



# Comparison of partial least square algorithms in hierarchical latent variable model with missing data

Simulation: Transactions of the Society for Modeling and Simulation International  
2020, Vol. 96(10) 825–839  
© The Author(s) 2020  
DOI: 10.1177/0037549720944467  
journals.sagepub.com/home/sim



Hao Cheng<sup>1,2,3,4</sup>

## Abstract

Missing data is almost inevitable for various reasons in many applications. For hierarchical latent variable models, there usually exist two kinds of missing data problems. One is manifest variables with incomplete observations, the other is latent variables which cannot be observed directly. Missing data in manifest variables can be handled by different methods. For latent variables, there exist several kinds of partial least square (PLS) algorithms which have been widely used to estimate the value of latent variables. In this paper, we not only combine traditional linear regression type PLS algorithms with missing data handling methods, but also introduce quantile regression to improve the performances of PLS algorithms when the relationships among manifest and latent variables are not fixed according to the explored quantile of interest. Thus, we can get the overall view of variables' relationships at different levels. The main challenges lie in how to introduce quantile regression in PLS algorithms correctly and how well the PLS algorithms perform when missing manifest variables occur. By simulation studies, we compare all the PLS algorithms with missing data handling methods in different settings, and finally build a business sophistication hierarchical latent variable model based on real data.

## Keywords

Partial least square, hierarchical latent variable model, missing data, quantile regression

## 1 Introduction

In recent years, latent variable models have emerged as an important method in various applications.<sup>1–4</sup> The importance of modeling extra constructs representing other latent variables instead of manifest variables has been recognized by many experts and researchers.<sup>5–7</sup> In this case, we consider building a hierarchical latent variable model to establish high-level constructs that reflect other latent variables. Hence, the hierarchical latent variable model (Figure 1) usually contains two layers of latent variables.

Layer 1 in Figure 1 displays the relationship between the first-order latent variables  $\xi_j$  and manifest variables  $x_{jh}$ , ( $j = 1, 2, 3; h = 1, 2, 3$ ). We use  $\lambda_{jh}$  as the factor loading coefficients linking the manifest variables to the first-order latent variable with error terms  $\epsilon_{jh}$ . Hence, layer 1 can be written as the following Equation (1), which is the so-called measurement model. Here,  $\epsilon_{jh}$  is a random measurement error variable with mean 0 and fixed variance for the  $h$ th manifest variable  $x_{jh}$  under the  $j$ th first-order latent variable  $\xi_j$ :

$$x_{jh} = \lambda_{jh}\xi_j + \epsilon_{jh} \quad (1)$$

Layer 2 in Figure 1 displays the relationship between the second-order latent variable  $\eta$  and first-order latent variables  $\xi_j$ ,  $j = 1, 2, 3$ .  $\beta_j$  are the path coefficients linking the first-order latent variables to the second-order latent variable with error terms  $\delta_j$ . Hence, layer 2 can be written as the Equation (2), which is the so-called structural model. Here,  $\delta_j$  is a random measurement error variable with mean 0 and fixed variance for the  $j$ th first-order latent variable  $\xi_j$ :

$$\xi_j = \beta_j\eta + \delta_j \quad (2)$$

<sup>1</sup>National Academy of Innovation Strategy, China Association for Science and Technology, China

<sup>2</sup>School of Statistics, Renmin University of China, China

<sup>3</sup>Department of Biostatistics, Columbia University, USA

<sup>4</sup>Needham Research Institute, Cambridge University, UK

## Corresponding author:

Hao Cheng, National Academy of Innovation Strategy, China Association for Science and Technology, Fuxing Road 3, Haidian District, Beijing, 100863, China.

Email: chenghao0524@yeah.net

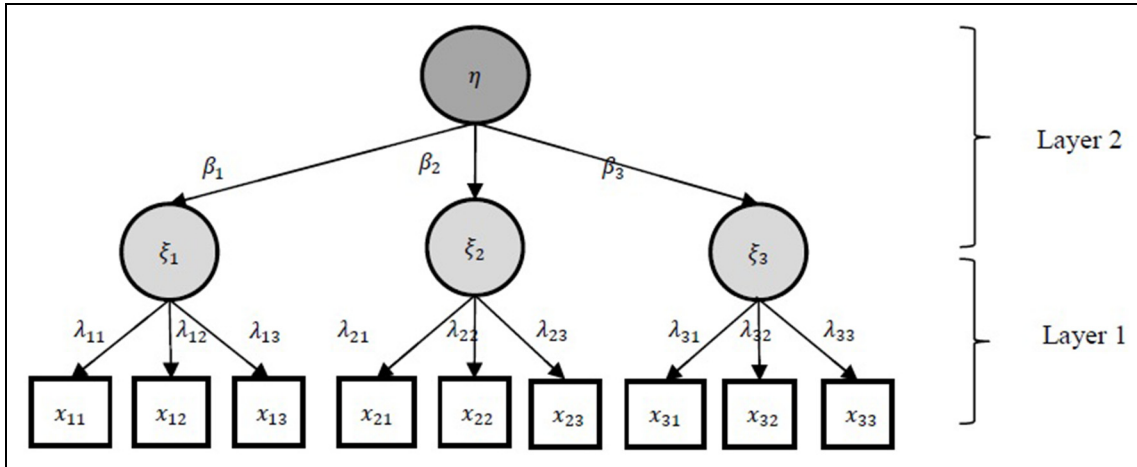


Figure 1. Hierarchical latent variable model.

In the hierarchical latent variable model, there exist two kinds of variables, manifest variables and latent variables, which may present two kinds of missing data problems.

The first missing data problem is about manifest variables with incomplete observations. This seems more common in many data sets. In hierarchical latent variable models, the relatively suitable and widely used missing data handling methods include complete case analysis (CC), mean value replacement (MEAN), and others.<sup>8</sup> However, CC can diminish the number of observations and likely leads to biased results. MEAN decreases the variability of data and may reduce the possibility of finding meaningful relationships. Multiple imputation (MI) may lead to low computing efficiency and fractional imputation (FI) needs at least part of the response variable observations.<sup>9-12</sup> In this paper, we consider the weighted  $k$ -nearest neighbors imputation method (KN) and  $k$ -nearest neighbors imputation method based on median value (KNM) along with CC and MEAN. In the case of categorical variables, KNM uses the most frequent value instead of the median value. KN means use a weighted average of the  $k$ -nearest neighbors. Equation (3) is used to calculate the weights. We denote  $w$  as the weights and  $-D(k, x)$  as the Euclidean distance between the case with missing values  $x$  and the neighbor  $k$ :

$$w = \exp(-D(k, x)) \tag{3}$$

The second missing data problem is latent variables that cannot be observed directly. As one of the most widely used methods to obtain the values of latent variables, partial least square (PLS) methods are very powerful for avoiding the joint normal distribution and independence assumptions, and estimate path coefficients and loading coefficients of the model, even if the sample size is relatively small (such as 50).<sup>13-26</sup> There exist three well-

known PLS algorithms for estimating second-order (and higher) latent variable models. These are the repeated indicators algorithm (RI), the two-step algorithm (TS), and the hybrid algorithm (H). All of these algorithms use traditional linear regression-like models and the weights are estimated as simple or multiple linear regression parameters. Hence, they cannot show different levels of relationship between latent variables and manifest variables, but only provide average effects. To improve upon the previously mentioned limitations, we consider quantile regression estimation to expand the category of PLS. Here, we consider linear quantile regression as  $Q_Y(\tau) = x^T \beta_\tau$ ,  $\forall \tau \in (0, 1)$ , where  $Q_Y(\tau)$  is the  $\tau$ th quantile of a response variable  $Y$ , and  $x$  is the covariate vector.<sup>27,28</sup> To capture the changing relationships at the explored quantile of interest and the overall view of the structural relationship among all variables at different levels with missing data, we propose two kinds of PLS algorithm with missing data handling functions based on quantile regression. Therefore, we compare all five PLS algorithms in different models and settings.

Based on the previous, the main contributions or innovations of this paper can be summarized as: (1) We propose quantile regression type hierarchical latent variable models and corresponding PLS algorithms when parts of manifest variables contain missing data. (2) We compare different PLS algorithms with missing data (PLSMD) algorithms based on both linear regression and quantile regression models under different settings. (3) We write R code to accomplish different PLSMD algorithms with missing data handling functions. (4) We use PLSMD algorithms to build business sophistication hierarchical latent variable models based on real data. Therefore, there exist challenges in how to apply quantile regression correctly to both hierarchical latent variable models and PLS algorithms. As we can see in Equations (1) and (2), the random errors  $\epsilon_{jt}$

and  $\delta_j$  are required to have mean 0 and certain fixed variances. When quantile regressions are introduced in hierarchical latent variable models, such assumptions should be considered first. As we know, one obvious advantage of quantile regression is no extra distribution assumptions on random errors. Therefore, we can hold the same assumptions about random errors for our new quantile regression type hierarchical latent variable models. For PLSMD algorithms based on quantile regression, we simply set a class of quantile levels and use linear quantile regression to accomplish the estimation of the weights or coefficients and latent variables' scores at each quantile level.

The rest of the paper is organized as follows. We describe our PLSMD algorithms in Section 2, and conduct a simulation study in Section 3. In Section 4 we apply all five PLSMD algorithms with four different missing data handling methods to build a business sophistication hierarchical latent variable model based on real data from the Global Innovation Index 2018.

## 2 Estimation with PLS algorithms

### 2.1 The general PLS procedure with missing data

As a useful tool for model investigation with a high level of abstraction, PLS estimation specifies the estimates of the latent variables to be weighted aggregates and scales the estimated latent variable scores  $\xi_j$  and  $\eta$  to unit variance.<sup>13</sup> When there are missing data in manifest variables, we should diagnose all the manifest variables with missing data first, choose appropriate and fast missing data handling methods from those we mentioned before and imputing the missing values. Here, we denote  $x_{jh}^{new}$  (contains the existing values and imputed values) as the new manifest variables with complete observations. Based on the complete data, the parameter estimation of general PLS follows a double approximation of the latent variables: external estimation and internal estimation.<sup>26</sup>

In external estimation,  $\xi_j^{ext}$  is obtained as the product of the block of manifest variables  $x_{jh}^{new}$  with the external weights  $w_{jh}$  (which represent the estimation of measurement coefficients  $\lambda_{jh}$ ).  $\eta$  is obtained as the product of the block of the external estimation of first-order latent variables  $\xi_j^{ext}$  with the external weights  $w_j$ :

$$\xi_j^{ext} = \sum_{h=1}^{H_j} w_{jh} x_{jh}^{new} \quad (4)$$

$$\eta^{ext} = \sum_{j=1}^J w_j \xi_j^{ext} \quad (5)$$

In internal estimation,  $\xi_j^{int}$  is obtained as the product of the external estimation of  $\xi_j^{ext}$  with the internal weights  $e_j$ .  $\eta^{int}$  is obtained as the product of the external estimation of  $\eta^{ext}$  with the internal weights  $e$ :

$$\xi_j^{int} = e_j \xi_j^{ext} \quad (6)$$

$$\eta^{int} = e \eta^{ext} \quad (7)$$

To estimate the inner weights  $e_j$ , we need to calculate the correlation between each pair of external estimations  $\xi_j^{ext}$  and  $\eta^{ext}$  and the sign of the correlation  $e_j = \text{sign}(\text{cor}(\xi_j^{ext}, \eta^{ext}))$ , which is called the centroid scheme.<sup>17</sup>

In the external weights-updating procedure, we need to consider the kind of relationship with their latent variables in Figure 1. The manifest variables are a reflection of the latent variables. In other words, the manifest variables can be treated as the response variable of the first-order latent variables, and first-order latent variables can be treated as the response variable of the second-order latent variable. Modified by quantile regression, we get new estimates of external weights by the following two equations:

$$w_{jh} = \text{argmin} \sum_{i=1}^N \{x_{jh,i}^{new} - \xi_{j,i}^{int} w_{jh,i}\}^2 \quad (8)$$

$$w_j = \text{argmin} \sum_{i=1}^N \{\xi_{j,i}^{ext} - \eta^{int} w_{j,i}\}^2 \quad (9)$$

The weight estimation is an iterative procedure between the external estimation and internal estimation. The whole PLS procedure will not stop until it reaches the maximum number of iterations or the change in the outer weights between two consecutive iterations is smaller than this stop criterion value at the same time.<sup>20,21</sup> Maximum iterations represents the maximum number of iterations that will be used for calculating the PLS results. Maximum iterations should be sufficiently large and stop criterion should be sufficiently small. In this paper, maximum iterations is set to 200 and minimum change in the external weights between two consecutive iterations is less than  $10^{-5}$ .

Algorithm 1 shows the steps of the general PLSMD procedure for the hierarchical latent variable model.

### 2.2 The PLSMD algorithms

**2.2.1 Linear regression-based PLSMD algorithms: RIMD, TSMD, and HMD.** The RI approach, TS approach, and H approach algorithms are all based on linear regression to establish structural equation models or hierarchical latent variable models, which are constructed by Equations (1) and (2).<sup>29-31</sup> One of the main differences in these algorithms lies in the assignment patterns of manifest variables. RI assigns all the manifest variables of the first-order latent variables to the second-order latent variable at the same time. TS uses manifest variables only for the first-order latent variable and then uses the estimated scores of first-order latent variables for the second-order

**Algorithm 1.** The general PLS procedure with missing data (PLSMD) for hierarchical latent variable model

---

Step 1	Handle missing data.
Step 1.1	Select manifest variables with missing data and calculate missing rates.
Step 1.2	Choose appropriate and fast imputation methods to handle missing values.
Step 2	Initialize outer weights $\mathbf{w}_{jh}^{(1)}$ and $\mathbf{w}_j^{(1)}$ .
Step 3	External estimation. Use Equations (4) and (5) to calculate $\xi_j^{\text{ext},(l)}$ , $\eta^{\text{ext},(l)}$ for the $l$ th iteration.
Step 4	Internal estimation.
Step 4.1	Choose a centroid scheme and calculate $e^{(l)}$ .
Step 4.2	Use Equations (6) and (7) to calculate $\xi_j^{\text{int},(l)}$ , $\eta^{\text{int},(l)}$ for the $l$ th iteration.
Step 5	Update the external weights under a set of quantile levels.
Step 5.1	Define a set of quantile levels.
Step 5.2	Estimate the external weights among first-order latent variables and manifest variables by Equation (8).
Step 5.3	Estimate the external weights among second-order and first-order latent variables by Equation (9).
Step 6	Repeat steps 3–5 until the stop criterion or the maximum number of iterations are reached.

---

latent variable. H randomly splits the manifest variables so that part of them are assigned to first-order latent variables while the others are assigned to the second-order latent variable. A disadvantage of RI is a possible bias of the estimates because it relates variables of the same type. TS estimates any second-order construct in stage 2 without considering the first-order latent variable scores in stage 1. In H, all manifest variables are randomly assigned to first- and second-order latent variables without replacement, which may lead to the uncertainty of structural relationship each time. In this paper, we combine missing data handling methods with the previously mentioned three PLS algorithms and denote them as RIMD, TSMD, and HMD, respectively.

**2.2.2 Quantile regression-based PLSMD algorithms: PLSMD<sub>0</sub> and PLSMD<sub>τ</sub>** All of the previously mentioned three PLSMD algorithms are based on linear regression models, which offer conditional mean views of the relationship between the response and its covariates. However, quantile regression, which is now an indispensable and versatile tool for statistical research, broadens conditional mean views by allowing covariate effects to be examined at different quantiles and makes the estimates more resistant to outliers.<sup>32,33</sup> Therefore, we consider the following two PLSMD algorithms: one-stage PLS algorithm based on quantile regression with missing data handling methods (PLSMD<sub>0</sub>) and two-stage PLS algorithm based on quantile regression with missing data handling methods (PLSMD<sub>τ</sub>).

PLSMD<sub>0</sub> still contains three parts: external estimation, internal estimation, and external weights updating. PLSMD<sub>0</sub> is different from the PLSMD algorithm in Section 2.1 in two respects. The first difference is rebuilding the hierarchical latent variable based on quantile regression. Thus models (1) and (2) can be modified as  $Q_{x_{jh}^{\text{new}}}(\tau) = \lambda_{jh,\tau} \xi_j$  and  $Q_{\xi_j}(\tau) = \beta_{j,\tau} \eta$ , respectively. Where  $Q_{x_{jh}^{\text{new}}}(\tau)$  stands for the  $\tau$ th quantile of the manifest variables  $x_{jh}^{\text{new}}$ ,  $Q_{\xi_j}(\tau)$  stands for the  $\tau$ th quantile of the first-

order latent variables  $\xi_j$ . In addition, we denote  $\beta_{j,\tau}$  as the path coefficient at quantile level  $\tau$ ,  $\lambda_{jh,\tau}$  as the factor loading coefficient at quantile level  $\tau$ . The second difference lies in the external weights-updating procedure. Instead of using the least square method, we take  $\rho_\tau(u) = u(\tau - I(u < 0))$  as the loss function to calculate the external weights in each iteration. Modified by quantile regression, we get new estimates of external weights by  $\mathbf{w}_{jh,\tau} = \text{argmin} \sum_{i=1}^N \rho_\tau \{x_{jh,i}^{\text{new}} - \xi_{j,i}^{\text{int}} \mathbf{w}_{jh,\tau,i}\}$  and  $\mathbf{w}_{j,\tau} = \text{argmin} \sum_{i=1}^N \rho_\tau \{\xi_{j,i}^{\text{ext}} - \eta^{\text{int}} \mathbf{w}_{j,\tau,i}\}$  for all  $\tau \in (0, 1)$ .

PLSMD<sub>τ</sub> is just like the existing TSMD. In the first step, we estimate the first-order latent variable scores by using the first step of the existing TS approach. That is, we can get the score of a first-order latent variable by taking the first principal component of its indicators and the principal component analysis (PCA) scores of first-order latent variables are subsequently used as indicators for the second-order latent variable in a separate hierarchical latent variable model. In the second step, we modify the existing step 2 by  $Q_{\xi_j}(\tau) = \beta_{j,\tau} \eta$ ,  $\forall \tau \in (0, 1)$ , where  $Q_{\xi_j}(\tau)$  stands for the  $\tau$ th quantile of the first-order latent variables  $\xi_j$ . Different from PLSMD<sub>0</sub>,  $\xi_j$  are manifest variables instead of latent variables in PLSMD<sub>τ</sub>, and  $\eta$  can be treated as the first-order latent variable of  $\xi_j$ . The parameter estimation is similar as PLSMD<sub>0</sub>. Modified by quantile regression instead of traditional linear regression, we get new estimates of external weights by the following:  $\mathbf{w}_{j,\tau} = \text{argmin} \sum_{i=1}^N \rho_\tau \{\xi_{j,i}^{\text{ext}} - \eta^{\text{int}} \mathbf{w}_{j,\tau,i}\}$ ,  $\forall \tau \in (0, 1)$ .

## 3 Simulations

### 3.1 Model

Here we consider the following two equations to generate data. We refer to Ciavolino and Nitti's simulation plan and assume the path coefficients are 0.8 and loading coefficients are 0.7.<sup>26</sup> The second-order latent variable  $\eta$  follows a standard normal distribution  $N(0, 1)$ . The error term  $\epsilon_{jh}$  follows a continuous uniform  $U(-1, 1)$ . The error term  $\delta_j$  follows a univariate normal distribution

$N(0, Var_{\delta_j})$ . Where  $J$  is the number of first-order latent variables and  $H_j$  is the number of observed variables for the  $j$ th first-order latent variable  $\xi_j$ :

$$x_{jh} = 0.7\xi_j + \epsilon_{jh}, \quad \forall j = 1, \dots, J, \quad h = 1, \dots, H_j \quad (10)$$

$$\xi_j = 0.8\eta + \delta_j, \quad \forall j = 1, \dots, J \quad (11)$$

Here we use  $R^2$  as the ratio of  $Var(model)$  to  $Var(total)$ . And we calculate  $Var_{\delta_j}$  by  $R^2$ :

$$\begin{aligned} R^2 &= \frac{Var(model)}{Var(total)} = \frac{Var(model)}{Var(model) + Var(error)} \\ &= \frac{Var(\beta_j \eta)}{Var(\beta_j \eta) + Var(\delta_j)} = \frac{\beta_j^2 Var(\eta)}{\beta_j^2 Var(\eta) + Var(\delta_j)} \end{aligned} \quad (12)$$

Hence we get  $Var_{\delta_j}$  by the following equation:

$$Var_{\delta_j} = \beta_j^2 Var(\eta) \left( \frac{1}{R^2} - 1 \right) \quad (13)$$

Here,  $\beta_j$  equals 0.8,  $Var(\eta)$  equals 1, and  $R^2$  equals 0.8. Hence, we can calculate  $Var_{\delta_j}$  as  $(1/R^2 - 1)\beta_j^2 = 0.16$ . Finally, we can generate the first-order latent variable scores and all the manifest variables according to models (10) and (11).

### 3.2 Settings

To investigate the performance of PLSMD algorithms with different missing data handling methods, we mainly consider the following three aspects.

#### 1. Balanced model or unbalanced model (B or U).

We design different numbers of manifest variables and also consider whether they are assigned evenly to first-order latent variables or not. If the manifest variables are assigned evenly to first-order latent variables (balanced model), we set 15 manifest variables in total and assign 5 manifest variables for each first-order latent variable. If the manifest variables are not assigned evenly to first-order latent variables (unbalanced model), we set 18 manifest variables in total and assign 4, 6, and 8 manifest variables for each first-order latent variable respectively. Hence, we get the following two settings:

S1.1 : Balanced model (B): (5, 5, 5)

S1.2 : Unbalanced model (U): (4, 6, 8)

#### 2. Number of missing variables (NMV).

We set different numbers of manifest variables with missing data for both the balanced model and the unbalanced model. Here we assume that there are three manifest variables with incomplete observations first, and then six manifest variables with

incomplete observations. For the balanced/unbalanced models, we assign one manifest variable with incomplete observations for each first-order latent variable first, and then assign two manifest variables with incomplete observations for each first-order latent variable:

S2.1 : One manifest variable with missing data for each first-order latent variable (O)

S2.2 : Two manifest variables with missing data for each first-order latent variable (T)

#### 3. Missing rates (MR).

Here, part of the manifest variables are missing at random (MAR) and the other manifest variables are completely observed. We define  $p(\delta_i|x_i) = \max\{0, [(x_i + 1.65)/10]^{1/20}\}$  as the missing probability function when the missing rate is approximately 0.1 and  $p(\delta_i|x_i) = \max\{0, [(x_i + 1.2)/10]^{1/20}\}$  as the missing probability function when the missing rate is approximately 0.2:

S3.1 : Missing rate equals 10% (0.1)

S3.2 : Missing rate equals 20% (0.2)

Based on these aspects, we conducted the following numerical investigations and simulations (see Table 1). In all investigations, we choose the Monte Carlo sample size as 200. The methods are the RI approach (RIMD), two-step approach (TSMD), hybrid approach (HMD), new two-stage PLS algorithm (PLSMD<sub>T</sub>), and new one-stage PLS algorithm (PLSMD<sub>O</sub>). According to all the settings and methods, we mainly compared the following two aspects:

1. We compared the estimation accuracy and efficiency of path coefficients of different algorithms with different missing data handling methods.
2. We compared the latent variables' prediction accuracy of different algorithms with different missing data handling methods.

**Table 1.** Simulation plan.

	Balanced or unbalanced	NMV	MR
Simu. 01	Balanced	1	0.1
Simu. 02	Balanced	1	0.2
Simu. 03	Balanced	2	0.1
Simu. 04	Balanced	2	0.2
Simu. 05	Unbalanced	1	0.1
Simu. 06	Unbalanced	1	0.2
Simu. 07	Unbalanced	2	0.1
Simu. 08	Unbalanced	2	0.2

MR, missing rates; NMV, number of missing manifest variables.

**Table 2.** Mean biases (MB), standard errors (SE), and mean squared errors (MSE) of the estimated path coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> under Simu.01 and 02 from 200 Monte Carlo replicates with sample size 500.

	CC			MEAN			KN			KNM		
	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE
Simu. 01												
RIMD	0.074	0.015	0.006	0.086	0.011	0.008	0.089	0.011	0.008	0.089	0.011	0.008
	0.072	0.013	0.005	0.085	0.010	0.007	0.088	0.010	0.008	0.088	0.011	0.008
	0.071	0.015	0.005	0.084	0.011	0.007	0.087	0.011	0.008	0.087	0.011	0.008
TSMD	0.074	0.012	0.006	0.085	0.010	0.007	0.089	0.010	0.008	0.089	0.010	0.008
	0.072	0.012	0.005	0.084	0.010	0.007	0.088	0.010	0.008	0.087	0.010	0.008
	0.071	0.013	0.005	0.083	0.010	0.007	0.087	0.010	0.008	0.087	0.010	0.008
HMD	-0.069	0.024	0.005	-0.051	0.018	0.003	-0.042	0.018	0.002	-0.042	0.018	0.002
	-0.069	0.025	0.005	-0.049	0.019	0.003	-0.037	0.020	0.002	-0.038	0.020	0.002
	-0.072	0.024	0.006	-0.052	0.019	0.003	-0.043	0.019	0.002	-0.043	0.019	0.002
PLSMD <sub>O<sub>25</sub></sub>	0.073	0.031	0.006	0.080	0.024	0.007	0.086	0.024	0.008	0.086	0.024	0.008
	0.066	0.028	0.005	0.074	0.023	0.006	0.081	0.024	0.007	0.080	0.023	0.007
	0.067	0.031	0.005	0.075	0.026	0.006	0.081	0.024	0.007	0.081	0.024	0.007
PLSMD <sub>O<sub>50</sub></sub>	0.073	0.024	0.006	0.086	0.019	0.008	0.089	0.018	0.008	0.089	0.018	0.008
	0.072	0.026	0.006	0.083	0.019	0.007	0.088	0.019	0.008	0.088	0.018	0.008
	0.071	0.028	0.006	0.083	0.020	0.007	0.087	0.019	0.008	0.087	0.019	0.008
PLSMD <sub>T<sub>25</sub></sub>	0.072	0.031	0.006	0.080	0.024	0.007	0.085	0.024	0.008	0.086	0.024	0.008
	0.065	0.028	0.005	0.076	0.023	0.006	0.080	0.024	0.007	0.080	0.024	0.007
	0.067	0.031	0.006	0.076	0.026	0.006	0.081	0.024	0.007	0.081	0.024	0.007
PLSMD <sub>T<sub>50</sub></sub>	0.073	0.023	0.006	0.086	0.019	0.008	0.089	0.017	0.008	0.088	0.018	0.008
	0.073	0.026	0.006	0.084	0.018	0.007	0.088	0.019	0.008	0.088	0.018	0.008
	0.072	0.028	0.006	0.082	0.020	0.007	0.087	0.019	0.008	0.087	0.019	0.008
Simu. 02												
RIMD	0.060	0.018	0.004	0.085	0.011	0.007	0.088	0.011	0.008	0.088	0.011	0.008
	0.058	0.018	0.004	0.083	0.011	0.007	0.086	0.011	0.008	0.086	0.011	0.008
	0.058	0.019	0.004	0.082	0.012	0.007	0.085	0.012	0.007	0.085	0.012	0.007
TSMD	0.060	0.015	0.004	0.083	0.010	0.007	0.088	0.010	0.008	0.088	0.010	0.008
	0.057	0.015	0.004	0.082	0.010	0.007	0.086	0.010	0.008	0.086	0.010	0.008
	0.056	0.017	0.003	0.081	0.010	0.007	0.085	0.010	0.007	0.085	0.010	0.007
HMD	-0.101	0.029	0.011	-0.059	0.019	0.004	-0.049	0.018	0.003	-0.049	0.018	0.003
	-0.098	0.031	0.011	-0.056	0.020	0.003	-0.042	0.020	0.002	-0.042	0.020	0.002
	-0.102	0.030	0.011	-0.061	0.020	0.004	-0.050	0.019	0.003	-0.050	0.019	0.003
PLSMD <sub>O<sub>25</sub></sub>	0.052	0.037	0.004	0.070	0.024	0.005	0.078	0.025	0.007	0.078	0.025	0.007
	0.047	0.034	0.003	0.064	0.024	0.005	0.073	0.024	0.006	0.073	0.024	0.006
	0.050	0.038	0.004	0.066	0.028	0.005	0.073	0.027	0.006	0.075	0.027	0.006
PLSMD <sub>O<sub>50</sub></sub>	0.061	0.029	0.005	0.084	0.020	0.007	0.088	0.020	0.008	0.087	0.019	0.008
	0.059	0.029	0.004	0.083	0.019	0.007	0.087	0.018	0.008	0.088	0.018	0.008
	0.056	0.033	0.004	0.080	0.020	0.007	0.085	0.020	0.008	0.086	0.020	0.008
PLSMD <sub>T<sub>25</sub></sub>	0.052	0.036	0.004	0.070	0.025	0.005	0.077	0.024	0.006	0.077	0.025	0.006
	0.047	0.034	0.003	0.065	0.023	0.005	0.071	0.024	0.006	0.072	0.023	0.006
	0.050	0.037	0.004	0.067	0.028	0.005	0.073	0.027	0.006	0.074	0.027	0.006
PLSMD <sub>T<sub>50</sub></sub>	0.060	0.029	0.004	0.083	0.019	0.007	0.088	0.019	0.008	0.087	0.019	0.008
	0.060	0.029	0.004	0.082	0.019	0.007	0.088	0.018	0.008	0.088	0.018	0.008
	0.055	0.033	0.004	0.079	0.020	0.007	0.085	0.020	0.008	0.086	0.019	0.008

CC, complete case analysis; HMD hybrid algorithm with missing data handling methods; KN, weighted *k*-nearest neighbors imputation method; KNM, *k*-nearest neighbors imputation method based on median value; MEAN, mean imputation method; PLSMD<sub>O<sub>25</sub></sub> and PLSMD<sub>O<sub>50</sub></sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T<sub>25</sub></sub> and PLSMD<sub>T<sub>50</sub></sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

### 3.3 Results

**3.3.1 Comparisons of path coefficients' estimation accuracy and efficiency.** Tables 2–5 present the mean biases, standard errors, and mean square errors of the estimated path

coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub> and PLSMD<sub>T</sub> under *Simu.01 – 02*, *Simu.03 – 04*, *Simu.05 – 06*, and *Simu.07 – 08*, respectively from 200 Monte Carlo replicates with sample size 500. We also run

**Table 3.** Mean biases (MB), standard errors (SE), and mean squared errors (MSE) of the estimated path coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> under Simu.03 and 04 from 200 Monte Carlo replicates with sample size 500.

	CC			MEAN			KN			KNM		
	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE
Simu. 03												
RIMD	0.068	0.019	0.005	0.083	0.011	0.007	0.089	0.011	0.008	0.088	0.011	0.008
	0.063	0.019	0.004	0.080	0.011	0.007	0.087	0.011	0.008	0.086	0.011	0.008
	0.065	0.019	0.005	0.080	0.012	0.007	0.086	0.012	0.008	0.086	0.012	0.008
TSMD	0.067	0.015	0.005	0.081	0.010	0.007	0.088	0.010	0.008	0.088	0.010	0.008
	0.063	0.017	0.004	0.080	0.010	0.006	0.087	0.010	0.008	0.087	0.010	0.008
	0.064	0.017	0.004	0.079	0.011	0.006	0.086	0.011	0.008	0.086	0.011	0.007
HMD	-0.080	0.030	0.007	-0.063	0.019	0.004	-0.042	0.019	0.002	-0.043	0.019	0.002
	-0.087	0.032	0.009	-0.065	0.021	0.005	-0.044	0.021	0.002	-0.045	0.021	0.002
	-0.086	0.029	0.008	-0.073	0.020	0.006	-0.052	0.019	0.003	-0.052	0.019	0.003
PLSMD <sub>O<sub>25</sub></sub>	0.060	0.040	0.005	0.067	0.024	0.005	0.075	0.026	0.006	0.076	0.026	0.006
	0.057	0.040	0.005	0.063	0.025	0.005	0.072	0.023	0.006	0.073	0.023	0.006
	0.058	0.038	0.005	0.064	0.027	0.005	0.074	0.025	0.006	0.074	0.026	0.006
PLSMD <sub>O<sub>50</sub></sub>	0.067	0.029	0.005	0.079	0.020	0.007	0.089	0.019	0.008	0.088	0.020	0.008
	0.065	0.031	0.005	0.083	0.020	0.007	0.089	0.019	0.008	0.089	0.020	0.008
	0.064	0.032	0.005	0.079	0.020	0.007	0.088	0.019	0.008	0.088	0.019	0.008
PLSMD <sub>T<sub>25</sub></sub>	0.060	0.040	0.005	0.068	0.024	0.005	0.076	0.025	0.006	0.077	0.026	0.007
	0.056	0.039	0.005	0.063	0.025	0.005	0.072	0.023	0.006	0.073	0.023	0.006
	0.059	0.036	0.005	0.064	0.026	0.005	0.074	0.025	0.006	0.074	0.025	0.006
PLSMD <sub>T<sub>50</sub></sub>	0.067	0.028	0.005	0.079	0.020	0.007	0.089	0.019	0.008	0.088	0.019	0.008
	0.065	0.032	0.005	0.082	0.019	0.007	0.089	0.019	0.008	0.089	0.020	0.008
	0.064	0.032	0.005	0.079	0.019	0.007	0.088	0.018	0.008	0.087	0.019	0.008
Simu. 04												
RIMD	0.048	0.026	0.003	0.079	0.012	0.006	0.085	0.012	0.007	0.085	0.012	0.007
	0.042	0.025	0.002	0.076	0.012	0.006	0.082	0.012	0.007	0.082	0.012	0.007
	0.045	0.026	0.003	0.076	0.013	0.006	0.082	0.013	0.007	0.082	0.013	0.007
TSMD	0.047	0.021	0.003	0.076	0.011	0.006	0.085	0.011	0.007	0.085	0.011	0.007
	0.041	0.022	0.002	0.074	0.011	0.006	0.083	0.011	0.007	0.082	0.011	0.007
	0.043	0.023	0.002	0.074	0.012	0.006	0.082	0.012	0.007	0.082	0.012	0.007
HMD	-0.123	0.040	0.017	-0.080	0.020	0.007	-0.056	0.020	0.004	-0.056	0.020	0.004
	-0.131	0.042	0.019	-0.082	0.022	0.007	-0.058	0.022	0.004	-0.059	0.022	0.004
	-0.131	0.041	0.019	-0.101	0.022	0.011	-0.073	0.022	0.006	-0.073	0.022	0.006
PLSMD <sub>O<sub>25</sub></sub>	0.035	0.051	0.004	0.050	0.025	0.003	0.061	0.025	0.004	0.063	0.025	0.005
	0.028	0.053	0.004	0.044	0.027	0.003	0.056	0.025	0.004	0.055	0.025	0.004
	0.029	0.050	0.003	0.042	0.028	0.003	0.057	0.027	0.004	0.057	0.026	0.004
PLSMD <sub>O<sub>50</sub></sub>	0.047	0.038	0.004	0.076	0.021	0.006	0.084	0.019	0.007	0.084	0.020	0.007
	0.046	0.042	0.004	0.077	0.022	0.006	0.087	0.021	0.008	0.086	0.020	0.008
	0.040	0.044	0.004	0.075	0.021	0.006	0.082	0.020	0.007	0.083	0.021	0.007
PLSMD <sub>T<sub>25</sub></sub>	0.034	0.051	0.004	0.049	0.025	0.003	0.060	0.024	0.004	0.061	0.024	0.004
	0.028	0.052	0.003	0.044	0.026	0.003	0.054	0.025	0.004	0.054	0.025	0.004
	0.030	0.048	0.003	0.043	0.028	0.003	0.054	0.027	0.004	0.054	0.026	0.004
PLSMD <sub>T<sub>50</sub></sub>	0.046	0.038	0.004	0.076	0.021	0.006	0.084	0.019	0.007	0.084	0.019	0.007
	0.048	0.041	0.004	0.077	0.022	0.006	0.087	0.020	0.008	0.086	0.020	0.008
	0.040	0.044	0.003	0.075	0.021	0.006	0.083	0.021	0.007	0.083	0.020	0.007

CC, complete case analysis; HMD hybrid algorithm with missing data handling methods; KN, weighted *k*-nearest neighbors imputation method; KNM, *k*-nearest neighbors imputation method based on median value; MEAN, mean imputation method; PLSMD<sub>O<sub>25</sub></sub> and PLSMD<sub>O<sub>50</sub></sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T<sub>25</sub></sub> and PLSMD<sub>T<sub>50</sub></sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

the same simulation process with an increased sample size of 1000, and obtain very similar results. For brevity, we do not report these in the paper. For PLSMD<sub>O</sub> and PLSMD<sub>T</sub>, we give the estimated path coefficients at quantile levels 0.25 and 0.50. The balanced model means each

first-order latent variable has five manifest variables. The unbalanced model means the first-order latent variables have four, six, and eight manifest variables, respectively. According to Tables 2–5, we get the following main conclusions:

**Table 4.** Mean biases (MB), standard errors (SE), and mean squared errors (MSE) of the estimated path coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> under Simu.05 and 06 from 200 Monte Carlo replicates with sample size 500.

	CC			MEAN			KN			KNM		
	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE
Simu. 05												
RIMD	0.027	0.017	0.001	0.039	0.014	0.002	0.045	0.014	0.002	0.045	0.014	0.002
	0.078	0.014	0.006	0.090	0.010	0.008	0.092	0.010	0.009	0.092	0.010	0.009
	0.116	0.012	0.014	0.127	0.008	0.016	0.128	0.008	0.016	0.128	0.008	0.016
TSMD	0.069	0.012	0.005	0.080	0.010	0.006	0.084	0.010	0.007	0.084	0.010	0.007
	0.079	0.013	0.006	0.091	0.009	0.008	0.094	0.009	0.009	0.094	0.009	0.009
	0.083	0.013	0.007	0.095	0.010	0.009	0.098	0.010	0.010	0.098	0.010	0.010
HMD	-0.108	0.026	0.012	-0.085	0.021	0.008	-0.073	0.022	0.006	-0.073	0.022	0.006
	-0.049	0.024	0.003	-0.026	0.018	0.001	-0.019	0.018	0.001	-0.019	0.019	0.001
	-0.003	0.020	0.000	0.014	0.016	0.000	0.020	0.016	0.001	0.019	0.016	0.001
PLSMD <sub>O.25</sub>	0.064	0.030	0.005	0.070	0.025	0.005	0.076	0.022	0.006	0.076	0.023	0.006
	0.075	0.028	0.006	0.086	0.022	0.008	0.090	0.023	0.009	0.090	0.023	0.009
	0.080	0.028	0.007	0.088	0.022	0.008	0.094	0.023	0.009	0.093	0.023	0.009
PLSMD <sub>O.50</sub>	0.069	0.027	0.005	0.080	0.021	0.007	0.085	0.021	0.008	0.085	0.021	0.008
	0.082	0.025	0.007	0.091	0.021	0.009	0.094	0.020	0.009	0.095	0.020	0.009
	0.083	0.026	0.008	0.095	0.019	0.009	0.097	0.019	0.010	0.098	0.019	0.010
PLSMD <sub>T.25</sub>	0.064	0.030	0.005	0.071	0.024	0.006	0.076	0.022	0.006	0.076	0.022	0.006
	0.075	0.028	0.006	0.086	0.022	0.008	0.089	0.022	0.008	0.089	0.022	0.008
	0.080	0.028	0.007	0.089	0.023	0.008	0.093	0.023	0.009	0.093	0.023	0.009
PLSMD <sub>T.50</sub>	0.068	0.027	0.005	0.080	0.020	0.007	0.085	0.020	0.008	0.086	0.021	0.008
	0.082	0.025	0.007	0.091	0.020	0.009	0.094	0.020	0.009	0.094	0.020	0.009
	0.083	0.026	0.008	0.095	0.018	0.009	0.098	0.019	0.010	0.098	0.019	0.010
Simu. 06												
RIMD	0.009	0.021	0.001	0.035	0.015	0.001	0.042	0.014	0.002	0.042	0.014	0.002
	0.064	0.019	0.004	0.088	0.010	0.008	0.091	0.010	0.008	0.091	0.010	0.008
	0.108	0.014	0.012	0.127	0.008	0.016	0.128	0.008	0.016	0.128	0.008	0.016
TSMD	0.055	0.014	0.003	0.077	0.010	0.006	0.083	0.010	0.007	0.083	0.010	0.007
	0.066	0.016	0.005	0.089	0.009	0.008	0.093	0.009	0.009	0.093	0.009	0.009
	0.070	0.015	0.005	0.094	0.010	0.009	0.097	0.010	0.009	0.097	0.010	0.009
HMD	-0.140	0.031	0.020	-0.091	0.022	0.009	-0.077	0.022	0.006	-0.077	0.022	0.006
	-0.076	0.030	0.007	-0.029	0.019	0.001	-0.022	0.019	0.001	-0.022	0.019	0.001
	-0.027	0.025	0.001	0.008	0.016	0.000	0.015	0.016	0.000	0.015	0.016	0.000
PLSMD <sub>O.25</sub>	0.045	0.039	0.004	0.059	0.025	0.004	0.069	0.024	0.005	0.070	0.024	0.005
	0.059	0.035	0.005	0.079	0.024	0.007	0.084	0.024	0.008	0.084	0.024	0.008
	0.066	0.032	0.005	0.084	0.022	0.008	0.089	0.022	0.008	0.089	0.022	0.008
PLSMD <sub>O.50</sub>	0.054	0.030	0.004	0.077	0.021	0.006	0.083	0.020	0.007	0.084	0.019	0.007
	0.068	0.030	0.006	0.090	0.020	0.008	0.093	0.020	0.009	0.093	0.019	0.009
	0.069	0.031	0.006	0.093	0.019	0.009	0.097	0.019	0.010	0.097	0.019	0.010
PLSMD <sub>T.25</sub>	0.046	0.037	0.003	0.060	0.026	0.004	0.068	0.024	0.005	0.069	0.024	0.005
	0.059	0.036	0.005	0.079	0.023	0.007	0.083	0.024	0.007	0.084	0.024	0.008
	0.065	0.032	0.005	0.084	0.021	0.008	0.088	0.022	0.008	0.088	0.022	0.008
PLSMD <sub>T.50</sub>	0.055	0.030	0.004	0.077	0.020	0.006	0.084	0.020	0.007	0.084	0.019	0.007
	0.069	0.029	0.006	0.090	0.020	0.008	0.093	0.020	0.009	0.093	0.020	0.009
	0.069	0.031	0.006	0.093	0.019	0.009	0.097	0.019	0.010	0.097	0.019	0.010

CC, complete case analysis; HMD hybrid algorithm with missing data handling methods; KN, weighted *k*-nearest neighbors imputation method; KNM, *k*-nearest neighbors imputation method based on median value; MEAN, mean imputation method; PLSMD<sub>O.25</sub> and PLSMD<sub>O.50</sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T.25</sub> and PLSMD<sub>T.50</sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

1. **All models under Simu.01 – 08.**  
Based on Tables 2–5, the biases of the estimated path coefficients using TSMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> are comparable, no matter which missing data handling method we choose. At quantile level

0.25, the biases of PLSMD<sub>O</sub> and PLSMD<sub>T</sub> are slightly smaller than other PLSMD algorithms. The variances of RIMD and TSMD are always smaller than other PLSMD algorithms for each missing data handling method. The estimated



**Table 5.** Mean biases (MB), standard errors (SE), and mean squared errors (MSE) of the estimated path coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> under Simu.07–08 from 200 Monte Carlo replicates with sample size 500.

	CC			MEAN			KN			KNM		
	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE	MB	SE	MSE
Simu. 07												
RIMD	0.018	0.022	0.001	0.030	0.014	0.001	0.042	0.014	0.002	0.042	0.014	0.002
	0.069	0.019	0.005	0.087	0.011	0.008	0.092	0.010	0.008	0.091	0.010	0.008
	0.112	0.014	0.013	0.127	0.008	0.016	0.129	0.008	0.017	0.129	0.008	0.017
	0.062	0.016	0.004	0.074	0.010	0.006	0.083	0.010	0.007	0.083	0.010	0.007
TSMD	0.072	0.016	0.005	0.088	0.010	0.008	0.093	0.009	0.009	0.093	0.009	0.009
	0.076	0.015	0.006	0.092	0.010	0.009	0.097	0.010	0.010	0.097	0.010	0.009
	-0.123	0.031	0.016	-0.104	0.022	0.011	-0.078	0.022	0.007	-0.079	0.022	0.007
HMD	-0.063	0.031	0.005	-0.037	0.019	0.002	-0.022	0.019	0.001	-0.023	0.019	0.001
	-0.014	0.025	0.001	0.009	0.016	0.000	0.018	0.016	0.001	0.018	0.016	0.001
	0.055	0.034	0.004	0.059	0.026	0.004	0.068	0.026	0.005	0.068	0.026	0.005
PLSMD <sub>O<sub>25</sub></sub>	0.064	0.035	0.005	0.074	0.023	0.006	0.082	0.021	0.007	0.081	0.022	0.007
	0.070	0.036	0.006	0.083	0.023	0.007	0.089	0.022	0.008	0.089	0.022	0.008
	0.063	0.030	0.005	0.072	0.022	0.006	0.083	0.020	0.007	0.083	0.020	0.007
PLSMD <sub>O<sub>50</sub></sub>	0.073	0.031	0.006	0.090	0.019	0.008	0.097	0.019	0.010	0.096	0.019	0.010
	0.075	0.029	0.007	0.095	0.019	0.009	0.099	0.019	0.010	0.099	0.019	0.010
	0.056	0.033	0.004	0.059	0.026	0.004	0.068	0.026	0.005	0.068	0.025	0.005
PLSMD <sub>T<sub>25</sub></sub>	0.064	0.036	0.005	0.074	0.022	0.006	0.081	0.022	0.007	0.081	0.021	0.007
	0.069	0.037	0.006	0.084	0.023	0.008	0.089	0.022	0.008	0.089	0.022	0.008
	0.063	0.030	0.005	0.072	0.021	0.006	0.083	0.020	0.007	0.084	0.019	0.007
PLSMD <sub>T<sub>50</sub></sub>	0.072	0.031	0.006	0.089	0.019	0.008	0.096	0.019	0.010	0.096	0.019	0.010
	0.076	0.029	0.007	0.095	0.019	0.009	0.099	0.019	0.010	0.099	0.019	0.010
Simu. 08												
RIMD	-0.007	0.031	0.001	0.019	0.016	0.001	0.032	0.015	0.001	0.033	0.016	0.001
	0.052	0.024	0.003	0.084	0.010	0.007	0.089	0.009	0.008	0.089	0.010	0.008
	0.099	0.021	0.010	0.128	0.008	0.016	0.129	0.008	0.017	0.129	0.008	0.017
	0.042	0.022	0.002	0.067	0.012	0.005	0.079	0.011	0.006	0.079	0.012	0.006
TSMD	0.053	0.020	0.003	0.085	0.009	0.007	0.092	0.009	0.008	0.092	0.009	0.008
	0.058	0.021	0.004	0.088	0.010	0.008	0.094	0.010	0.009	0.094	0.010	0.009
	-0.167	0.045	0.030	-0.122	0.023	0.016	-0.093	0.024	0.009	-0.093	0.024	0.009
HMD	-0.102	0.042	0.012	-0.047	0.020	0.003	-0.031	0.020	0.001	-0.031	0.020	0.001
	-0.048	0.036	0.004	0.000	0.017	0.000	0.010	0.017	0.000	0.010	0.017	0.000
	0.028	0.046	0.003	0.032	0.026	0.002	0.048	0.026	0.003	0.051	0.025	0.003
PLSMD <sub>O<sub>25</sub></sub>	0.038	0.044	0.003	0.062	0.027	0.005	0.073	0.025	0.006	0.072	0.024	0.006
	0.047	0.048	0.004	0.072	0.024	0.006	0.081	0.025	0.007	0.080	0.025	0.007
	0.041	0.040	0.003	0.059	0.022	0.004	0.074	0.020	0.006	0.075	0.020	0.006
PLSMD <sub>O<sub>50</sub></sub>	0.057	0.039	0.005	0.090	0.019	0.008	0.097	0.017	0.010	0.097	0.019	0.010
	0.061	0.039	0.005	0.092	0.018	0.009	0.097	0.018	0.010	0.098	0.018	0.010
	0.030	0.046	0.003	0.033	0.026	0.002	0.048	0.025	0.003	0.048	0.025	0.003
PLSMD <sub>T<sub>25</sub></sub>	0.039	0.044	0.003	0.062	0.026	0.004	0.071	0.024	0.006	0.070	0.024	0.006
	0.044	0.047	0.004	0.072	0.024	0.006	0.078	0.023	0.007	0.078	0.024	0.007
	0.044	0.039	0.004	0.061	0.022	0.004	0.074	0.020	0.006	0.075	0.020	0.006
PLSMD <sub>T<sub>50</sub></sub>	0.058	0.037	0.005	0.089	0.020	0.008	0.096	0.018	0.010	0.096	0.019	0.010
	0.062	0.037	0.005	0.092	0.018	0.009	0.097	0.018	0.010	0.098	0.018	0.010

CC, complete case analysis; HMD hybrid algorithm with missing data handling methods; KN, weighted *k*-nearest neighbors imputation method; KNM, *k*-nearest neighbors imputation method based on median value; MEAN, mean imputation method; PLSMD<sub>O<sub>25</sub></sub> and PLSMD<sub>O<sub>50</sub></sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T<sub>25</sub></sub> and PLSMD<sub>T<sub>50</sub></sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

results with KN and KNM are very similar in all simulations. Except HMD, the mean biases of other PLSMD algorithms with MEAN, KN, and KNM are larger than CC, but their variances are smaller.

2. **Balanced models under Simu.01 – 04.**

According to Tables 2 and 3, the biases of the estimated path coefficients using RIMD, TSMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> are comparable. At quantile level 0.25, the biases of PLSMD<sub>O</sub> and

**Table 6.** Average correlation coefficients between the predicted latent variable scores and the defined true scores from 200 Monte Carlo replicates with sample size 500.

B1	B1		U1	
	0.1	0.2	0.1	0.2
	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM
RIMD	0.917, 0.928, 0.929, 0.929	0.905, 0.926, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.909, 0.930, 0.930, 0.930
TSMD	0.917, 0.927, 0.929, 0.929	0.905, 0.925, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.910, 0.929, 0.931, 0.931
HMD	0.871, 0.880, 0.886, 0.886	0.851, 0.874, 0.881, 0.881	0.866, 0.882, 0.885, 0.885	0.848, 0.881, 0.883, 0.883
PLSMD <sub>O<sub>0.25</sub></sub>	0.917, 0.927, 0.929, 0.929	0.904, 0.925, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.910, 0.929, 0.931, 0.931
PLSMD <sub>O<sub>0.50</sub></sub>	0.917, 0.927, 0.929, 0.929	0.905, 0.925, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.909, 0.929, 0.931, 0.931
PLSMD <sub>T<sub>0.25</sub></sub>	0.917, 0.927, 0.929, 0.929	0.905, 0.925, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.910, 0.929, 0.931, 0.931
PLSMD <sub>T<sub>0.50</sub></sub>	0.917, 0.927, 0.929, 0.929	0.905, 0.925, 0.927, 0.927	0.921, 0.931, 0.932, 0.932	0.910, 0.929, 0.931, 0.931

B2	B2		U2	
	0.1	0.2	0.1	0.2
	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM	CC, MEAN, KN, KNM
RIMD	0.911, 0.923, 0.926, 0.926	0.891, 0.919, 0.922, 0.922	0.915, 0.927, 0.929, 0.929	0.897, 0.924, 0.926, 0.926
TSMD	0.911, 0.921, 0.925, 0.925	0.891, 0.914, 0.919, 0.919	0.915, 0.926, 0.930, 0.929	0.897, 0.920, 0.925, 0.925
HMD	0.861, 0.862, 0.878, 0.878	0.831, 0.842, 0.862, 0.862	0.858, 0.882, 0.885, 0.885	0.831, 0.880, 0.882, 0.882
PLSMD <sub>O<sub>0.25</sub></sub>	0.911, 0.921, 0.925, 0.925	0.891, 0.914, 0.920, 0.920	0.915, 0.926, 0.930, 0.929	0.897, 0.920, 0.925, 0.925
PLSMD <sub>O<sub>0.50</sub></sub>	0.911, 0.921, 0.925, 0.925	0.891, 0.915, 0.920, 0.920	0.915, 0.926, 0.930, 0.929	0.897, 0.921, 0.925, 0.925
PLSMD <sub>T<sub>0.25</sub></sub>	0.911, 0.921, 0.925, 0.925	0.891, 0.914, 0.919, 0.919	0.915, 0.926, 0.930, 0.929	0.897, 0.920, 0.925, 0.925
PLSMD <sub>T<sub>0.50</sub></sub>	0.911, 0.921, 0.925, 0.925	0.891, 0.914, 0.919, 0.919	0.915, 0.926, 0.930, 0.929	0.897, 0.920, 0.925, 0.925

0.1 and 0.2, the missing rates. B1, balanced (5, 5, 5) with one missing manifest variable for each first-order latent variable. B2, balanced (5, 5, 5) with two missing manifest variables for each first-order latent variable. U1, unbalanced (4, 6, 8) with one missing manifest variable for each first-order latent variable. U2, unbalanced (4, 6, 8) with two missing manifest variables for each first-order latent variable. CC, complete case analysis. MEAN, mean imputation method. KN, weighted  $k$ -nearest neighbors imputation method. KNM,  $k$ -nearest neighbors imputation method based on median value. HMD, hybrid algorithm with missing data handling methods; PLSMD<sub>O<sub>0.25</sub></sub> and PLSMD<sub>O<sub>0.50</sub></sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T<sub>0.25</sub></sub> and PLSMD<sub>T<sub>0.50</sub></sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

PLSMD<sub>T</sub> are slightly smaller than other PLSMD algorithms. With the NMV and MR increasing, the mean biases of HMD with CC and MEAN get worse, but in the same simulation the mean biases of HMD are improved with KN and KNM when compared with CC and MEAN.

### 3. Unbalanced models under *Simu.05* – 08.

According to Tables 4 and 5, the biases of the estimated path coefficients using RIMD and HMD vary a lot from one to another. For example, the mean biases of RIMD in Table 5 with CC are 0.018, 0.069, and 0.112. The mean biases of HMD in Table 5 with CC are  $-0.123$ ,  $-0.063$ , and 0.014. The values of 0.112 and the absolute value of  $-0.123$  are much larger than 0.018 and 0.014, and also larger than the other PLSMD algorithms. In addition, the biases of HMD get improved with KN and KNM when compared with CC and MEAN.

Based on the previous conclusions, the performances of TSMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> in unbiasedness are comparable in all simulations. PLSMD<sub>O</sub> and PLSMD<sub>T</sub> not only capture the overview of structural relationships between

variables, but also have relatively smaller biases at quantile level 0.25. The performance of RIMD and TSMD in efficiency are relatively better. The performance of CC in unbiasedness seems better than that of other missing data handling methods, while its performance in efficiency is worse. HMD is more sensitive to the differences in model type (balanced or unbalanced), the NMV, MR, and missing data handling methods. RIMD performs badly in unbalanced models.

### 3.3.2 Comparisons of latent variables' prediction accuracy.

Table 6 displays the average correlation coefficients between the predicted scores of second-order latent variables and the defined true scores from 200 Monte Carlo replicates with sample size 500. This table consists of four parts. The upper left corner B1 presents the average correlation coefficients under balanced (5, 5, 5) with one missing manifest variable for each first-order latent variable. The lower left corner B2 presents the average correlation coefficients under balanced (5, 5, 5) with two missing manifest variables for each first-order latent variable. The upper right corner U1 presents the average correlation coefficients under unbalanced (4, 6, 8), with one

missing manifest variable for each first-order latent variable, and the lower right corner U2 presents the average correlation coefficients under unbalanced (4, 6, 8) with two missing manifest variables for each first-order latent variable.

Table 6 shows that, except HMD, the average correlation coefficients of all the other PLS algorithms are comparable in all models and settings. This finding indicates that RIMD, TSMD, HMD, PLSMD<sub>T</sub>, and PLSMD<sub>O</sub> have similar prediction accuracy with different missing data handling methods. HMD has relatively obvious smaller average correlation coefficients when compared with other PLS algorithms, which displays a relatively poor prediction ability when compared with other PLSMD algorithms. From CC to MEAN to KN to KNM, the average correlation coefficients become larger in all models and settings. And there is very little difference between KN and KNM. Therefore, in our simulations, both KN and KNM have a relatively better property of predicting latent variables' scores for different PLSMD algorithms when compared to CC and MEAN.

#### 4 Application to real data study

In this section we illustrate the performance of all the PLSMD algorithms using business sophistication data from the Global Innovation Index 2018 (GII 2018). GII 2018 provides detailed metrics about the innovation performance of 126 countries and economies around the world. Its 80 indicators explore a broad vision of innovation, including political environment, education, infrastructure, and business sophistication. Here, we use the data of business sophistication to investigate the performance of our algorithms.

As one of five enabling pillars of the Innovation Input Sub-Index, business sophistication (BS) tries to capture the level of business sophistication to assess how conducive firms are to innovation activity from three dimensions: knowledge workers (KW), innovation linkages (IL), and knowledge absorption (KA). Both BS and its three dimensions cannot be observed directly and belong to latent variables. However, all three dimensions have their manifest indicators or variables. We calculate the MR (%) of all the manifest variables. The first dimension, KW, has five indicators in total. They are employment in knowledge-intensive services (EI, 10.32%), firms offering formal training (FO, 28.57%), gross expenditure on R&D (GERD) performed by business enterprise (GP, 28.57%), GERD financed by business enterprise (GF, 24.60%), and females employed with advanced degrees (FE, 15.87%). The second dimension, IL, also has five indicators. They are university–industry research collaboration (UR, 5.56%), state of cluster development (SO, 5.56%), GERD financed by abroad (GB, 21.43%), joint venture/strategic

alliance deals (JV, 11.11%), and patent families filed in two offices (PF, 9.52%). The indicators of the third dimension, KA, are intellectual property payments (IP, 9.52%), high-tech imports (HI, 2.38%), ICT services imports (IS, 1.59%), foreign direct investment net inflows (FD, 0.79%), and research talent in business enterprise (RT, 34.13%). Finally, we get the BS model with three dimensions and 15 manifest indicators. (More information about indicators can be seen in the 2018 Global Innovation Index.)

Figure 2 shows the BS hierarchical latent variable model.

For  $\forall \tau \in (0, 1)$ , the BS hierarchical latent variable model in Figure 2 can be written as

$$Q_{(EI, FO, GP, GF, FE)^T}(\tau) = (l_{11}, l_{12}, l_{13}, l_{14}, l_{15})^T * KW \quad (14)$$

$$Q_{(UR, SO, GB, JV, PE)^T}(\tau) = (l_{21}, l_{22}, l_{23}, l_{24}, l_{25})^T * IL \quad (15)$$

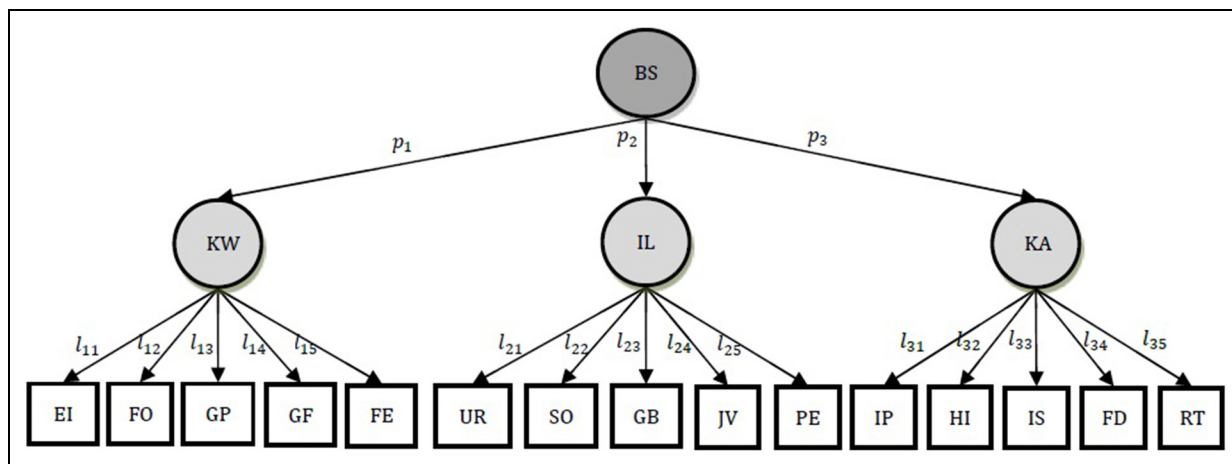
$$Q_{(IP, HI, IS, FD, RT)^T}(\tau) = (l_{31}, l_{32}, l_{33}, l_{34}, l_{35})^T * KA \quad (16)$$

$$Q_{(KW, IL, KA)^T}(\tau) = (p_1, p_2, p_3)^T * BS \quad (17)$$

Models (14)–(16) are the measurement models and model (17) is the structural model for the BS hierarchical latent variable model in Figure 2. Here, we run 200 bootstraps. Considering all 126 countries are developed to different levels, we want to investigate the overall view of relationship and apply RIMD, TSMD, HMD, PLSMD<sub>T</sub>, and PLSMD<sub>O</sub> to obtain the estimated path coefficients, loading coefficients and the scores of first-order latent variables and second-order latent variables. Raw estimation (RE), mean biases (MB), standard errors (SE), and mean squared errors (MSE) of the estimated path coefficients using different PLS algorithms from 200 Monte Carlo replicates are listed in Table 7 (here, the related results of PLSMD<sub>T</sub> and PLSMD<sub>O</sub> are at quantile levels 0.25, 0.50, and 0.75).

This table consists of nine layers. The first layer is the RIMD estimator with missing data handling methods CC, MEAN, KN, and KNM. The latter layers present the same information from the other estimators. Based on Table 7, we find that both RIMD and TSMD have relatively smaller estimated SE than other PLSMD methods with all missing data handling methods except CC. For all the PLSMD algorithms, the estimated SE of CC are obviously larger than MEAN, KN, and KNM. These conclusions indicate that the performance of CC in efficiency is much worse than other missing data handling methods, no matter which PLSMD algorithm we choose.

According to the RE, we can find the most important factor (first-order latent variable) for BS. The most important factor with CC is very different from other missing data handling methods for each PLSMD algorithm. The most important factor with KN and KNM are the same for each PLSMD algorithm. The most important factor with MEAN



**Figure 2.** BS hierarchical latent variable model.

BS, business sophistication; KW, knowledge workers; IL, innovation linkages; KA, knowledge absorption; EI, employment in knowledge-intensive services; FO, firms offering formal training; GP, gross expenditure on R&D (GERD) performed by business enterprise; GF, GERD financed by business enterprise; FE, females employed with advanced degrees; UR, university–industry research collaboration; SO, state of cluster development; GB, GERD financed by abroad; JV, joint venture/strategic alliance deals; PF, patent families filed in two offices; IP, intellectual property payments; HI, high-tech imports; IS, ICT services imports; FD, foreign direct investment net inflows; RT, research talent in business enterprise.

is different from KN and KNM using  $PLSMD_O$  and  $PLSMD_T$  at quantile levels 0.50 and 0.75. The most important factor is the second first-order latent variable IL using RIMD, TSMD, HMD, and  $PLSMD_{T_{25}}$  with the missing data handling methods MEAN, KN, and KNM. The most important factor using  $PLSMD_O$  at quantile level 0.25 is KW with MEAN, KN, and KNM. By  $PLSMD_{O_{50}}$ , IL (0.935), KW (0.930), and KW (0.935) are the most important factors for MEAN, KN and KNM, respectively. By  $PLSMD_{O_{50}}$  and  $PLSMD_{T_{50}}$ , IL (0.935 and 0.935), KW (0.930 and 0.976), and KW (0.935 and 0.975) are the most important factors for MEAN, KN, and KNM, respectively. By  $PLSMD_{O_{75}}$  and  $PLSMD_{T_{75}}$ , KW (0.936 and 0.950), KA (1.057 and 0.974), and KA (0.998 and 1.002) are the most important factors for MEAN, KN, and KNM respectively. According to all linear regression type PLSMD algorithms and missing data handling methods, all the countries should pay more attention to IL. If we want to investigate the most important factors for BS at different levels, both  $PLSMD_O$  and  $PLSMD_T$  will give us the overview of the structural relationship between variables. Therefore, different countries can focus on different important factors according to the RE in Table 7. For brevity, the estimated loading coefficients and predicted latent variable scores are omitted here, but available upon request from the author.

## 5 Discussion

In this paper, we investigate two kinds of missing data problems in hierarchical latent variable models: latent

variables and missing manifest variables. We compare five PLSMD algorithms with consideration of four missing manifest variable handling methods through simulation studies. For missing manifest variables, complete case analysis (CC), mean value replacement (MEAN), weighted  $k$ -nearest neighbors imputation method (KN), and  $k$ -nearest neighbors imputation method based on median value (KNM) are very popular missing data handling methods. For latent variables without any direct observations, the three well-known PLSMD algorithms (RIMD, TSMD, HMD), which are based on a linear regression type model, cannot be used to investigate the overall view of the structural relationship between variables at different levels. Hence, we modify the PLS procedure by quantile regression and investigate two PLSMD algorithms ( $PLSMD_O$  and  $PLSMD_T$ ) based on a hierarchical latent variable model through simulations.

According to the simulation investigation, we find the following conclusions. (1) In all settings, the performance of TSMD,  $PLSMD_O$ , and  $PLSMD_T$  in unbiasedness are comparable. But  $PLSMD_O$  and  $PLSMD_T$  have relatively smaller biases at quantile level 0.25. (2) In all eight simulations, RIMD and TSMD perform relatively better than other PLSMD algorithms in SE with different missing data handling methods. (3) RIMD performs badly in unbalanced models. (4) HMD is more sensitive to the difference in model type (balanced or unbalanced), the NMV, MR, and missing data handling methods. (5) The performance of CC in unbiasedness seems better than other missing data handling methods, while its performance in efficiency is worse. (6) HMD has relatively poor prediction ability

**Table 7.** Raw estimation (RE), standard errors (SE), lower confidence limit (LCL) and upper confidence limit (UCL) of the estimated path coefficients using RIMD, TSMD, HMD, PLSMD<sub>O</sub> and PLSMD<sub>T</sub> with the missing data handling methods CC, MEAN, KN, and KNM from 200 Monte Carlo replicates.

	CC			MEAN			KN			KNM			
	RE	SE	UCL	LCL	UCL	RE	SE	LCL	UCL	RE	SE	LCL	UCL
RIMD	0.944	0.177	0.888	0.920	0.890	0.022	0.886	0.890	0.914	0.023	0.908	0.912	0.911
	0.885	0.308	0.782	0.838	0.934	0.017	0.929	0.932	0.929	0.019	0.925	0.928	0.925
	0.893	0.392	0.772	0.843	0.857	0.020	0.857	0.861	0.892	0.019	0.890	0.893	0.896
TSMD	0.904	0.049	0.894	0.903	0.885	0.016	0.884	0.887	0.901	0.018	0.897	0.901	0.899
	0.825	0.087	0.794	0.809	0.909	0.017	0.904	0.908	0.914	0.017	0.909	0.913	0.909
	0.870	0.226	0.792	0.833	0.877	0.023	0.870	0.874	0.900	0.019	0.893	0.896	0.895
HMD	0.806	0.259	0.723	0.770	0.714	0.039	0.708	0.715	0.742	0.043	0.737	0.745	0.749
	0.870	0.283	0.746	0.798	0.827	0.023	0.825	0.829	0.845	0.022	0.841	0.845	0.842
	0.829	0.210	0.759	0.797	0.621	0.051	0.632	0.642	0.747	0.049	0.747	0.756	0.762
PLSMD <sub>O<sub>25</sub></sub>	0.866	0.185	0.799	0.833	0.900	0.045	0.897	0.906	0.929	0.046	0.892	0.901	0.897
	0.217	0.342	0.300	0.362	0.879	0.073	0.891	0.905	0.821	0.087	0.843	0.858	0.860
	0.890	0.094	0.886	0.904	0.883	0.052	0.868	0.877	0.861	0.053	0.848	0.858	0.860
PLSMD <sub>O<sub>50</sub></sub>	0.896	0.090	0.887	0.904	0.898	0.043	0.891	0.899	0.930	0.044	0.909	0.917	0.911
	0.423	0.399	0.421	0.493	0.935	0.050	0.920	0.929	0.925	0.041	0.903	0.910	0.901
	0.958	0.147	0.852	0.879	0.874	0.069	0.844	0.857	0.907	0.051	0.907	0.916	0.915
PLSMD <sub>O<sub>75</sub></sub>	1.098	0.158	1.001	1.029	0.936	0.056	0.924	0.934	0.929	0.065	0.942	0.954	0.971
	0.866	0.251	0.766	0.811	0.924	0.038	0.922	0.929	0.939	0.038	0.935	0.942	0.941
	0.891	0.277	0.715	0.765	0.902	0.092	0.942	0.959	1.057	0.074	0.978	0.992	0.983
PLSMD <sub>T<sub>25</sub></sub>	0.901	0.186	0.819	0.853	0.830	0.059	0.854	0.865	0.882	0.065	0.869	0.881	0.878
	0.202	0.344	0.332	0.395	0.877	0.065	0.871	0.883	0.900	0.122	0.829	0.852	0.854
	0.882	0.214	0.821	0.860	0.740	0.097	0.716	0.734	0.761	0.081	0.746	0.761	0.756
PLSMD <sub>T<sub>50</sub></sub>	0.937	0.071	0.905	0.918	0.918	0.040	0.918	0.926	0.976	0.043	0.947	0.955	0.948
	0.437	0.332	0.467	0.527	0.935	0.062	0.920	0.931	0.909	0.047	0.891	0.900	0.895
	1.000	0.257	0.807	0.854	0.877	0.077	0.814	0.828	0.892	0.062	0.870	0.881	0.881
PLSMD <sub>T<sub>75</sub></sub>	1.118	0.159	0.990	1.019	0.950	0.057	0.941	0.952	0.960	0.056	0.967	0.978	0.991
	0.949	0.232	0.778	0.820	0.925	0.035	0.927	0.933	0.956	0.035	0.946	0.952	0.949
	0.952	0.323	0.823	0.881	0.918	0.092	0.967	0.984	0.974	0.077	1.007	1.021	1.013

CC, complete case analysis; HMD hybrid algorithm with missing data handling methods; KN, weighted k-nearest neighbors imputation method; KNM, k-nearest neighbors imputation method based on median value; MEAN, mean imputation method; PLSMD<sub>O<sub>25</sub></sub>, one-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; PLSMD<sub>T<sub>25</sub></sub> and PLSMD<sub>T<sub>50</sub></sub>, two-stage PLS algorithm with missing data handling methods at quantile levels 0.25 and 0.50; RIMD, repeated indicators algorithm with missing data handling methods; TSMD, two-step algorithm with missing data handling methods.

when compared with other PLSMD algorithms. However, RIMD, TSMD, HMD, PLSMD<sub>O</sub>, and PLSMD<sub>T</sub> are comparable in latent variable score prediction accuracy with different missing data handling methods.

In our research, the type of manifest variables we focus on are continuous variables for hierarchical latent variable models. In the future, we will make more efforts with other types of manifest variables, such as categorical data or other kinds of discrete data, and consider more complex data problems.<sup>34</sup> We will apply our methods to more real data problems.<sup>35–40</sup> Another future work is about investigating how different PLSMD algorithms can be used in longitudinal data analysis or dynamic structural equation models. Last, but not least, all the PLSMD algorithms should be expanded to large-scale data.<sup>41</sup>


### Acknowledgements

The author is grateful to the reviewers and editors for their many helpful comments and suggestions. In addition, the author wants to thank his father and mother's upbringing and support since he was born, and his wife Yujie Liu's patience, care, and love.

### Funding

The author gratefully acknowledges The Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (16XNH102).

### ORCID iD

Hao Cheng  <https://orcid.org/0000-0003-1143-9601>

### References

- Merkle EC, Furr D and Rabe-Hesketh S. Bayesian comparison of latent variable models: conditional versus marginal likelihoods. *Psychometrika* 2019; 84(3): 802–829.
- Xie YF, Ma MD, Zhang YN, et al. Factors associated with health literacy in rural areas of central china: structural equation model. *BMC Health Serv Res* 2019; 19(1): 1–8.
- Zhao ZG, Wang PL, Li QH, et al. Input trajectory adjustment within batch runs based on latent variable models. *Ind Eng Chem Res* 2019; 58(34): 15562–15572.
- Cheng H, Yi DH, Si JS, et al. Establishment of comprehensive indicators in TCM pectoral-qi with experts diagnosis and self-test technology. *Medicine* 2018; 97(7): e9916.
- Becker JM, Klein K and Wetzels M. Formative hierarchical latent variable models in PLS-SEM: recommendations and guidelines. *Long Range Plann* 2012; 45: 359–394.
- Bollen KA. *Structural equations with latent variables*. New York: Wiley, 1989.
- Sammel MD and Ryan LM. Latent variable models with fixed effects. *Biometrics* 1996; 52(2): 650–663.
- Little RJA and Rubin DB. *Statistical analysis with missing data*. New York: Wiley, 1987
- Wei Y, Ma Y and Carroll RJ. Multiple imputation in quantile regression. *Biometrika* 2012; 99: 423–438.
- Kim JK. Parametric fractional imputation for missing data analysis. *Biometrika* 2011; 98: 119–132.
- Cheng H and Wei Y. A fast imputation algorithm in quantile regression. *Comput Stat* 2018; 33(4): 1017–1036.
- Cheng H. An application research of inverse probability weighted multiple imputation method on factors of residents income in China. *Statistics and Information Forum* 2019; 7: 26–34.
- Lohmöller JB. *Latent variable path modeling with partial least squares*. Heidelberg: Physica-Verlag, 1989.
- Wetzels M, Schröder GO and Oppen CV. Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. *MIS Q* 2009; 33(1): 177–195.
- Robert WG, Bruce RK and Herman OAW. Partial least squares path modeling with latent variables. *Anal Chim Acta* 1979; 112(4): 417–421.
- Hair JF, Hult GTM, Ringle CM, et al. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks: SAGE Publications, 2014.
- Wold H. Soft modeling: the basic design and some extensions. In: Jöreskog KG and Wold H (eds) *Systems under indirect observations: part II*. Amsterdam: North-Holland, 1982, pp.1–54
- Ciavolino E and Al-Nasser AD. Comparing generalized maximum entropy and partial least squares methods for structural equation models. *J Nonparametr Stat* 2009; 21(8): 1017–1036.
- Claes C, Peter H and Anders HW. Robustness of partial least squares method for estimating latent variable quality structures. *J Appl Stat* 1999; 26(4): 435–446.
- Esposito VV, Chin WW, Henseler J, et al. *Handbook of partial least squares. concepts, methods and applications*. New York: Springer, 2010.
- Fornell C and Bookstein FL. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *J Market Res* 1982; 19(4): 440–452.
- Guinot C, Latreille J and Tenenhaus M. PLS path modeling and multiple table analysis: application to the cosmetic habits of women in Ile-de-France. *Chemom Intell Lab Syst* 2001; 58(2): 247–259.
- Henseler J, Ringle CM and Sinkovics RR. The use of partial least squares path modeling in international marketing. In: Sinkovics RR and Ghauri PN (eds) *Advances in International Marketing* Bingley: Emerald Publishing, 2009, pp.277–320.
- Henseler J, Ringle CM and Sarstedt M. Using partial least squares path modeling in international advertising research: basic concepts and recent issues. In: Okazaki S (ed) *Handbook of research in international advertising*. Cheltenham: Edward Elgar Publishing, 2012, pp.252–276.
- Sanchez G. *PLS path modeling with R*. Berkeley, CA: Trowchez Editions, 2013.
- Ciavolino E and Nitti M. Simulation study for PLS path modeling with high-order construct: a job satisfaction model evidence. In: *Advanced dynamic modeling of economic and social systems*. Berlin: Springer, 2013, pp.185–207.
- Tenenhaus M. l'approche PLS. *Revue de Statistique Applique* 1999; 47(2): 5–40.
- Tenenhaus M, Esposito Vinzi V, Chatelin YM, et al. PLS path modeling. *Comput Stat Data Anal* 2005; 48: 159–205.

29. Chatelin YM, Vinzi Esposito V and Tenenhaus M. State-of-art on PLS path modeling through the available software, [www.hec.fr/Recherche/Cahiers-de-recherche/State-of-arton-PLS-Path-Modeling-through-the-available-software](http://www.hec.fr/Recherche/Cahiers-de-recherche/State-of-arton-PLS-Path-Modeling-through-the-available-software) (2002).
30. Ringle CM, Wende S and Becker JM. *SmartPLS 3*. Boenningstedt: SmartPLS GmbH, 2015.
31. Ciavolino E and Nitti M. Using the hybrid two-step estimation approach for the identification of second-order latent variable models. *J Appl Stat* 2013; 40(3): 508–526.
32. Koenker R and Bassett GJ. Regression quantiles. *Econometrica* 1978; 46: 33–50.
33. Koenker R. *Quantile regression*. Cambridge: Cambridge University Press, 2005.
34. Cui R, Bucur IG, Groot P, et al. A novel Bayesian approach for latent variable modeling from mixed data with missing values. *Stat Comput* 2019; 29(5): 977–993.
35. Hu J, Zhang WQ, Xing F, et al. Research on the measurement and evaluation of national economic and social development from the perspective of the Belt and Road Initiative. *Statistics and Information Forum* 2018; 6: 43–53.
36. Xia WL and Ding PQ. Establishment of provincial innovation and entrepreneurship environment evaluation indicators system: evaluate the 31 provinces of China. *Statistics and Information Forum* 2017; 4: 63–72.
37. Zhou YD, Meng XC and Yu ZQ. Evaluation index system of the five-pronged approach “Five in One” for the regional comprehensive development. *Statistics and Information Forum* 2018; 5: 19–25.
38. Sun XD and Ni RX. Onboard attributes/criteria of cruise ships and cruisers satisfaction evaluation. *Statistics and Information Forum* 2017; 10: 116–122.
39. Liu M and Chen Z. Research on indicator system for measuring E-commerce. *Statistics and Information Forum* 2008; 7: 20–28.
40. Yuan XL, Jing XJ, Zhao ZH, et al. The construction of evaluation system of regional economic growth quality: empirical analysis based on Shaanxi province data. *Statistics and Information Forum* 2017; 6: 42–47.
41. Kang Q. Financial risk assessment model based on big data. *International Journal of Modeling, Simulation, and Scientific Computing* 2019; 10(4): 106–113.

### Author biography

**Hao Cheng** is an assistant research fellow at the National Academy of Innovation Strategy, China Association for Science and Technology. He obtained his PhD from Renmin University of China and has visited Columbia University, Needham Research Institute of Cambridge University, the London School of Economics and Political Science, and ISCTE-IUL as a visiting scholar.