y_{ij} is not missing. We let $\mathbf{Y}_i^{\text{obs}} = \{Y_{ij} : j \in \mathcal{B}_i\}$ be the observed indicators for θ_i and let $\mathbf{Y}_i^{\text{mis}} = \{Y_{ij} : j \notin \mathcal{B}_i\}$ be the missing ones. Similarly, we let $\mathcal{A}_i$ be the set indicating all the observed variables in $(Z_{i,p_1+1}, \ldots, Z_{ip})^\top$, and let $\mathbf{Z}_i^{\text{obs}} = \{Z_{ij} : j \in \mathcal{A}_i\}$ and $\mathbf{Z}_i^{\text{mis}} = \{Z_{ij} : j \notin \mathcal{A}_i\}$. Under the MAR assumption, the log-likelihood function for Ξ, Λ_1, ..., and Λ_{p_1} takes the form $l_1(\Xi, \Lambda_1, \ldots, \Lambda_{p_1}) = \sum_{i=1}^N \log f_i(\mathbf{w}_i, \mathbf{z}_i^{\text{obs}}|\Xi, \Lambda_1, \ldots, \Lambda_{p_1})$, where $f_i(\mathbf{w}_i, \mathbf{z}_i^{\text{obs}}|\Xi, \Lambda_1, \ldots, \Lambda_{p_1}) = \int \cdots \int f(\mathbf{z}_i|\Xi)(\prod_{j=1}^{p_1} q_j(\mathbf{w}_{ij}|z_{ij}; \Lambda_j))(\prod_{j \notin \mathcal{A}_i} dz_{ij})$. Note that the integrals in $f_i(\mathbf{w}_i, \mathbf{z}_i^{\text{obs}}|\Xi, \Lambda_1, \ldots, \Lambda_{p_1})$ are with respect to $\mathbf{Z}_i^{\text{mis}}$.

The maximum likelihood estimator for $\Xi, \Lambda_1, \ldots, \Lambda_{p_1}$ is given by

$$\begin{aligned} (\hat{\Xi}, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_{p_1}) = \underset{\Xi, \Lambda_1, \ldots, \Lambda_{p_1}}{\arg\max} \quad & l_1(\Xi, \Lambda_1, \ldots, \Lambda_{p_1}) \\ \text{subject to} \quad & \Sigma_{jj} = 1, \ j = 1, \ldots, p, \\ & d_j > 0, j \notin \mathcal{D}, \ c_{j1} < c_{j2} < \ldots < c_{jK_j}, j \in \mathcal{D}. \end{aligned} \tag{6}$$

We note that this optimisation problem involves high-dimensional integrals and constraints. We adopt a stochastic proximal gradient algorithm proposed in Zhang and Chen (2022). In this algorithm, the integrals are handled by Monte Carlo sampling of the missing values, and the unknown parameters are updated by stochastic proximal gradient descent, in which constraints are handled. The details of this algorithm can be found in the online supplementary material.

Given $\hat{\Xi}$, $\hat{\Lambda}_1, \ldots, \hat{\Lambda}_{p_1}$ from (6), one can estimate the rest of the unknown parameters, including the regression coefficients in (1) that are of major interest. We denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$. Let $\mathbf{y}_i^{\text{obs}}$ be the realisation of $\mathbf{Y}_i^{\text{obs}}$, $i = 1, \ldots, N$. The log-likelihood for $\boldsymbol{\beta}$, β_0, σ^2, and Δ takes the form

$$l_2(\boldsymbol{\beta}, \beta_0, \sigma^2, \Delta) = \sum_{i=1}^N \log\left[\int \cdots \int \left(\prod_{j \notin \mathcal{A}_i} \mathrm{d}z_{ij}\right) f(\mathbf{z}_i|\hat{\Xi}) \left(\prod_{j=1}^{p_1} q_j(\mathbf{w}_{ij}|z_{ij}; \hat{\Lambda}_j)\right) f_i(\mathbf{y}_i^{\text{obs}}|\mathbf{z}_i; \boldsymbol{\beta}, \beta_0, \sigma^2, \Delta)\right], \tag{7}$$

where $f_i(\mathbf{y}_i^{\text{obs}}|\mathbf{z}_i; \boldsymbol{\beta}, \beta_0, \sigma^2, \Delta)$ is the conditional density function of $\mathbf{Y}_i^{\text{obs}}$ given $\mathbf{Z}_i = \mathbf{z}_i$

$$\begin{aligned} f_i(\mathbf{y}_i^{\text{obs}}|\mathbf{z}_i; \boldsymbol{\beta}, \beta_0, \sigma^2, \Delta) = & \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right. \\ & \left. \times \int \cdots \int \mathrm{d}\theta_i \left(\prod_{j \notin \mathcal{B}_i} \mathrm{d}y_{ij}\right) h(\mathbf{y}_i|\theta_i; \Delta) \exp\left(-\frac{(\theta_i - (\beta_0 + \boldsymbol{\beta}_1^\top g_1(z_{i1}) + \cdots + \boldsymbol{\beta}_p^\top g_p(z_{ip}))^2}{2\sigma^2}\right)\right]. \end{aligned} \tag{8}$$

We estimate $\boldsymbol{\beta}$, β_0 , σ^2 and Δ by maximising $l_2(\boldsymbol{\beta}, \beta_0, \sigma^2, \Delta)$. Similar to the optimisation (6), the maximisation of $l_2(\boldsymbol{\beta}, \beta_0, \sigma^2, \Delta)$ also involves high-dimensional integrals. We carry out this optimisation using a stochastic Expectation-Maximisation (EM) algorithm[2] (Nielsen, 2000; Zhang et al., 2020). The details are given in the online supplementary material.

4 Variable selection via knockoffs

4.1 Problem setup and knockoffs

As mentioned previously, our goal is to solve a model selection problem, i.e. to find the non-null predictors for which $\|\boldsymbol{\beta}_j\| \neq 0$. We hope to control the statistical error in the model selection to assure that most of the discoveries are indeed true and replicable. This is typically achieved by controlling for a certain risk function, such as the false discovery rate, the k-familywise error rate, and the per familywise error (PFER); see Janson and Su (2016) and Candès et al. (2018). Let $\hat{\mathcal{S}}$ and $\mathcal{S}^* \subset \{1, \ldots, p\}$ be the selected and true non-null predictors, respectively. The current study concerns the control of PFER, defined as $\mathbb{E}|\hat{\mathcal{S}} \setminus \mathcal{S}^*|$, where $|\cdot|$ denotes the number of elements in a set.

The knockoff method is a general framework for controlled variable selection. The key to a knockoff method is the construction of knockoff variables, where the knockoff variables mimic the dependence structure within the original variables but are null variables (i.e. not associated with the response variable). They serve as negative controls in the variable selection procedure that help identify the truly important predictors while controlling for a certain risk function, such as the PFER. Many knockoff methods have been developed (Barber & Candès, 2015, 2019; Candès et al., 2018; Fan et al., 2019, 2020; Janson & Su, 2016; Romano et al., 2020; Sesia et al., 2019). Many knockoff methods are based on the model-X knockoff framework (Candès et al., 2018), which is very flexible and can be extended to the current setting involving missing data and mixed-type predictors. However, one drawback of the model-X knockoffs is that it only takes one draw of the knockoff variables through Monte Carlo sampling. As a result, this procedure often suffers from high uncertainty brought about by the Monte Carlo error, even though the risk function is controlled. To alleviate this uncertainty, which has important implications on the interpretability of the variable selection results, we adopt the derandomised knockoff method (Ren et al., 2023). This method can substantially reduce the Monte Carlo error by aggregating the selection results across multiple runs of a knockoff algorithm. In what follows, we first introduce the way of constructing knockoff variables under the joint model described in the above section and then introduce a derandomised knockoff procedure for controlling PFER.

4.2 Constructing knockoffs with missing data

We extend the concept of knockoffs to the missing data setting. To control the variable selection error with the knockoff procedure introduced below, a stronger MAR condition is needed. It is called the SMAR condition as introduced in Definition 1 below.

Definition 1 (SMAR condition). Let $\mathbf{X}_i = (\mathbf{X}_{i1}^\top, \ldots, \mathbf{X}_{ip}^\top)^\top$, such that $\mathbf{X}_{ij} = \mathbf{W}_{ij}$ if $j = 1, \ldots, p_1$ and $\mathbf{X}_{ij} = Z_{ij}$ otherwise. Consider the conditional distribution of $\mathcal{A}_i$ given $\mathbf{X}_i$. Let $q(\boldsymbol{\alpha}|\mathbf{x}_i)$ denote the conditional probability mass function of $\mathcal{A}_i$ given $\mathbf{X}_i$. We say the SMAR condition holds with respect to the non-null variables $\mathcal{S}^*$, if $q(\boldsymbol{\alpha}|\mathbf{x}) = q(\boldsymbol{\alpha}|\mathbf{x}')$ holds, for any $\boldsymbol{\alpha}$, $\mathbf{x} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_{p_1}^\top)^\top$ and $\mathbf{x}' = (\mathbf{x}_1'^\top, \ldots, \mathbf{x}_p'^\top)^\top$ satisfying $\{\mathbf{x}_j : j \in \{1, \ldots, p\} \cap \mathcal{S}^*\} = \{\mathbf{x}_j' : j \in \{1, \ldots, p\} \cap \mathcal{S}^*\}$.

This SMAR condition says that the probability of being missing is the same within groups defined by the observed non-null variables. It is stronger than MAR because MAR only requires

[2] The stochastic proximal gradient algorithm used for the optimisation problem (6) can also be used to solve the current optimisation problem. The stochastic EM algorithm is chosen as it tends to converge empirically faster for the current problem.

$q(\boldsymbol{\alpha}|\mathbf{x}) = q(\boldsymbol{\alpha}|\mathbf{x}')$ to hold, for any $\boldsymbol{\alpha}$, $\mathbf{x} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_{p_1}^\top)^\top$, and $\mathbf{x}' = (\mathbf{x}_1'^\top, \ldots, \mathbf{x}_p'^\top)^\top$ satisfying $\{\mathbf{x}_i : i \in \boldsymbol{\alpha}\} = \{\mathbf{x}_i' : i \in \boldsymbol{\alpha}\}$, i.e. the probability of being missing is the same within groups defined by the observed variables, regardless of whether they are in $\mathcal{S}^*$ or not. On the other hand, the SMAR condition is weaker than completely missing at random (MCAR), as MCAR implies that $q(\boldsymbol{\alpha}|\mathbf{x}) = q(\boldsymbol{\alpha}|\mathbf{x}')$ for all $\boldsymbol{\alpha}$, $\mathbf{x}$, and $\mathbf{x}'$. Throughout the rest, the SMAR condition is assumed.

Definition 2 (Knockoffs). Suppose that the SMAR condition in Definition 1 holds and the true values of Ξ, Λ_1,..., Λ_{p_1}, $\boldsymbol{\beta}$, β_0, σ^2, and Δ are known. Under the setting in Section 3, we say that $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$ is a knockoff copy of $(\mathbf{W}_i, \mathbf{Z}_i^{\text{obs}})$, if there exists a random vector $\tilde{\mathbf{W}}_i = (\tilde{\mathbf{W}}_{i1}^\top, \ldots, \tilde{\mathbf{W}}_{ip_1}^\top)^\top$, where $\tilde{\mathbf{W}}_{ij} = (\tilde{\mathbf{W}}_{ij1}, \ldots, \tilde{\mathbf{W}}_{ijl_j})^\top$ for each $j = 1, \ldots, p_1$, as well as underlying variables $\mathbf{Z}_i^* = (Z_{i1}^*, \ldots, Z_{ip}^*)^\top$ and $\tilde{\mathbf{Z}}_i^* = (\tilde{Z}_{i1}^*, \ldots, \tilde{Z}_{ip}^*)^\top$, such that

(1) $\mathbf{Z}_i^{\text{obs}} = \{F_j(Z_{ij}^*) : j \in \mathcal{A}_i\}$ and $\tilde{\mathbf{Z}}_i^{obs} = \{F_j(\tilde{Z}_{ij}^*) : j \in \mathcal{A}_i\}$;
(2) $\mathbf{Y}_i^{\text{obs}}$, $\mathbf{W}_i$, and $\tilde{\mathbf{Z}}_i^*$ are conditionally independent given $\mathbf{Z}_i^*$;
(3) Let $\mathbf{Z}_i = (F_1(Z_{i1}^*, \ldots, F_p(Z_{ip}^*))^\top$. Then $\mathbf{Z}_i$ follows the Gaussian copula model (2), $\mathbf{W}_{ij}$ given Z_{ij} follows the conditional model (5) for each $j = 1, \ldots, p_1$, and $\mathbf{Y}_i^{\text{obs}}$ given $\mathbf{Z}_i$ follows the conditional model (8).
(4) Let $\tilde{\mathbf{Z}}_i = (F_1(\tilde{Z}_{i1}^*), \ldots, F_p(\tilde{Z}_{ip}^*))^\top$. Then for each $j = 1, \ldots, p_1$, the conditional probability density/mass function of $\tilde{\mathbf{W}}_{ij}$ given $\tilde{Z}_{ij}$ is the same as (5).
(5) For any subset $\mathcal{S} \subset \{1, \ldots, p\}$, $(\mathbf{Z}_i^*, \tilde{\mathbf{Z}}_i^*)_{\text{swap}(\mathcal{S})}$ and $(\mathbf{Z}_i^*, \tilde{\mathbf{Z}}_i^*)$ are identically distributed.

Recall that $F_j(\cdot)$ is the transformation between each predictor and its underlying variable. In addition, the vector $(\mathbf{Z}_i^*, \tilde{\mathbf{Z}}_i^*)_{\text{swap}(\mathcal{S})}$ is obtained from $(\mathbf{Z}_i^*, \tilde{\mathbf{Z}}_i^*)$ by swapping the entries Z_{ij}^* and $\tilde{Z}_{ij}^*$ for each $j \in \mathcal{S}$; for example, with $p = 3$ and $\mathcal{S} = \{1, 3\}$, $(Z_{i1}^*, Z_{i1}^*, Z_{i3}^*, \tilde{Z}_{i1}^*, \tilde{Z}_{i2}^*, \tilde{Z}_{i3}^*)_{\text{swap}(\{1,3\})} = (\tilde{Z}_{i1}^*, Z_{i2}^*, \tilde{Z}_{i3}^*, Z_{i1}^*, \tilde{Z}_{i2}^*, Z_{i3}^*)$. We compare the current definition of knockoffs under a missing data setting with the standard definition for model-X knockoffs in Candès et al. (2018). The model-X knockoff framework assumes no missing data in predictors $\mathbf{Z}_i$, and $\mathbf{Z}_1$,..., $\mathbf{Z}_N$ are independent and identically distributed. Therefore, the definition of model-X knockoff omits the subscript i. On the other hand, the current analysis depends on $\mathcal{A}_i$, which differs across observations. Consequently, knockoffs are defined for each $\mathbf{Z}_i^{\text{obs}}$. When there are no unobservable predictors and no missing data, i.e. $p_1 = 0$ and $\mathcal{A}_i = \{1, \ldots, p\}$, $i = 1, \ldots, N$, the current definition coincides with the definition in Candès et al. (2018). Note that stronger conditions are needed for the construction of knockoffs when there exist missing data. These conditions (e.g. SMAR) are needed to ensure that the joint distribution of $\mathbf{Y}_i^{\text{obs}}$, $\mathcal{A}_i$, $\mathbf{W}_i$, $\mathbf{Z}_i^{\text{obs}}$, $\tilde{\mathbf{W}}_i$, and $\tilde{\mathbf{Z}}_i^{\text{obs}}$ remains identical when swapping the null indices, which is essential for establishing the exchangeability property (Candès et al., 2018) for controlling variable selection error. Specifically, under Definition 2, $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$ and $\mathbf{Y}_i^{\text{obs}}$ are likely not conditionally independent given $(\mathbf{W}_i, \mathbf{Z}_i^{\text{obs}})$. Consequently, when constructing the knockoff variables $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$, one needs information from not only $(\mathbf{W}_i, \mathbf{Z}_i^{\text{obs}})$ but also $\mathbf{Y}_i^{\text{obs}}$, to compensate for the missing information. In other words, the joint distribution of $\mathbf{Y}_i^{\text{obs}}$, $\mathbf{W}_i$, and $\mathbf{Z}_i^{\text{obs}}$ is needed to construct $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$.

In what follows, we present an algorithm for constructing knockoffs $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$ under Definition 2. To ensure the exact satisfaction of Definition 2, we assume that the true model parameters are known. In practice, we plug an estimate of the parameters into the algorithm; see Section 4.4 for theoretical justifications and further discussions.

Algorithm 1 (Constructing knockoff copies)

Input: Observed data $\mathbf{Y}_i^{\text{obs}}$, $\mathbf{W}_i$, and $\mathbf{Z}_i^{\text{obs}}$, $i = 1, \ldots, N$, the true model parameters Ξ of the Gaussian copula model, the true parameters $\Lambda_1, \ldots, \Lambda_{p_1}$ in the measurement models for Z_{ij}, and the true parameters $\boldsymbol{\beta}, \beta_0, \sigma^2, \Delta$ in the conditional model of $\mathbf{Y}_i$ given $\mathbf{Z}_i$.

Step 1: Sample underlying variables $\mathbf{Z}_i^*$ from their conditional distribution given $\mathbf{Y}_i^{\text{obs}}$, $\mathbf{W}_i$, and $\mathbf{Z}_i^{\text{obs}}$.

Step 2: Sample $\tilde{\mathbf{Z}}_i^*$ given $\mathbf{Z}_i^*$, where $(\mathbf{Z}_i^*, \tilde{\mathbf{Z}}_i^*)$ jointly follows a multivariate normal distribution with mean zero and covariance matrix

$$\mathbf{G} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathbf{S} \\ \boldsymbol{\Sigma} - \mathbf{S} & \boldsymbol{\Sigma} \end{pmatrix},$$

where $\boldsymbol{\Sigma}$ is the correlation matrix in the Gaussian copula model, and $\mathbf{S}$ is a diagonal matrix specified in such a way that the joint covariance matrix $\mathbf{G}$ is positive semidefinite. The construction of $\mathbf{S}$ is based on the minimise the reconstructability (MVR) procedure (Spector & Janson, 2022).

Step 3: Obtain $\tilde{\mathbf{Z}}_i$ from $\tilde{\mathbf{Z}}_i^*$, where $\tilde{Z}_{ij} = F_j(\tilde{Z}_{ij}^*)$ for each $j = 1, \ldots, p$.

Step 4: Sample $\tilde{\mathbf{W}}_{ij}$ from the conditional distribution $q_j(\cdot \mid \tilde{Z}_{ij}; \Lambda_j)$ for each $j = 1, \ldots, p_1$.

Output: Knockoff copy $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$, where $\tilde{\mathbf{Z}}_i^{\text{obs}} = \{\tilde{Z}_{ij}^{obs} : j \in \mathcal{A}_i\}$.

Proposition 1 The output $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$ from Algorithm 1 satisfies Definition 2.

The proof of this proposition is given in the online supplementary material. Figure 2 below gives the path diagram for the generation of knockoff copies. We provide several remarks on the algorithm. This algorithm allows for mixed types of predictors under a Gaussian copula model, which extends the multivariate Gaussian model for knockoff construction considered in Barber and Candès (2015) and Candès et al. (2018). When all the predictors are continuous, the Gaussian copula model degenerates to the multivariate Gaussian model. In that case, and if there is no missing data, then Algorithm 1 coincides with the knockoff construction method in Candès et al. (2018), except that Candès et al. (2018) uses the mean absolute correlation (MAC) procedure to construct the $\mathbf{S}$ matrix.

When $\mathbf{W}_i$ or $\mathbf{Z}_i^{\text{obs}}$ contains binary or ordinal variables, the sampling of $\mathbf{Z}_i^*$ is not straightforward. However, we can obtain approximate samples via Gibbs sampling. Thanks to the underlying multivariate normality assumption, each step of the Gibbs sampler only involves sampling from univariate normal or truncated normal distributions. Details of the Gibbs sampler are given in the online supplementary material. We compute the diagonal matrix $\mathbf{S}$ in Step 2 of the algorithm using the MVR procedure (Spector & Janson, 2022), which tends to be more powerful than the MAC procedure adopted in Barber and Candès (2015) and Candès et al. (2018).

4.3 Variable selection via derandomised knockoffs

We now describe a knockoff procedure for variable selection with a controlled PFER. Suppose that knockoff copies $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$, $i = 1, \ldots, N$, have been obtained using Algorithm 1. For ease of exposition, we denote $\mathbf{Z}^{\text{obs}} = \{\mathbf{Z}_i^{\text{obs}}\}_{i=1}^N$, $\tilde{\mathbf{Z}}^{\text{obs}} = \{\tilde{\mathbf{Z}}_i^{\text{obs}}\}_{i=1}^N$, $\mathbf{W} = \{\mathbf{W}_i\}_{i=1}^N$, $\tilde{\mathbf{W}} = \{\tilde{\mathbf{W}}_i\}_{i=1}^N$, and $\mathbf{Y}^{\text{obs}} = \{\mathbf{Y}_i^{\text{obs}}\}_{i=1}^N$. We define a knockoff statistic that measures the importance of each predictor.

Definition 3 (Knockoff statistic). Consider a statistic T_j taking the form $T_j = t_j((\mathbf{W}, \mathbf{Z}^{\text{obs}}), (\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}}), \mathbf{Y}^{\text{obs}})$ for some function t_j, where $(\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}})$ are knockoffs satisfying Definition 2. This statistic is called a knockoff statistic for the jth predictor if it satisfies the *flip-sign property*; that is for any

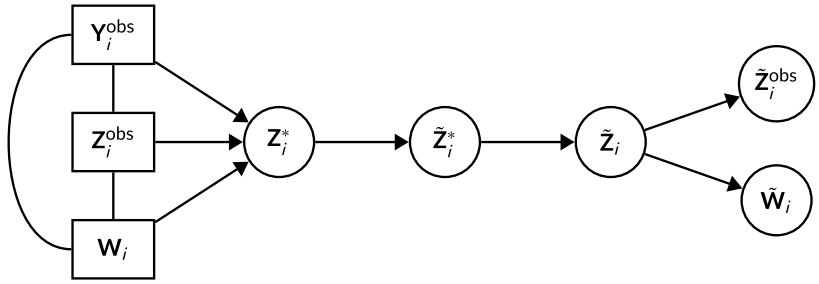

Figure 2. Path diagram for constructing $(\tilde{\mathbf{W}}_i, \tilde{\mathbf{Z}}_i^{\text{obs}})$.

subset $\mathcal{S} \subset \{1, \ldots, p\}$,

$$t_j\left(\{(\mathbf{W}, \mathbf{Z}^{\text{obs}}), (\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}})\}_{\text{swap}(\mathcal{S})}, \mathbf{Y}^{\text{obs}}\right) = \begin{cases} t_j\left((\mathbf{W}, \mathbf{Z}^{\text{obs}}), (\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}}), \mathbf{Y}^{\text{obs}}\right), & j \notin \mathcal{S}, \\ -t_j\left((\mathbf{W}, \mathbf{Z}^{\text{obs}}), (\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}}), \mathbf{Y}^{\text{obs}}\right), & j \in \mathcal{S}, \end{cases}$$

where $\{(\mathbf{W}, \mathbf{Z}^{\text{obs}}), (\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}})\}_{\text{swap}(\mathcal{S})}$ is obtained by swapping

1. the entries of $\mathbf{W}_{ij}$ and $\tilde{\mathbf{W}}_{ij}$ for each $j \in \mathcal{S} \cap \{1, 2, \ldots, p_1\}$, $i = 1, \ldots, N$;
2. the entries Z_{ij} and $\tilde{Z}_{ij}$ for each $j \in \mathcal{S} \cap \mathcal{A}_i$, $i = 1, \ldots, N$.

The flip-sign property in Definition 3 is key to guaranteeing valid statistical inference from finite samples. However, to achieve a good power, T_j should also provide evidence regarding whether $\|\boldsymbol{\beta}_j\| = 0$. See Section 3 of Candès et al. (2018) for a generic method of constructing T_j and specific examples. In this study, we will focus on knockoff statistics constructed based on the likelihood function. More specifically, we incorporate the knockoff variables into the general latent variable model defined in Section 3. That is, the measurement model remains the same, while the structural model becomes

$$\theta_i | \mathbf{Z}_i, \tilde{\mathbf{Z}}_i \sim N(\beta_0 + \boldsymbol{\beta}_1^\top g_1(Z_{i1}) + \cdots + \boldsymbol{\beta}_p^\top g_p(Z_{ip}) + \boldsymbol{\gamma}_1^\top g_1(\tilde{Z}_{i1}) + \cdots + \boldsymbol{\gamma}_p^\top g_p(\tilde{Z}_{ip}), \sigma^2),$$

where $\mathbf{Z}_i$ and $\tilde{\mathbf{Z}}_i$ are defined in Definition 2. Since $\tilde{\mathbf{Z}}_i^*$ and $\mathbf{Y}_i^{\text{obs}}$ are conditionally independent given $\mathbf{Z}_i^*$, the true value of $\boldsymbol{\gamma}_j$ is 0, $j = 1, \ldots, p$, though these parameters will be estimated when constructing the knockoff statistics. The general latent variable model corresponding to this is depicted by the path diagram shown in Figure 3.

Suppose that the values of Ξ, Λ_1,..., Λ_{p_1}, β_0, σ^2, and Δ are known. The likelihood function for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ under this extended latent regression model takes the form

$$\tilde{l}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \log\left\{\int \cdots \int \left[\left(\prod_{j \notin \mathcal{A}_i} \mathrm{d}z_{ij}\right)\left(\prod_{j \notin \mathcal{A}_i} \mathrm{d}\tilde{z}_{ij}\right) f(\mathbf{z}_i, \tilde{\mathbf{z}}_i | \Xi) \times \left(\prod_{j=1}^{p_1} q_j(\mathbf{w}_{ij} | z_{ij}; \Lambda_j)\right)\left(\prod_{j=1}^{p_1} q_j(\tilde{\mathbf{w}}_{ij} | \tilde{z}_{ij}; \Lambda_j)\right) f_i(\mathbf{y}_i^{\text{obs}} | \mathbf{z}_i, \tilde{\mathbf{z}}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \beta_0, \sigma^2, \Delta)\right]\right\}. \quad (9)$$

Here, $f_i(\mathbf{y}_i^{\text{obs}} | \mathbf{z}_i, \tilde{\mathbf{z}}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \Delta)$ is the density function of the conditional distribution of $\mathbf{Y}_i^{\text{obs}}$ given

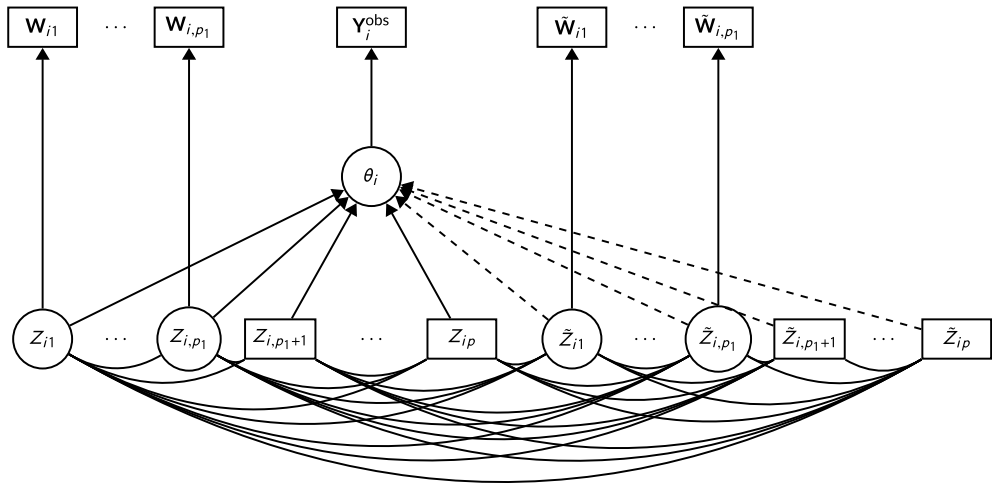

Figure 3. Path diagram for the general latent variable model involving knockoff variables. The interpretation is similar to that of Figure 1. The directed edges from the knockoff variables $\tilde{Z}_{ij}$s to θ_i are drawn with dashed lines, as the true values of the corresponding coefficients are zero.

$\mathbf{Z}_i = \mathbf{z}_i$ and $\tilde{\mathbf{Z}}_i = \tilde{\mathbf{z}}_i$; that is,

$$f(\mathbf{y}_i^{\text{obs}}|\mathbf{z}_i, \tilde{\mathbf{z}}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \beta_0, \sigma^2, \Delta) = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \int \cdots \int \left[\mathrm{d}\theta_i \left(\prod_{j \notin \mathcal{B}_i} \mathrm{d}y_{ij} \right) \right. \right.$$
$$\left. \left. \times h(\mathbf{y}_i|\theta_i; \Delta) \exp\left(- \frac{(\theta_i - (\beta_0 + \boldsymbol{\beta}_1^\top g_1(z_{i1}) + \cdots + \boldsymbol{\gamma}_1^\top g_1(\tilde{z}_{i1}) + \cdots + \boldsymbol{\gamma}_p^\top g_p(\tilde{z}_{ip}))^2}{2\sigma^2} \right) \right] \right\}.$$

In addition, $f(\mathbf{z}_i, \tilde{\mathbf{z}}_i|\Xi)$ denotes the density function of the Gaussian copula model for $(\mathbf{Z}_i, \tilde{\mathbf{Z}}_i)$, noting that this density function is completely determined by the parameters Ξ of the Gaussian copula model for $\mathbf{Z}_i$; see the online supplementary material for the specific form of $f(\mathbf{z}_i, \tilde{\mathbf{z}}_i|\Xi)$.

A knockoff statistic T_j is constructed based on $\tilde{l}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Specifically, consider the maximum likelihood estimator based on $\tilde{l}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})$

$$(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) = \arg\max_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \ \tilde{l}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}). \tag{10}$$

Then, a knockoff statistic can be constructed as

$$T_j = \text{sign}(\|\tilde{\boldsymbol{\beta}}_j^\dagger\| - \|\tilde{\boldsymbol{\gamma}}_j^\dagger\|) \max\left\{ \|\tilde{\boldsymbol{\beta}}_j^\dagger\|/\sqrt{p_j}, \|\tilde{\boldsymbol{\gamma}}_j^\dagger\|/\sqrt{p_j} \right\}, \tag{11}$$

where p_j is the dimension of $\boldsymbol{\beta}_j$ (or equivalently that of $\boldsymbol{\gamma}_j$), and $\tilde{\boldsymbol{\beta}}_j^\dagger = \text{Cov}(g_j(Z_{ij}))^{\frac{1}{2}}\tilde{\boldsymbol{\beta}}_j$ and $\tilde{\boldsymbol{\gamma}}_j^\dagger = \text{Cov}(g_j(Z_{ij}))^{\frac{1}{2}}\tilde{\boldsymbol{\gamma}}_j$ are standarised coefficients.

Proposition 2 Assume that the values of Ξ, Λ_1,..., Λ_{p_1}, β_0, σ^2, and Δ are known and the knockoffs satisfy Definition 2. Then T_j given by (11) satisfies Definition 3.

The proof of this proposition is given in the online supplementary material. Similar to the estimation of the latent regression model without knockoffs, the optimisation problem (10) can be solved using a stochastic EM algorithm. We remark that the statistic (11) is a special case of the

Lasso coefficient-difference statistic given in Candès et al. (2018) when the Lasso penalty is set to zero. Since the sample size N is often much larger than p in ILSA applications, this likelihood-based knockoff statistic performs well in our simulation study and real data analysis. For higher-dimensional settings, a Lasso coefficient-difference statistic may be preferred; see Candès et al. (2018).

We now adapt the derandomised knockoff method (Ren et al., 2023) to the current problem. This method achieves PFER control by aggregating the results from multiple runs of a baseline algorithm proposed in Janson and Su (2016). This baseline algorithm is summarised in Algorithm 2 below.

Algorithm 2 (Baseline algorithm for PFER control Janson & Su, 2016)

Input: Observed data $\mathbf{Y}^{\text{obs}}$, $\mathbf{W}$, and $\mathbf{Z}^{\text{obs}}$, a PFER level $v \in \mathbb{Z}_+$, the true model parameters Ξ of the Gaussian copula model, the true parameters $\Lambda_1,\ldots,\Lambda_{p_1}$ in the measurement models for Z_{ij}, and the true parameters $\boldsymbol{\beta}$, β_0, σ^2, Δ in the conditional model of $\mathbf{Y}_i$ given $\mathbf{Z}_i$.

Step 1: Generate knockoffs $(\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}^{\text{obs}})$ using Algorithm 1.

Step 2: Compute a set of knockoff statistics $T_1, \ldots, T_p$ using equations (10) and (11).

Step 3: Compute the threshold $\tau = \inf\{t > 0 : 1 + |\{j : T_j < -t\}| = v\}$. We let $\tau = -\infty$ if the set on the right-hand side is an empty set.

Output: $\hat{\mathcal{S}} = \{j : T_j > \tau\}$.

Proposition 3 $\hat{\mathcal{S}}$ given by Algorithm 2 satisfies $\mathbb{E}|\hat{\mathcal{S}} \setminus \mathcal{S}^*| \leq v$, i.e. the PFER can be controlled at level v.

Algorithm 3 (Derandomised knockoffs Ren et al., 2023)

Input: Observed data $\mathbf{Y}^{\text{obs}}$, $\mathbf{W}$, and $\mathbf{Z}^{\text{obs}}$, the number of runs M of the baseline algorithm, a selection threshold η, a PFER level $v \in \mathbb{Z}_+$, the true model parameters Ξ of the Gaussian copula model, the true parameters Λ_1, …, Λ_{p_1} in the measurement models for Z_{ij}, and the true parameters $\boldsymbol{\beta}$, β_0, σ^2, Δ in the conditional model of $\mathbf{Y}_i$ given $\mathbf{Z}_i$.

Step 1: For each $m = 1, \ldots, M$, run Algorithm 2 independently and obtain the selection set $\hat{\mathcal{S}}^{(m)}$.

Step 2: For each $j = 1, \ldots, p$, compute the selection frequency $\Pi_j = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(j \in \hat{\mathcal{S}}^{(m)})$.

Output: $\hat{\mathcal{S}} = \{j \in \{1, \ldots, p\} : \Pi_j \geq \eta\}$.

Following the theoretical result in Ren et al. (2023) when the threshold η is chosen properly, Algorithm 3 guarantees to control PFER at level v. We provide a simplified version of this result in Proposition 4 below.

Proposition 4 If for any $\eta \in (0, 1)$, the condition $\mathbb{P}(\Pi_j \geq \eta) \leq \mathbb{E}[\Pi_j]$ holds for every $j \notin \mathcal{S}^*$, then $\hat{\mathcal{S}}$ given by Algorithm 3 satisfies $\mathbb{E}|\hat{\mathcal{S}} \setminus \mathcal{S}^*| \leq v$, i.e. the PFER can be controlled at level v. In particular, assuming that the probability mass function of Π_j is monotonically non-increasing for each $j \notin \mathcal{S}^*$, $\mathbb{P}(\Pi_j \geq \eta) \leq \mathbb{E}[\Pi_j]$ holds for $M = 31$ and $\eta = 1/2$.

While noting that other choices are possible, we set $M = 31$ and $\eta = 1/2$, which is also the default choice in Ren et al. (2023). We also note that the statistics Π_j, $j = 1, \ldots, p$, rank the importance of the predictors. The predictors with $\Pi_j \geq \eta$ are selected as the non-null variables.

4.4 A three-step procedure when model parameters are unknown and its robustness

The knockoff procedure described previously requires the true joint model for $\mathbf{Y}_i$, $\mathbf{W}_i$, and $\mathbf{Z}_i$, which is infeasible in practice. When the true model is known, the variable selection problem becomes trivial since the null and non-null variables can be directly identified from the true model. In

practice, we first estimate the model parameters and then conduct variable selection based on the estimated model. This procedure involves three steps. First, estimate the parameters Ξ in the Gaussian copula model as well as the parameters $\Lambda_1, \ldots, \Lambda_{p_1}$ in the measurement models for Z_{ij}. This is done by the maximum likelihood estimator (6). Second, estimate the parameters $\boldsymbol{\beta}$, β_0, σ^2, and Δ based on the log-likelihood (7), where the estimated Gaussian copula model $\hat{\Xi}$ and the estimated measurement models $\hat{\Lambda}_1, \ldots, \hat{\Lambda}_{p_1}$ are plugged in. Third, select variables by plugging the estimated parameters $\hat{\Xi}$, $\hat{\Lambda}_j$s, $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_0$, $\hat{\sigma}^2$, and $\hat{\Delta}$ into Algorithm 2 or 3.

Empirically, simulation results in Section 5 show that PFER is well controlled when we apply the above three-step procedure. Theoretically, by plugging into the estimated model rather than the true model, the PFER can no longer be exactly controlled as described in Propositions 3 and 4. Following a similar proof strategy as in Barber et al. (2020), we show that this procedure is robust, in the sense that the resulting PFER is controlled near v if the plug-in model is sufficiently accurate. Note that Barber et al. (2020) only consider the robustness of model-X knockoffs for controlling false discovery rate and does not cover PFER.

More precisely, we use $\mathbb{P}$ and $\mathbb{Q}$ to denote the true and plug-in models, respectively. Consider a pair of i and j, satisfying $j \in (\{1, \ldots, p_1\} \cup \mathcal{A}_i)$. We consider $\mathbf{X}_i$ in Definition 1. We further let $\mathbf{X}_{i,-j}^{\text{obs}} = \{\mathbf{X}_{ik} : k \in (\{1, \ldots, p_1\} \cup \mathcal{A}_i) \backslash \{j\}\}$. We also define $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{X}}_i^{\text{obs}}$ similar to $\mathbf{X}_i$ and $\mathbf{X}_i^{\text{obs}}$, respectively, but with $\mathbf{W}_{ij}$ replaced by $\tilde{\mathbf{W}}_{ij}$ and with Z_{ij} replaced by $\tilde{Z}_{ij}$. Let $\mathbb{P}_{ij}(\mathbf{x}_{ij}|\mathbf{x}_{i,-j}^{\text{obs}}, \mathbf{y}_i^{\text{obs}})$ denote the conditional density function of $\mathbf{X}_{ij}$ given $\mathbf{X}_{i,-j}^{\text{obs}} = \mathbf{x}_{i,-j}^{\text{obs}}$ and $\mathbf{Y}_i^{\text{obs}} = \mathbf{y}_i^{\text{obs}}$ under the true model $\mathbb{P}$. Let $\mathbb{Q}_{ij}(\tilde{\mathbf{x}}_{i,-j}^{\text{obs}}, \tilde{\mathbf{x}}_{ij}|\mathbf{x}_{i,-j}^{\text{obs}}, \mathbf{x}_{ij}, \mathbf{y}_i^{\text{obs}})$ denote the conditional density function of $(\tilde{\mathbf{X}}_{i,-j}^{\text{obs}}, \tilde{\mathbf{X}}_{ij})$ given $\mathbf{X}_{i,-j}^{\text{obs}} = \mathbf{x}_{i,-j}^{\text{obs}}$, $\mathbf{X}_{ij} = \mathbf{x}_{ij}$, and $\mathbf{Y}_i^{\text{obs}} = \mathbf{y}_i^{\text{obs}}$ under the plug-in model $\mathbb{Q}$. We define

$$\hat{\text{KL}}_j = \sum_{i:j \in (\{1,\ldots,p_1\} \cup \mathcal{A}_i)} \log\left(\frac{\mathbb{P}_{ij}(\mathbf{X}_{ij}|\mathbf{X}_{i,-j}^{\text{obs}}, \mathbf{Y}_i^{\text{obs}}) \cdot \mathbb{Q}_{ij}(\tilde{\mathbf{X}}_{i,-j}^{\text{obs}}, \tilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j}^{\text{obs}}, \mathbf{X}_{ij}, \mathbf{Y}_i^{\text{obs}})}{\mathbb{P}_{ij}(\tilde{\mathbf{X}}_{ij}|\mathbf{X}_{i,-j}^{\text{obs}}, \mathbf{Y}_i^{\text{obs}}) \cdot \mathbb{Q}_{ij}(\tilde{\mathbf{X}}_{i,-j}^{\text{obs}}, \mathbf{X}_{ij}|\mathbf{X}_{i,-j}^{\text{obs}}, \tilde{\mathbf{X}}_{ij}, \mathbf{Y}_i^{\text{obs}})}\right).$$

Here, $\{i : j \in (\{1, \ldots, p_1\} \cup \mathcal{A}_i)\} = \{1, \ldots, p\}$ if $j \in \{1, \ldots, p_1\}$, and $\{i : j \in (\{1, \ldots, p_1\} \cup \mathcal{A}_i)\} = \{i : j \in \mathcal{A}_i\}$ otherwise. Note that the numerator inside of the logarithm corresponds to the true data generation mechanism for $(\mathbf{X}_{ij}, \tilde{\mathbf{X}}_i^{\text{obs}})$, and the denominator corresponds to that when switching the roles of $\mathbf{X}_{ij}$ and $\tilde{\mathbf{X}}_{ij}$. $\hat{\text{KL}}_j$ can be viewed as an observed Kullback–Leibler (KL) divergence that measures the discrepancy between the true model $\mathbb{P}$ and its approximation $\mathbb{Q}$, with $\hat{\text{KL}}_j = 0$ when $\mathbb{Q} = \mathbb{P}$. We remark that this definition of $\hat{\text{KL}}_j$ is consistent with that in Barber et al. (2020). However, the $\hat{\text{KL}}_j$ in Barber et al. (2020) can be further simplified with a pairwise exchangeable property of their procedure under a model-X knockoff setting without missing data, while this pairwise exchangeable property does not always hold for the current procedure due to the involvement of $\mathbf{Y}^{\text{obs}}$ and thus, the current $\hat{\text{KL}}_j$ cannot be further simplified.

Theorem 1 Under the definitions above, for any $\epsilon \geq 0$, consider the null variables for which $\hat{\text{KL}}_j \leq \epsilon$. If we use a modified Algorithm 2 that generates knockoffs under the plug-in model $\mathbb{Q}$ which is assumed to be independent of data, then the expected number of rejections that correspond to such nulls obeys $\mathbb{E}|\{j : j \in \hat{\mathcal{S}} \setminus \mathcal{S}^* \text{ and } \hat{\text{KL}}_j \leq \epsilon\}| \leq v\mathrm{e}^{\epsilon}$. In particular, $\hat{\text{KL}}_j = 0$, when $\mathbb{Q} = \mathbb{P}$.

When $\mathbb{P} = \mathbb{Q}$, we can set $\epsilon = 0$, and thus, Theorem 1 implies Proposition 3. This property of robustness carries over to the derandomised procedure. We define $\Pi_j^{\dagger} = (\sum_{m=1}^{M} \mathbb{I}(j \in \hat{\mathcal{S}}^{(m)} \text{ and } \hat{\text{KL}}_j^{(m)} \leq \epsilon))/M$, where $\hat{\mathcal{S}}^{(m)}$ is the selection in the mth run of modified Algorithm 3 that generates knockoffs under the plug-in model $\mathbb{Q}$, and $\hat{\text{KL}}_j^{(m)}$ is the corresponding observed KL divergence based on the knockoffs from the mth run.

Theorem 2 Under the definitions above, for any $\epsilon \geq 0$, consider the null variables for which $\hat{\text{KL}}_j^{(m)} \leq \epsilon$ for all $m = 1, \ldots, M$. We use a modified Algorithm 3 where

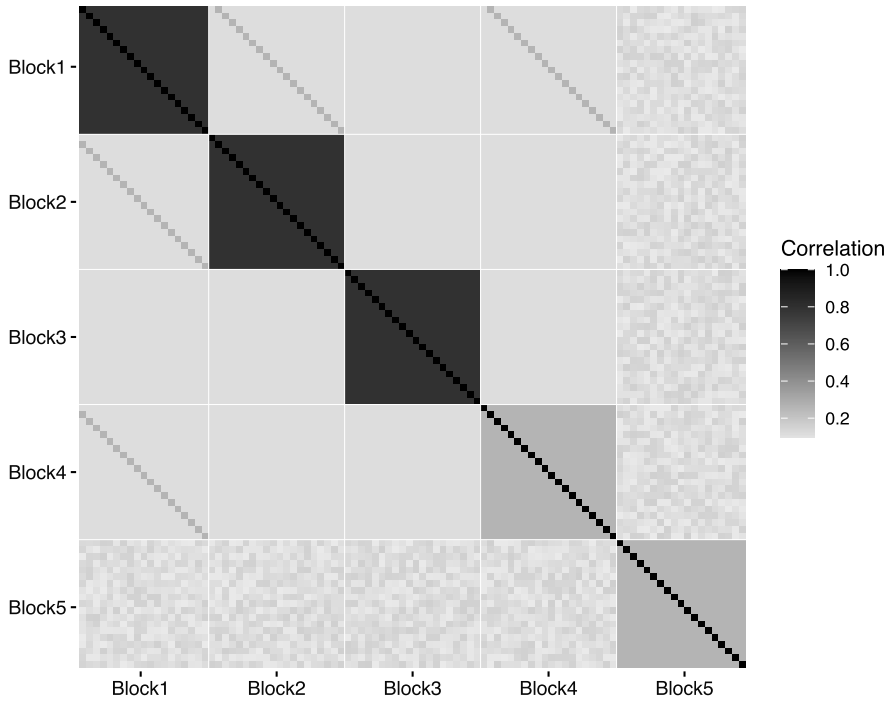

Figure 4. Heatmap of the designed correlation matrix in the simulation study.

knockoffs are generated under the plug-in model $\mathbb{Q}$ which is assumed to be independent of data, and obtain selections $\hat{\mathcal{S}}$. If the condition $\mathbb{P}(\Pi_j^\dagger \geq \eta) \leq \mathbb{E}[\Pi_j^\dagger]$ holds for every $j \notin \mathcal{S}^*$, then $\mathbb{E}|\{j : j \in \hat{\mathcal{S}} \setminus \mathcal{S}^*$ and $\hat{\mathrm{KL}}_j^{(m)} \leq \epsilon, m = 1, \ldots, M\}| \leq ve^\epsilon$.

If the probability mass function of $\Pi_j^\dagger$ is monotonically non-increasing for each $j \notin \mathcal{S}^*$, $\mathbb{P}(\Pi_j^\dagger \geq \eta) \leq \mathbb{E}[\Pi_j^\dagger]$ holds for $M = 31$ and $\eta = 1/2$.

5 Simulation study

In this section, we conduct a simulation study to evaluate the performance of the proposed knockoff method. We check if the PFER can be controlled at the targeted level when the three-step procedure described in Section 4.4 is applied. The power of variable selection will also be assessed.

We set $p = 100$, $J = 60$, and consider $N \in \{1{,}000, 2{,}000, 4{,}000\}$ for comparing power under different sample sizes. It leads to three settings. For each setting, we generate 100 independent replications. The data are generated as follows. We divide the predictors into five blocks, each containing 10 continuous variables and 10 binary variables. Ordinal variables or unobservable variables are not included in this study for simplicity. In Section F.3 of the online supplementary material, we present a simulation study that includes unobservable variables.

We consider the following design for the correlation matrix $\mathbf{\Sigma}$ of the underlying variables $\mathbf{Z}_i^*$, which is similar to the one used in Grund et al. (2021) that concerns analysing missing data in ILSAs. This correlation matrix mimics the correlation structure in ILSA data. (a) Within block 1, the correlation between every pair of variables is 0.6. (b) Within block 2, the correlation between every pair of variables is 0.6. For the 10-by-10 submatrix recording the correlations between variables in blocks 1 and 2, the diagonal entries are set to be 0.3, and the off-diagonal entries are set to be 0.15. (c) Within block 3, the correlation between every pair of variables is 0.6. The variables in block 3 have a correlation of 0.15 with each variable in blocks 1 and 2. (d) Within block 4, the correlation between every pair of variables is 0.3. For the 10-by-30 submatrix recording the correlations between variables in block 4 and those in blocks 1 to 3, all the entries take a value of 0.15, except that the diagonal entries of the 10-by-10 submatrix

Table 1. Simulation results

			$v = 1$	$v = 2$	$v = 3$	$v = 4$	$v = 5$
N = 1,000	PFER	Baseline	0.68	1.72	2.78	4.01	5.04
		DRM	0.01	0.10	0.33	0.55	0.83
	TPR	Baseline	54.0%	65.5%	70.9%	74.2%	77.2%
		DRM	59.9%	69.1%	74.6%	78.3%	80.7%
N = 2,000	PFER	Baseline	1.28	2.52	3.63	4.98	5.85
		DRM	0.06	0.33	0.69	1.15	1.60
	TPR	Baseline	81.7%	88.9%	91.1%	93.5%	94.0%
		DRM	83.8%	90.3%	93.1%	95.2%	95.8%
N = 4,000	PFER	Baseline	0.76	1.84	3.06	4.17	5.41
		DRM	0.14	0.53	0.95	1.51	1.95
	TPR	Baseline	95.3%	98.9%	99.2%	99.4%	99.5%
		DRM	97.7%	99.4%	99.6%	99.6%	99.7%

Note. Here, 'Baseline' refers to the baseline algorithm, Algorithm 2, and 'DRM' refers to derandomised knockoffs, Algorithm 3. v refers to the nominal PFER level.

corresponding to blocks 4 and 1 are set to 0.3. (e) Within block 5, the correlation between every pair of variables is 0.3. For the 10-by-40 submatrix recording the correlations between variables in block 5 and those in blocks 1 to 4, the entries are generated independently from a uniform distribution over the interval [0.1, 0.2]. The same correlation matrix is used in all 100 replications. The heat map of this correlation matrix is given in Figure 4. This correlation matrix has a maximal eigenvalue of 22.73 and a minimal eigenvalue of 0.11.

The rest of the Gaussian copula model is set as follows. For continuous variables, we set $c_j = 0$ and $d_j = 1$. For the binary variables, we set their threshold parameters c_{j1} to take one of the values in $(-1.2, -0.3, 0, 0.3, 1.2)$ iteratively (i.e. $c_{11,1} = -1.2$, $c_{12,1} = -0.3$ and so on). Regarding the parameters in the structural model, we set the intercept $\beta_0 = 0$, $\beta_j = 0.5$ for $j = 1, 22, 43, 64, 85$, -0.5 for $j = 11, 32, 53, 74, 95$, and 0 for the rest of the variables. Under this setting, the non-zero coefficients are distributed uniformly among the variables. We further set $\sigma^2 = 1$ for the residual variance.

Data missingness is generated following the SMAR condition. For each observation i, we generate a random variable R_i from a categorical distribution with support $\{1, 2, \ldots, 5\}$, satisfying $P(R_i = k) = 0.2$, for all $k = 1, \ldots, 5$. The data missingness is determined by R_i and the non-null variables. Let $\mathcal{S}_k^*$ denote the set of non-null variables in the kth block. For observation i, when $R_i = k$, we let all the variables in $\mathcal{S}_k^*$ be observed. For each of the rest of the variables j, its probability of being missing is given by $(1 + \exp(1 - (\sum_{j' \in \mathcal{S}_k^*} Z_{ij'})/2))^{-1}$. Under this setting, around 33% of the entries of the data matrix for predictors are missing.

Finally, we generate the parameters in the measurement model with only dichotomous items. We sample a_j's from a uniform distribution $U[0.5, 1.5]$, and b_j's from uniform distribution $U[-2, 0]$, where the range of these distributions is chosen to guarantee that $a_j\theta_i + b_j$ to be in a suitable range. When generating the responses, a matrix sampling design is adopted. Here, all the items are divided into three equal-sized blocks. Each observation is randomly assigned one of the three blocks, and the responses to the rest of the two blocks are missing completely at random.

We apply the three-step procedure described in Section 4.4, including both the baseline procedure based on Algorithm 2 and the derandomised procedure based on Algorithm 3. In this simulation study, we assume that all item parameters are known and fix them to their true values in both (7) and (9). In addition, we include an l_2-penalty on $\boldsymbol{\beta}$ in equation (7), as well as an l_2-penalty on $(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ in equation (9) during the estimation to mitigate the problem of overfitting. More details are given in the online supplementary material. Different target levels are considered, including $v \in \{1, 2, \ldots, 5\}$. Our results are given in Table 1. Two performance metrics are reported,

Table 2. Results from applying Algorithm 3 to PISA data

Name	Type	Description	Estimate	SE
$v = 1$				
ANXTES	C	Personality: test anxiety.	−0.0542	0.0097
BELONG	C	Subjective well-being: sense of belonging to school.	−0.0454	0.0108
DISCLI	C	Disciplinary climate in science classes.	0.0657	0.0104
CPSVAL	C	Collaboration and teamwork dispositions: value cooperation.	−0.0862	0.0110
EBSCIT	C	Enquiry-based science teaching and learning practices.	−0.0561	0.0117
EISCED	O	ISCED (International Standard Classification of Education) level student expects to complete. (0/1/2 = [level 2 or 3A]/[level 4 or 5B]/[level 5A or 6])	0.1733 0.0615	0.0350 0.0299
ENVAWA	C	Environmental awareness.	0.0640	0.0109
ENVOPT	C	Environmental optimism.	−0.0886	0.0090
EPIST	C	Epistemological beliefs.	0.0887	0.0101
GENDER	B	Student's gender. (0/1 = female/male)	0.1884	0.0211
JOYSCI	C	Enjoyment of science.	0.0889	0.0124
OUT.JOB	B	Whether work for pay outside the school. (0/1 = no/yes)	0.2076	0.0290
OUT.PAR	B	Whether talk to parents outside the school. (0/1 = no/yes)	−0.1373	0.0307
OUT.SPO	B	Whether exercise or do a sport outside the school. (0/1 = no/yes)	0.1966	0.0223
OUT.STU	B	Whether study for school or homework outside the school. (0/1 = no/yes)	0.1188	0.0202
PERFEE	C	Perceived feedback.	−0.1373	0.0125
REPEAT	B	Whether the student has ever repeated a grade. (0/1 = no/yes)	−0.2391	0.0350
SCI.CHE	B	Whether attended chemistry courses in this or last school year. (0/1 = no/yes)	0.1109	0.0207
TMINS	C	Learning time in class per week (minutes).	0.0949	0.0098
UNFAIR	C	Teacher unfairness.	−0.0542	0.0108
LANGAH	B	Whether language at home different from the test language. (0/1 = no/yes)	−0.1022	0.0280
TDSCIT	C	Teacher-directed science instruction.	0.0504	0.0112
EISEIO	C	Student's expected International Socio-economic Index of occupational status.	0.0448	0.0109
OUTHOU	C	Out-of-school study time per week (hours).	−0.0448	0.0106
COOPER	C	Collaboration and teamwork dispositions: enjoy cooperation.	0.0537	0.0113
INSTSC	C	Instrumental motivation.	−0.0388	0.0097
SCIEEF	C	Science self-efficacy.	0.0408	0.0105
FISEIO	C	ISEI (International Socio-economic Index) of occupational status of father.	0.0436	0.0122
MISEIO	C	ISEI (International Socio-economic Index) of occupational status of mother.	0.0373	0.0113
CULTPO	C	Cultural possessions at home.	0.0388	0.0114
CHONUM	O	Whether can choose the number of school science course(s) they study. (0/1/2 = no, not at all/ yes, to a certain degree/yes, can choose freely)	0.0995 −0.0288	0.0230 0.0361
SCI.PHY	B	Whether attended physics courses in this or last school year. (0/1 = no/yes)	−0.0718	0.0201
SKIDAY	O	The frequency student skipped a whole school day in the last two full weeks of school. (0/1/2 = [none]/[one or two times]/[three or more times])	−0.0398 −0.1312	0.0198 0.0452

(*continued*)

Table 2. Continued

Name	Type	Description	Estimate	SE
CHODIF	O	Whether can choose the level of difficulty for school science course(s). (0/1/2 = no, not at all/ yes, to a certain degree/yes, can choose freely)	0.0677 0.0459	0.0227 0.0305
DAYPEC	O	Averaged days that student attends physical education classes each week. (0/1/2/3 = [0]/[1 or 2]/[3 or 4]/[5 or more])	−0.0561 0.0150 −0.0740	0.0361 0.0409 0.0278
ADINST	C	Adaption of instruction.	0.0243	0.0140
$v = 2$				
SCI.EAR	B	Whether attended earth and space courses in this or last school year. (0/1 = no/yes)	−0.0549	0.0221
OUT.NET	B	Whether use Internet outside the school. (0/1 = no/yes)	0.0663	0.0254
ARRLAT	O	The frequency of arriving late for school in the last two full weeks of school. (0/1/2 = [none]/[one or two times]/[three or more times])	−0.0731 −0.0118	0.0211 0.0349
GRADE	O	Student's grade. (0/1/2 = lower than modal grade/not lower than modal grade/higher than modal grade.)	0.1125 −0.0090	0.0368 0.0242
HEDRES	C	Home educational resources.	−0.0216	0.0108
INTBRS	C	Interest in broad science topics.	0.0232	0.0120
CHOCOU	O	Whether can choose the school science course(s) they study. (0/1/2 = no, not at all/yes, to a certain degree/yes, can choose freely)	0.0516 −0.00612	0.0223 0.0291
OUT.VED	B	Whether watch TV/DVD/Video outside the school. (0/1 = no/yes)	0.0400	0.0204
DUECEC	O	Duration in early childhood education and care of students. (0/1/2/3 = [less than two years]/[at least two but less than three years]/[at least three but less than four years]/[at least four years])	0.0496 −0.0313 −0.0797	0.0268 0.0313 0.0416
$v = 3$				
SCI.GEN	B	Whether attended general, integrated, or comprehensive science courses in this or last school year. (0/1 = no/yes)	0.0372	0.0210
EMOSUP	C	Parents' emotional support.	−0.0180	0.0114
DAYMPA	O	Number of days with moderate physical activities for a total of at least 60 minutes per week. (0/1/2/3/4/5/6/7 = 0/1/2/3/4/5/6/7)	0.0070 0.0453 0.0457 −0.0218 −0.0048 0.0555 0.0006	0.0446 0.0457 0.0405 0.0370 0.0344 0.0363 0.0340
Unselected				
OUT.MEA	B	Whether have meals before school or after school. (0/1 = no/yes)	0.0373	0.0195
OUT.GAM	B	Whether play video-games outside the school. (0/1 = no/yes)	0.0222	0.0230
TEASUP	C	Teacher support in science classes of students' choice.	0.0112	0.0126
FISCED	O	Father's education in ISCED level. (0/1/2/3/4 = [none or ISCED 1]/[ISCED 2]/[ISCED 3B or 3C]/[ISCED 3A or 4]/[ISCED 5B]/[ISCED 5A or ISCED 6])	0.0057 0.0051 −0.0145 0.0445	0.0433 0.0343 0.0304 0.0360
MISCED	O	Mother's education in ISCED level. (0/1/2/3/4 = [none or ISCED 1]/[ISCED 2]/[ISCED 3B or 3C]/[ISCED 3A or 4]/[ISCED 5B]/[ISCED 5A or ISCED 6])	−0.0376 0.0314 −0.0418 0.0382	0.0520 0.0355 0.0270 0.0280
MOTIVA	C	Achievement motivation.	0.0082	0.0100
OUT.FRI	B	Whether meet or talk to friends on the phone outside the school. (0/1 = no/yes)	0.0105	0.0216

(*continued*)

Table 2. Continued

Name	Type	Description	Estimate	SE
OUT.HOL	B	Whether work in the household outside the school. (0/1 = no/yes)	−0.0040	0.0222
OUT.REA	B	Whether read a book/newspaper/magazine outside the school. (0/1 = no/yes)	0.0159	0.0244
SCI.APP	B	Whether attended applied sciences and technology courses in this or last school year. (0/1 = no/yes)	0.0024	0.0293
SCI.BIO	B	Whether attended biology courses in this or last school year. (0/1 = no/yes)	−0.0148	0.0248
SCIACT	C	Index science activities.	0.0060	0.0114
SKICAL	O	The frequency of skipping some classes in the last two full weeks of school. (0/1/2 = [none]/[one or two times]/[three or more times])	−0.0192 −0.0090	0.0204 0.0390
WEALTH	C	Family wealth.	0.0053	0.0108

Note. The variables are ordered according to the value of Π_j when $v = 1$, from the largest to the smallest. For variables with the same Π_j values, they are ordered alphabetically. Continuous, binary, and ordinal variables are indicated by C, B, and O, respectively. For an ordinal variable Z_j, a coefficient corresponds to a dummy variable $\mathbb{I}(Z \geq k)$, for each non-baseline category $k = 1, \ldots, K_j$.

including (a) the average PFER, which is calculated by averaging $|\hat{\mathcal{S}} \setminus \mathcal{S}^*|$ over 100 replications and (b) the average True Positive Rate (TPR), which is calculated by averaging $|\hat{\mathcal{S}} \cap \mathcal{S}^*|/|\mathcal{S}^*|$. As we can see, the baseline algorithm controls the PFER around the nominal level, while the derandomised knockoff method tends to be more conservative, which gives an average PFER much smaller than the nominal level. On the other hand, the derandomised method tends to be more powerful than the baseline algorithm in the sense that it typically achieves a higher average TPR. This phenomenon is consistent with the findings in Ren et al. (2023) under linear and logistic regression settings.

6 Application to PISA 2015

We now apply the proposed method to the PISA 2015 dataset described in Section 2. Our results are given in Table 2. In this table, the predictors are ranked according to the value of Π_j when $v = 1$, from the largest to the smallest. For each predictor, we give the variable name, the variable type (continuous, binary, or ordinal), and a brief explanation of the variable. Further details about these variables are given in the online supplementary material. In addition, we present the estimated coefficients of these variables under the full model (i.e. the model with all the predictors) and their standard errors based on a non-parametric bootstrap procedure with 200 replications. For each continuous variable, the standardised estimated coefficient is given, which is the estimated coefficient multiplied by the standard deviation of the corresponding variable. Variable selection results with nominal PFER levels $v = 1, 2, 3$ are given in Table 2, for which 36, 45, and 48 predictors are selected, respectively. Note that by the construction of the derandomised knockoff method, these selection results are nested, in the sense that the variables selected with $v = t$ are also selected with $v = t + 1$, $t = 1, 2, \ldots$. We also point out that for the first 20 variables (ANXTES to UNFAIR), $\Pi_j = 1$, i.e. the variables are always selected by the baseline algorithm, and for the last 11 variables (FISCED to WEALTH), $\Pi_j = 0$, for any $v = 1, 2, 3$, i.e. they are never selected by the baseline algorithm.

We comment on some of the variable selection results. Several variables in the data concern the socioeconomic status of students' families, including the parents' occupational statuses (FISEIO, MISEIO), cultural possessions at home (CULTPO; e.g. books), parents' education levels (FISCED, MISCED), home educational resources (HEDRES), and family wealth (WEALTH), where FISEIO, MISEIO, FISCED, and MISCED are ordinal variables, and HEDRES, CULTPO, and WEALTH are continuous variables. These variables are positively correlated with each other (correlations/polyserial correlations between 0.22 and 0.69). It is interesting that parents' occupational statuses, cultural possessions, and home educational resources seem to be important in

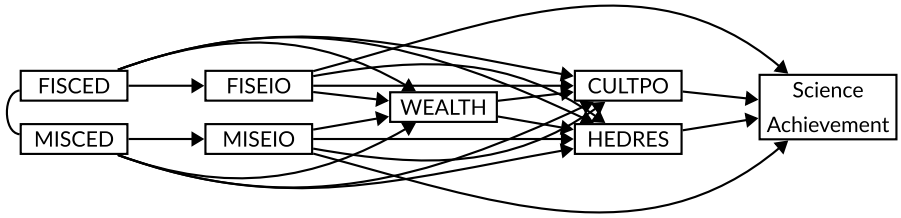

Figure 5. A hypothetical path diagram for several socioeconomic variables and science achievement.

explaining students' performance in science (statistically significant and selected when $v \leq 2$). Given the rest of the variables, it is found that the higher occupational status of the father/mother or the more cultural possessions is associated with better science performance. However, HEDRES has a negative coefficient, which seems to be counter-intuitive and is worth further investigation. On the other hand, parents' education levels and family wealth seem to be less important (statistically insignificant and not selected even when $v = 3$). These results may be interpreted by a hypothetical mediation model as shown in Figure 5, which remains to be validated using additional data and statistical path analysis. That is, WEALTH naturally has direct effects on CULTPO and HEDRES, which may have direct effects on students' science achievement.

Moreover, FISCED and MISCED naturally have a direct effect on FISEIO and MISEIO, respectively, and also possibly have direct effects on CULTPO, HEDRES, and WEALTH. However, there may not be direct paths from WEALTH, FISCED, or MISCED to students' science achievement. Students' science achievement may be largely influenced by genetic factors (e.g. intelligence) and environmental factors (e.g. education resources inside and outside home). It is possible that FISEIO, MISEIO, CULTPO, HEDRES, and the other variables in the current analysis have provided good proxies to these genetic and environmental factors. Given these variables, FISCED, MISCED, and WEALTH tend to be conditionally independent of students' science achievement. Several variables consider students' behaviours attending school, including whether the student has ever repeated a grade (REPEAT), the frequency of a student skipping a whole school day in the last two full weeks of school (SKIDAY), the frequency of the student arriving late for school in the last two full weeks of school (ARRLAT), and the frequency of the student skipping some classes in the last two full weeks of school (SKICLA), where REPEAT is a binary variable, and the other three are ordinal variables. These variables are positively correlated with each other (tetrachoric/polychoric correlations between 0.07 and 0.53). The signs of the estimated coefficients are all consistent with our intuition. For instance, a student tended to perform worse on the test if they had ever repeated a grade or if they often arrived at school late. Among these variables, REPEAT, ARRLAT, and SKIDAY seem to be important variables in the sense that they are all selected with $v \leq 2$. On the other hand, given these variables as well as the rest of the variables, the variable SKICAL seems to be irrelevant (not selected even with $v = 3$, and the coefficients are not significant).

A few variables are related to teachers and their teaching style, including enquiry-based teaching and learning (EBSCIT), teacher-directed science instruction (TDSCIT), perceived feedback (PERFEE), teacher unfairness (UNFAIR), adaptive instruction (ADINST), and teacher support in science classes of students' choice (TEASUP), all of which are continuous variables. Among these variables, ADINST, EBSCIT, TDSCIT, PERFEE, and UNFAIR are selected by our procedure with $v = 1$, while TEASUP is not selected. Variable UNFAIR has a negative coefficient, suggesting that teacher unfairness is associated with poor student performance after controlling for the other variables. ADINST has a positive coefficient, suggesting that teachers' flexibility with their lessons—tailoring the lessons to the students in their classes—tends to improve students' science performance. In addition, it is interesting to see that TDSCIT has a positive coefficient while EBSCIT has a negative coefficient, which suggests that enquiry-based teaching and learning seem to have a negative effect on students' science achievement while teacher-directed instruction has a positive effect. It is possible that enquiry-based teaching and learning can broaden students' interests and increase their enjoyment of science (correlation between EBSCIT and JOYSCI is 0.16 and that between EBSCIT and INTBRS is 0.13) but may be less efficient in developing students' science knowledge than teacher-directed instruction. Thus, a blended instruction model that

combines the two teaching modes may be preferred. Finally, PERFEE has a negative coefficient, which may seem counter-intuitive at first glance, as providing informative and encouraging feedback is essential for improving student outcomes. This result may be due to the confounding of school types, which are not included in the current analysis. That is, students in disadvantaged schools may be more likely to report that their teachers provide them with feedback (Chapter 2, OECD, 2016a). These students also tended to perform worse on the test, which resulted in a negative coefficient estimate.

Several variables concern students' attending of science courses in this or last school year, including chemistry (SCI.CHE), physics (SCI.PHY), earth and space (SCI.EAR), biology (SCI.BIO), general, integrated, or comprehensive science (SCI.GEN), and applied sciences and technology (SCI.APP). All these variables are binary. Among these variables, SCI.CHE and SCI.PHY are selected with $v = 1$, SCI.EAR is selected with $v = 2$, SCI.GEN is selected with $v = 3$, and the rest are not selected even with $v = 3$. For the selected variables, SCI.CHE and SCI.GEN have positive coefficients, while SCI.PHY and SCI.EAR have negative coefficients. We suspect that these results may be due to the different curriculum settings at different types of schools, which are not included in the current model. Besides, there are also variables that measure students' opportunity to learn science at school. In particular, data are available on whether students can choose the number (CHONUM) and level of difficulty (CHODIF) of science courses, and whether they can choose specific science courses (CHOCOU) at school. It turns out that CHONUM and CHODIF are selected with $v = 1$, while CHOCOU is selected with $v = 2$. More specifically, the estimated coefficients for CHOCOU, CHONUM, and CHODIF suggest that students with some freedom to choose the subject, number, and level of difficulty of science courses tended to perform better in the test.

Students' science achievement may also be related to their activities and received support outside of school. The current analysis includes variables on whether a student studies for school or homework (OUT.STU), talks to parents outside the school (OUT.PAR), works for pay (OUT.JOB), exercises or does sports (OUT.SPO), uses internet (OUT.NET), watches TV/DVD/Video (OUT.VED), plays video games (OUT.GAM), has meals (OUT.MEA), meets or talks to friends (OUT.FRI), works in the household (OUT.HOL), and reads a book/newspaper/magazine (OUT.REA) outside the school, and whether they receive emotional support from their parents (EMOSUP). All these variables are binary. Among these variables, OUT.STU, OUT.PAR, OUT.JOB, OUT.SPO, and are selected with $v = 1$, OUT.NET and OUT.VED are selected with $v = 2$, EMOSUP is selected with $v = 3$, and the rest are not selected. Among the selected variables, variables OUT.STU, OUT.JOB, OUT.SPO, OUT.NET, and OUT.VED have positive coefficients, suggesting that students with these outside-of-school activities also tended to perform better in the test after controlling for the rest of the variables. On the other hand, it is counter-intuitive that OUT.PAR and EMOSUP have negative coefficients, though the coefficient for EMOSUP is not statistically insignificant. This is worth future investigation.

Furthermore, the data contain variables that concern students' perceptions or attitudes towards science and related topics. They include the level of enjoying cooperation (COOPER), the level of valuing cooperation (CPSVAL), environmental awareness (ENVAWA), environmental optimism (ENVOPT), epistemological beliefs about science (EPIST), enjoyment of science (JOYSCI), instrumental motivation (INSTSC), science self-efficacy (SCIEEF), and interest in broad science topics (INTBRS). All these variables are continuous. They are all selected. Specifically, INTBRS is selected with $v = 2$, and the rest are selected with $v = 1$. The correlation between CPSVAL and COOPER is 0.45. It is interesting that CPSVAL has a negative coefficient, suggesting that controlling for the other variables, students who more appreciate the value of cooperation and teamwork tended to perform worse in the test. In contrast, COOPER has a positive coefficient, implying that controlling for the other variables, students who enjoy cooperation and teamwork tended to perform better. Variables ENVAWA and ENVOPT have a correlation −0.14. Interestingly, ENVAWA has a positive coefficient, and ENVOPT has a negative coefficient. Moreover, INSTSC, which measures students' perception that studying science in school is useful to their future lives and careers, has a negative coefficient. It seems slightly counter-intuitive. However, such a result is possible, given that variables like JOYSCI and INTBRS have been included in the regression model (correlation between INSTSC and JOYSCI is 0.34 and correlation between INSTSC and INTBRS is 0.24). It may be explained by a mediation model in Figure 6, where INSTSC

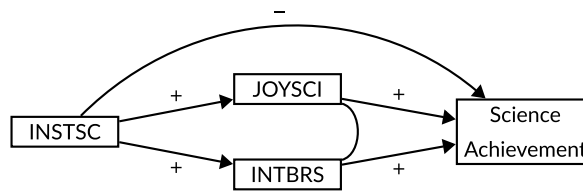

Figure 6. A hypothetical path diagram for INSTSC, JOYSCI, INTBRS, and science achievement.

has positive direct effects on JOYSCI and INTBRS, both of which further have positive effects on students' science achievement. However, given JOYSCI and INTBRS, the direct effect of INSTSC on science achievement is negative, possibly due to that INSTSC also brings pressure and stress to students when they learn science.

Finally, science achievement may also be related to other psychological factors. Specifically, the current analysis considers students' test anxiety level (ANXTES), sense of belonging to school (BELONG), expected education level (EISCED), expected occupational status (EISEIO), and motivation to achieve (MOTIVA), all of which are continuous except that EISCED is ordinal. All these variables are selected with $v = 1$, except for MOTIVA, which is not selected even when $v = 3$. For most of these variables, the signs of the estimated coefficients are consistent with our intuition. Specifically, ANXTES has a negative coefficient, suggesting a higher level of test anxiety is associated with poorer performance, controlling for the rest of the variables. EISCED and EISEIO have positive coefficients, which suggests that higher anticipation of the future is associated with high science achievement. However, it is less intuitive that BELONG has a negative coefficient, which may be due to not accounting for the school effect.

7 Discussions

In this article, we considered identifying non-cognitive predictors of students' academic performance based on complex data from ILSAs that involve many missing values, mixed data types, and measurement errors. This problem can naturally be formulated as a variable selection problem. However, existing statistical methods are not applicable due to the complex data structure. For instance, variable selection methods for linear regression do not solve the current problem due to that (a) the response variable—students' academic achievement—is not directly observable but measured by cognitive items and (b) there are many missing values in the predictors. We addressed these challenges by proposing a new model which combines a latent regression model and a Gaussian copula model. Furthermore, we proposed a derandomised knockoff method under a general latent variable model framework which includes the proposed latent regression model as a special case. This method tackles the multiple comparison issues of variable selection by controlling the PFER, a familywise error rate for variable selection. Theoretical properties of the proposed method were established. We focused on an application to PISA 2015 data, with the response variable being students' proficiency in the science domain. This analysis involved 5,685 students, 184 science items, and 62 non-cognitive variables that are of mixed types and contain many missing values. To our best knowledge, this is the first variable selection study of ILSAs that involves a dataset as large as the current one. With PFER level set to be $v = 1, 2, 3$, the proposed procedure selected 36, 45, and 48 variables, respectively. The model selection results are sensible, and signs of the parameter estimates for most of the selected variables are consistent with our intuition. The variable selection and parameter estimation results were examined from the perspectives of family socioeconomic status, school attending behaviours, teacher-related factors, science course resources and choices at school, out-of-school activities, perception and attitude towards science and related topics, and other psychological factors. These results provided insights into non-cognitive factors that are likely associated with students' science achievement, which can be useful to educators, policymakers, and other stakeholders.

The current analysis has several limitations that will be addressed in future research. First, the current application only considers the US sample and the science domain in PISA. It is of interest to investigate how the result of model selection varies across countries and knowledge domains. In particular, we expect the selection results to be substantially different across different countries due to cultural and

socio-economic differences. In addition, the non-null predictors for different knowledge domains may also differ, which can suggest tailored education strategies for different domains. Second, it is also of interest to extend the current analysis to other ILSAs, such as the TIMSS and PIRLS, to see how the results change with slightly different test designs and different student age groups.

The proposed method may be very useful for the scaling and reporting of ILSAs. First, the variable selection results establish a pathway between students' achievement in each subject domain and its possible influencing factors/causes. These results provide evidence that assists educators, policymakers, and related stakeholders to make informed education decisions. Second, it may improve the scaling methodology of ILSAs. Currently, a latent regression model is used in most ILSAs to estimate the performance distributions of populations (e.g. countries). This model, which is similar to the latent regression model in the current study, borrows information from non-cognitive background variables to compensate for the shortage of cognitive information. However, unlike the current model, the latent regression model adopted in ILSAs does not directly regress on the background variables. Instead, it first conducts a PCA step to reduce the background variables' dimensionality and then incorporates the derived PCA scores as predictors in latent regression. This approach is often criticised for lacking interpretability, as the principal components often lack substantive meanings. Instead of performing PCA, we recommend reducing the dimensionality of the background variables by variable selection and then fitting the latent regression model with the selected predictors. With the theoretical guarantee of our variable selection method and by reporting the selected variables, the estimation and reporting of performance distributions become more transparent and interpretable.

While we focus on an application to ILSAs, the proposed method also receives many other applications. For example, the method can also be used to identify neural determinants of visual short-term memory and to identify demographic correlates of psycho-pathological traits (Jacobucci et al., 2019). Moreover, the proposed Gaussian copula model can be used with other regression models, such as linear and generalised linear regression models, for solving estimation and variable selection problems involving massive missing data and mixed types of variables. It is thus widely applicable to real-world problems involving missing data, which are commonly encountered in the social sciences, such as social surveys, marketing, and public health.

From the methodological perspective, there are several directions worth future development and investigation. First, the current model fails to account for possible multilevel structures in the data; for example, students are nested within schools. From the analysis of PISA data, the signs of some estimated coefficients are not consistent with our intuition, which is likely due to not accounting for the school effect. Therefore, we believe that it is important to extend the current model by introducing random effects to model multilevel structures. New computation methods need to be developed accordingly. Second, the current analysis requires a relatively strong condition on data missingness, which is weaker than MCAR but stronger than MAR. In social science, data may often be missing not at random. In that case, one may simultaneously model the complete data distribution and the missing data mechanism (e.g. Kuha et al., 2018). Such a joint model can be incorporated into the current analysis framework for generating knockoffs and further controlling variable selection errors. Third, the knockoff method may be coupled with the multiple imputation method for missing data analysis, as the knockoff variables can naturally be viewed as missing data. Thus, one may extend the state-of-the-art multiple imputation methods (Liu et al., 2014; Van Buuren, 2018) to simultaneously impute missing data and knockoff copies and then use the imputed data for solving the variable selection problem. Finally, this article focuses on the PFER as the performance metric for variable selection. Other performance metrics may be explored, such as false discovery rate and k family-wise error rate, which may be more sensible in other applications. Making use of recent developments on knockoff methods (Ren & Barber, 2022; Ren et al., 2023), we believe that it is not difficult to extend the current method to these error metrics.

Acknowledgments

The authors thank the editor, associate editor, and anonymous reviewers for their valuable suggestions.

Conflict of interest: None declared.

Funding

Haolei Weng is partially supported by NSF DMS-1915099.

Data availability

The PISA 2015 dataset is available at https://www.oecd.org/pisa/data/2015database/. All code used in this article are available in the following public repository: https://github.com/ZilongXie/LatentRegMissingKnockoff.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Barber R. F., & Candès E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-AOS1337

Barber R. F., & Candès E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, *47*(5), 2504–2537. https://doi.org/10.1214/18-AOS1755

Barber R. F., Candès E. J., & Samworth R. J. (2020). Robust inference with knockoffs. *The Annals of Statistics*, *48*(3), 1409–1431. https://doi.org/10.1214/19-AOS1852

Birnbaum A. (1968). Some latent trait models. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Candès E., Fan Y., Janson L., & Lv J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(3), 551–577. https://doi.org/10.1111/rssb.12265

Chen Y., Li X., Liu J., & Ying Z. (2023). Item response theory – a statistical framework for educational and psychological measurement. *Statistical Science*. To appear.

Cohen J., & Cohen P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Dardanoni V., De Luca G., Modica S., & Peracchi F. (2015). Model averaging estimation of generalized linear models with imputed covariates. *Journal of Econometrics*, *184*(2), 452–463. https://doi.org/10.1016/j.jeconom.2014.06.002

Dardanoni V., Modica S., & Peracchi F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, *162*(2), 362–368. https://doi.org/10.1016/j.jeconom.2011.02.005

Duckworth A. L., & Yeager D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Embretson S. E., & Reise S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Fan J., Liu H., Ning Y., & Zou H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *79*(2), 405–421. https://doi.org/10.1111/rssb.12168

Fan Y., Demirkaya E., Li G., & Lv J. (2019). RANK: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, *115*(529), 362–379. https://doi.org/10.1080/01621459.2018.1546589

Fan Y., Lv J., Sharifvaghefi M., & Uematsu Y. (2020). IPAD: Stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, *115*(532), 1822–1834. https://doi.org/10.1080/01621459.2019.1654878

Farkas G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, *29*(1), 541–562. https://doi.org/10.1146/soc.2003.29.issue-1

Gonzalez E., & Rutkowski L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (IERI Monograph Series, Vol. 3, pp. 125–156). IEA-ETS Research Institute, Hamburg, Germany.

Gov.UK (2015). National curriculum in England: Science programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study.

Grund S., Lüdtke O., & Robitzsch A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, *46*(4), 430–465. https://doi.org/10.3102/1076998620959058

Han F., & Pan W. (2012). A composite likelihood approach to latent multivariate Gaussian modeling of SNP data with application to genetic association testing. *Biometrics*, *68*(1), 307–315. https://doi.org/10.1111/biom.2012.68.issue-1

Jacobucci R., Brandmaier A. M., & Kievit R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science*, *2*(1), 55–76. https://doi.org/10.1177/2515245919826527

Janson L., & Su W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, *10*(1), 960–975. https://doi.org/10.1214/16-EJS1129

Kuha J., Katsikatsou M., & Moustaki I. (2018). Latent variable modelling with non-ignorable item non-response: Multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 1169–1192. https://doi.org/10.1111/rssa.12350

Lee J., & Stankov L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50–64. https://doi.org/10.1016/j.lindif.2018.05.009

Little R. J., & Rubin D. B. (2019). *Statistical analysis with missing data*. Wiley.

Liu J., Gelman A., Hill J., Su Y.-S., & Kropko J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, *101*(1), 155–173. https://doi.org/10.1093/biomet/ast044

Mislevy R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381. https://doi.org/10.1007/BF02306026

Muraki E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

Nielsen S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*(3), 457–489. https://doi.org/10.2307/3318671

OECD (2016a). *PISA 2015 results (Volume II). Policies and practices for successful schools*. OECD Publishing.

OECD (2016b). *PISA 2015 technical report*. OECD publishing.

Ren Z., & Barber R. F. (2022). 'Derandomized knockoffs: Leveraging e-values for false discovery rate control', arXiv, arXiv:2205.15461, preprint: not peer reviewed.

Ren Z., Wei Y., & Candès E. (2023). Derandomizing knockoffs. *Journal of the American Statistical Association*, *118*(542), 948–958. https://doi.org/10.1080/01621459.2021.1962720

Richardson M., Abraham C., & Bond R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. https://doi.org/10.1037/a0026838

Romano Y., Sesia M., & Candès E. (2020). Deep knockoffs. *Journal of the American Statistical Association*, *115*(532), 1861–1872. https://doi.org/10.1080/01621459.2019.1660174

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 2), 461–464. https://doi.org/10.1214/aos/11763441360

Sesia M., Sabatti C., & Candès E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika*, *106*(1), 1–18. https://doi.org/10.1093/biomet/asy033

Singer J. D., Braun H. I., & Chudowsky N. (2018). *International education assessments: Cautions, conundrums, and common sense*. National Academy of Education.

Skrondal A., & Rabe-Hesketh S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Spector A., & Janson L. (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, *50*(1), 252–276. https://doi.org/10.1214/21-AOS2104

Van Buuren S. (2018). *Flexible imputation of missing data*. CRC Press.

von Davier M., Gonzalez E., Kirsch I., & Yamamoto K. (2012). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. Springer.

von Davier M., Gonzalez E., & Mislevy R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (IERI Monograph Series, Vol. 2, pp. 9–36). IEA-ETS Research Institute, Hamburg, Germany.

von Davier M., & Sinharay S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174–193. https://doi.org/10.3102/1076998609346970

Zhang S., & Chen Y. (2022). Computation for latent variable model estimation: A unified stochastic proximal framework. *Psychometrika*, *87*(4), 1473–1502. https://doi.org/10.1007/s11336-022-09863-9

Zhang S., Chen Y., & Liu Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 44–71. https://doi.org/10.1111/bmsp.v73.1