Determining the number of components in PLS regression on incomplete data set

Titin Agustin Nengsih, Frédéric Bertrand, Myriam Maumy-Bertrand, Nicolas Meyer

Titin Agustin Nengsih[1,2] / Frédéric Bertrand[1,a] / Myriam Maumy-Bertrand[1] / Nicolas Meyer[2,3]

Determining the number of components in PLS regression on incomplete data set

[1] IRMA, CNRS UMR 7501, Université de Strasbourg, 67084 Strasbourg, Cedex, France, E-mail: frederic.bertrand1@utt.fr. https://orcid.org/0000-0002-0837-8281.
[2] iCUBE, CNRS UMR 7357, Université de Strasbourg, 67400 Strasbourg, France
[3] GMRC, Public Health Department, Strasbourg University Hospital, Strasbourg, France

Abstract:
Partial least squares regression – or PLS regression – is a multivariate method in which the model parameters are estimated using either the SIMPLS or NIPALS algorithm. PLS regression has been extensively used in applied research because of its effectiveness in analyzing relationships between an outcome and one or several components. Note that the NIPALS algorithm can provide estimates parameters on incomplete data. The selection of the number of components used to build a representative model in PLS regression is a central issue. However, how to deal with missing data when using PLS regression remains a matter of debate. Several approaches have been proposed in the literature, including the Q^2 criterion, and the *AIC* and *BIC* criteria. Here we study the behavior of the NIPALS algorithm when used to fit a PLS regression for various proportions of missing data and different types of missingness. We compare criteria to select the number of components for a PLS regression on incomplete data set and on imputed data set using three imputation methods: multiple imputation by chained equations, *k*-nearest neighbour imputation, and singular value decomposition imputation. We tested various criteria with different proportions of missing data (ranging from 5% to 50%) under different missingness assumptions. Q^2-leave-one-out component selection methods gave more reliable results than *AIC* and *BIC*-based ones.

Keywords: imputation method, missing data, NIPALS, number of components, PLS regression

MSC 2010: 62G08, 68U20, 65C60

DOI: 10.1515/sagmb-2018-0059

1 Introduction

Missing data are present in many real-world data set and often cause problems in data analysis. Missing data can occur for many reasons, including uncollected data, mishandled samples, equipment errors, measurement errors, misunderstanding questionnaires, etc. (Grung & Manne, 1998; Folch-Fortuny, Arteaga & Ferrer, 2016). According to Little and Rubin (2002), missing data can be divided into three categories, namely: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). If the probability that the data is known depends neither on the observed value nor on the missing values, the data are said to be MCAR. In the case of MAR, missingness depends only on the values of the observed data. Lastly, data are said to be MNAR if missingness depends both on the observed and missing data values.

Many methods have been proposed for imputing missing data. The simplest ones rely on single value imputation, e.g. the mean over the complete cases in the study sample – known as mean imputation (Troyanskaya et al., 2001). More complex methods include regression-based imputation (Horton & Lipsitz, 2001), imputation based on Non-linear Iterative PArtial Least Squares (NIPALS) (Tenenhaus, 1998), multiple imputation (Rubin, 1987), *K*-Nearest Neighbours imputation (KNNimpute) (Dixon, 1979), Singular Value Decomposition-based imputation (SVDimpute) (Troyanskaya et al., 2001), and so on.

Partial least squares (PLS) regression was introduced in the 1970s by Wold (1966). It has gone from being popular in chemometrics (see Wold, Sjöström & Eriksson, 2001) to being commonly used in many research areas such as bioinformatics (Nguyen & Rocke, 2004), medicine (Yang et al., 2017), social sciences (Sawatsky, Clyde & Meek, 2015), and spectroscopy (Oleszko et al., 2017). PLS regression – in its classical form – is based on the NIPALS algorithm. The alternative estimation method for PLS regression is SIMPLS algorithm, for *Straightforward Implementation of a statistically inspired Modification to PLS* (see De Jong, 1993). The former has been implemented

Frédéric Bertrand is the corresponding author.
[a] **Present address:** ICD, CNRS FRE 2019, ROSAS, M2S, Université de Technologie de Troyes, 10004 Troyes, Cedex, France

in software such as SIMCA (Eriksson et al., 2002) and more recently in the `plsRglm` package (Bertrand, Meyer & Maumy-Bertrand, 2014).

The NIPALS algorithm was initially devised to carry out principal component analysis (PCA) on incomplete data set. It explains why its reliability under increasing proportions of missing data has been studied mainly in this setting (Nelson, Taylor & MacGregor, 1996; Grung & Manne, 1998; Arteaga & Ferrer, 2002). As in PCA, one of the justifications for using the NIPALS algorithm in PLS regression is that it enables models to be fitted on incomplete data set. This feature is long-known and frequently used as an argument to apply this algorithm preferentially. In this paper, we focus on univariate PLS, also known as PLS1, and the use of the NIPALS algorithm.

The goal of PLS regression is to predict a set of dependent variables from a set of independent variables. This prediction is obtained by extracting from the predictors a set of orthogonal factors called components that have the best predictive power. Determining the optimal number of components is thus a critical problem in PLS regression. Selecting a less-than-optimal number of components leads to a loss of information, whereas selecting a more-than-optimal number can lead to models with poor predictive ability (Wiklund et al., 2007).

Several papers have studied ways to determine the number of components to retain in the final PLS regression (see for instance Lazraq, Cléroux & Gauchi, 2003). In spectroscopy, for example, the sample spectrum is the sum of the spectra of the constituents multiplied by their concentration in the sample. This interpretation makes sense with the data explained by several components in PLS regression. As also mentioned by Goicoechea and Olivieri (1999a), the analyse of interest is embedded in a complex mixture of several components. Another interesting application of the number of components is the use of microscopic concepts such as molecules and reactions in chemical and biological data which closely correspond with the use of a number of components. This has been discussed by Burnham, Macgregor, and Viveros (1999), Burnham, Viveros, and Macgregor (1996), and Kvalheim (1992).

Though it is now considered a benchmark for incomplete data set analysis, the reliability of the NIPALS algorithm when estimating PLS regression parameters on incomplete data sets has been studied very little, despite its importance. In the context of PLS regression, details on to missing data, such as how to estimate scores on incomplete data, and the impact of missing data on PLS prediction have been reported by Nelson, Taylor, and MacGregor (1996), Rännar et al. (1995), and Serneels and Verdonck (2008). However, the sensitivity of the NIPALS algorithm to increasing missing data proportions does not seem to have been given much attention. Moreover, in the few papers that pertain to incomplete data issues, the reliability of the NIPALS algorithm under different missingness mechanisms described in Little and Rubin (1987) has been systematically ignored.

In summary, at least two things may affect parameter estimates for PLS regression: the proportion of missing data and the type of missingness. Both issues will be studied here.

Besides, we compare criteria for selecting the number of components in PLS regression. The most-used method for incomplete data sets is PLS regression with the NIPALS algorithm. Other methods for imputing data sets are multiple imputation by chained equations (MICE), *K*-Nearest Neighbours imputation, and Singular Value Decomposition imputation. The influence of the proportion of missing data and the type of missingness on the estimation of the number of components in a PLS regression is the primary purpose of the present study.

The paper is organized as follows. Section 2 describes the methods, presenting a brief description of PLS regression, cross-validation with missing values, and imputation methods. Section 3 describes the simulation study, and Section 4 gives the results of this study. Real data are presented in Section 5. We conclude with a general discussion in Section 6.

2 Partial least squares regression and related works

2.1 PLS regression

A complete description of PLS regression can be found in Wold (1966) and Höskuldsson (1988).

Suppose that $\mathbf{X}$ is an $n \times p$ data matrix of continuous p explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p$ where p can be greater than n (the number of observations) and $\mathbf{y}$ is an univariate response variable.

PLS1 replaces the OLS goal of finding $\boldsymbol{\beta}$ that maximizes squared correlation $\mathrm{cor}(\mathbf{X}\boldsymbol{\beta}, \mathbf{y})^2$ with an alternative goal of finding $\boldsymbol{\beta}$ with length $\|\boldsymbol{\beta}\| = 1$ maximizing covariance

$$\mathrm{cov}(\mathbf{X}\boldsymbol{\beta}, \mathbf{y})^2 = \mathrm{cor}(\mathbf{X}\boldsymbol{\beta}, \mathbf{y})^2 \times \mathrm{var}(\mathbf{X}\boldsymbol{\beta})$$

which effectively penalizes directions of low variance.

The PLS1 regression model is:

$$\mathbf{y} = \mathbf{Tq} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{T}$ is the $(n \times H)$ matrix of the H extracted score vectors (with columns $\mathbf{t}_h = (t_{1h}, \ldots, t_{nh})'$, $h = 1, \ldots, H$. -A' means the transpose of the $\mathbf{A}$ matrix or vector-), $\mathbf{q}$ is a vector -of loadings- of length (n), and $\boldsymbol{\epsilon}$ is the error vector.

As stated before, the PLS regression can be included in an objective function framework (Burnham, Viveros & Macgregor, 1996) and viewed as the solution to an iterated maximization problem. The first component is the solution to:

$$\max_{\mathbf{w}_1} \operatorname{Cov}(\mathbf{Xw}_1, \mathbf{y})^2 = \max_{\mathbf{w}_1} \mathbf{w}_1'\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w}_1 \quad \text{subject to} \quad \mathbf{w}_1'\mathbf{w}_1 = 1. \tag{2}$$

and the maximum for (2) is obtained at $\mathbf{w_1}$, the largest eigenvector of the matrix $\mathbf{X'yy'X}$. In order to obtain further weight vectors, the algorithm is repeated with deflated $\mathbf{X}$-matrix and $\mathbf{y}$-vector. The deflation process is defined for $i = 1, 2, \ldots, h-1$ as

$$\mathbf{X}_{i+1} = \left(\mathbf{I} - \frac{\mathbf{t}_i\mathbf{t}_i'}{\mathbf{t}_i'\mathbf{t}_i}\right)\mathbf{X}_i, \quad \mathbf{y}_{i+1} = \left(\mathbf{I} - \frac{\mathbf{t}_i\mathbf{t}_i'}{\mathbf{t}_i'\mathbf{t}_i}\right)\mathbf{y}_i. \tag{3}$$

where $\mathbf{X_1} = \mathbf{X}$, $\mathbf{y_1} = \mathbf{y}$ and $\mathbf{t}_i = \mathbf{Xw}_i$. Hence only the subspace in $\mathbf{X}$ that is orthogonal to the earlier linear combinations developed in the $\mathbf{X}$-space is used. The $\mathbf{y}$-space is projected onto the space orthogonal to the previous $\mathbf{X}$-components. Subsequent weights vectors $\mathbf{w_i}$ are chosen to satisfy (2) using deflated $\mathbf{X_i}$-matrices and $\mathbf{y_i}$-vectors in place of the original $\mathbf{X}$-matrix and $\mathbf{y}$-vector.

The objective function can be updated to include the iterative deflation steps as shown in (Burnham, Viveros & Macgregor, 1996) for both univariate and multivariate PLS:

$$\max_{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i} \left(\boldsymbol{\alpha}_i'\mathbf{X}'\mathbf{y}\boldsymbol{\beta}_i - \sum_{j=1}^{i-1} \frac{(\boldsymbol{\alpha}_i'\mathbf{X}'\mathbf{X}\mathbf{r}_j)(\mathbf{r}_j'\mathbf{X}'\mathbf{y}\boldsymbol{\beta}_i)}{\mathbf{r}_j'\mathbf{X}'\mathbf{X}\mathbf{r}_j} \right) \quad \text{subject to } \boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1, \boldsymbol{\beta}_i'\boldsymbol{\beta}_i = 1$$

where $\mathbf{Xr}_j$ is given by the Gram-Schmidt formula for determining an orthogonal basis for a set of vectors. As a consequence, $\mathbf{Xr}_j$ can be expressed as

$$\mathbf{Xr}_j = \mathbf{X}\boldsymbol{\alpha}_j - \sum_{k=1}^{j-1} \mathbf{Xr}_k \frac{\boldsymbol{\alpha}_j'\mathbf{X}'\mathbf{X}\mathbf{r}_k}{\mathbf{r}_k'\mathbf{X}'\mathbf{X}\mathbf{r}_k}$$

In addition, in our univariate response setting, we must have $\boldsymbol{\beta}_i = \beta_i = 1$. This leads to a simplified problem:

$$\max_{\boldsymbol{\alpha}_i} \left(\boldsymbol{\alpha}_i'\mathbf{X}'\mathbf{y} - \sum_{j=1}^{i-1} \frac{(\boldsymbol{\alpha}_i'\mathbf{X}'\mathbf{X}\mathbf{r}_j)(\mathbf{r}_j'\mathbf{X}'\mathbf{y})}{\mathbf{r}_j'\mathbf{X}'\mathbf{X}\mathbf{r}_j} \right) \quad \text{subject to } \boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1 \tag{4}$$

In its classical form and with complete data, the PLS model fitting process (Rosipal & Krämer, 2005) is based on the nonlinear iterative partial least squares (NIPALS) algorithm that finds a sequence of weights vector $(\mathbf{w}_i)_{1,\ldots,i}$ solution to (4).

As a result, this fitting process for the PLS regression finds iteratively the new variables $\mathbf{t_h}$. They are called components and are mutually orthogonal. The number of components H can be chosen from data through cross validation, a model selection criterion (see next Section 2.3). The $\mathbf{t_h}$ PLS components are linear combinations of the p variables of the matrix $\mathbf{X}$ with formulas :

$$\mathbf{T} = \mathbf{XW} \tag{5}$$

where $\mathbf{W}$ is a $p \times H$ matrix of weights. The columns of $\mathbf{W}$ are denoted as $\mathbf{w}_h = (w_{1h}, \ldots, w_{ph})'$, respectively, for $h = 1, \ldots, H$.

2.2 NIPALS-PLSR

In the chapter 6 of Tenenhaus book (Tenenhaus, 1998), the NIPALS algorithm divides the data set into a complete part and an incomplete part to build a matrix, called $\mathbf{X_h}$. The columns of the matrix $\mathbf{X_h}$ are noted $\mathbf{x}_{h1}, \ldots, \mathbf{x}_{hj}, \ldots, \mathbf{x}_{hp}$.

The PLS regression which has been estimated by NIPALS algorithm (NIPALS-PLSR) starts with (optionally) transformed, scaled, and centered $\mathbf{X}$ and $\mathbf{y}$. The $\mathbf{t_h}$ component is equal to Xw_h, where the weight $\mathbf{w_h}$ (see below in the step 4) is constructed step by step. The steps of NIPALS-PLSR follows in Algorithm Algorithm 1.

Algorithm 1 (NIPALS-PLSR algorithm in complete data set (Tenenhaus, 1998)).

Initialize $\mathbf{X}_0 = \mathbf{X}, \mathbf{y}_0 = \mathbf{y}$
for $h = 1$ to H **do**
  repeat
    $\mathbf{w}_h = \mathbf{X}'_{h-1}\mathbf{y}_{h-1}/\mathbf{y}'_{h-1}\mathbf{y}_{h-1}$
    Normalize $\mathbf{w_h}$ to 1
    $\mathbf{t}_h = \mathbf{X}_{h-1}\mathbf{w}_h/\mathbf{w}'_h\mathbf{w}_h$
    $\mathbf{p}_h = \mathbf{X}'_{h-1}\mathbf{t}_h/\mathbf{t}'_h\mathbf{t}_h$
    $\mathbf{X}_h = \mathbf{X}_{h-1} - \mathbf{t}_h\mathbf{p}'_h$
    $q_h = \mathbf{y}'_{h-1}\mathbf{t}_h/\mathbf{t}'_h\mathbf{t}_h$
    $\mathbf{y}_h = \mathbf{y}_{h-1} - \mathbf{t}_h q_h$
  until *Convergence of* $\boldsymbol{p}_h$
endfor

The NIPALS-PLSR algorithm can, therefore, be seen as a compromise between a multiple linear regression and a principal component analysis (Wold, Esbensen & Geladi, 1987), in which the first h components $\mathbf{t_h}$ are the principal components whose covariances with $\mathbf{y}$ are the largest.

There is an additional interest to use the NIPALS algorithm in PLS regression in the presence of incomplete data. Treatment of missing data with NIPALS-PLSR can be implicitly associated with a simple imputation method (Bastien & Tenenhaus, 2003). When data in any column or row of the matrix $\mathbf{X}$ is missing, the iterative regressions are performed using the available values, ignoring the missing ones (Tenenhaus, 1998).

The iteration of the NIPALS-PLSR for the h^{th} component follows in Algorithm Algorithm 2.

Algorithm 2 (NIPALS-PLSR algorithm in incomplete data set).

Initialize $\mathbf{X}_0 = \mathbf{X}, \mathbf{y}_0 = \mathbf{y}$
for $h = 1$ to H **do**
  repeat
    $w_{j,h} = \frac{\sum_{\{i:x_{ij,h} \text{ and } y_{i,h} exist\}} x_{ij,h}y_{i,h}}{\sum_{\{i:x_{ij,h} \text{ and } y_{i,h} exist\}} y^2_{i,h}}, j = 1, \ldots, p$
    Normalize $\mathbf{w_h}$ to 1
    $t_{i,h} = \frac{\sum_{\{j:x_{ij,h} \ exists\}} x_{ij,h}w_{j,h}}{\sum_{\{j:x_{ij,h} \ exists\}} w^2_{j,h}}, i = 1, \ldots, n$
    $p_{j,h} = \frac{\sum_{\{j:x_{ij,h} \ exists\}} x_{ij,h}t_{i,h}}{\sum_{\{j:x_{ij,h} \ exists\}} t^2_{i,h}}, j = 1, \ldots, p$
    $\mathbf{X}_h = \mathbf{X}_{h-1} - \mathbf{t}_h\mathbf{p}'_h$ for $x_{ij,h}$ existing
    $q_h = \frac{\sum_{\{i:y_{i,h} \ exists\}} y_{i,h}t_{i,h}}{\sum_{\{i:y_{i,h} \ exists\}} t^2_{i,h}}$
    $\mathbf{y}_h = \mathbf{y}_{h-1} - \mathbf{t}_h q_h$ for $y_{i,h}$ existing
  until *Convergence of* $\boldsymbol{p}_h$
endfor

2.3 Model selection: cross-validation and information criteria

Several papers have studied methods to determine the number of components to retain in the final model of PLS regression – see, for instance, Lazraq, Cléroux, and Gauchi (2003). In this present study, only selection of the number of components is considered, including all $\mathbf{x_j}$ variables on each of the first components, whatever their significance for these components. There exist several approaches in the literature to choose the h number of components to include in the final model: the Q^2 criterion, computed by cross-validation (Stone, 1974) and

information criteria like the Akaike Information Criterion (*AIC*) (Akaike, 1969) and the Bayesian Information Criterion (*BIC*) (Schwarz, 1978). We are going to describe them briefly below.

Cross-validation is a practical and reliable way to test this predictive significance (Wakeling & Morris, 1993; Tenenhaus, 1998). It has become the standard in PLS regression analysis and incorporated in one form or another in all available PLS regression software. Cross-validation is performed by dividing the complete data set in k complete subsets, and then developing several parallel models from reduced data with one of the groups deleted. It is called k-fold cross-validation. Five- to ten-fold cross-validation are common. If $k = n$, this approach is called leave-one-out (LOO) cross-validation.

If we study an incomplete data set, so cross-validation requires modifications. There are two methods: *standard* cross-validation or *adaptive* cross-validation. The first, called *standard* cross-validation, is to predict the response value of any row of the data set as if it featured missing data. In *adaptive* cross-validation, it predicts the response value for a row accordingly to the presence of missing data in that row: regular prediction for complete data if no missing data in the row, missing data specific prediction if there are missing data in the row (Bertrand, Meyer & Maumy-Bertrand, 2014).

The Q_h^2 criterion is defined for each h component as:

$$Q_h^2 = 1 - PRESS_h / RSS_{h-1}, \tag{6}$$

where $PRESS_h$ is the *PRedictive Error Sum of Squares* when the number of components containing h components and RSS_{h-1} is the *Residual Sum of Squares* associated to the model containing $(h-1)$ components. RSS_h can be calculated by $RSS_h = \sum_{i=1}^{n} \left(y_i - \hat{y}_{hi}\right)^2$ where $\hat{y}_{hi}$ is predicted value based on the model with h components.

PRESS is computed by cross validation with the below formula (Pérez-Enciso & Tenenhaus, 2003):

$$PRESS_h = \sum_{i=1}^{n} \left(y_{h-1,i} - \hat{y}_{h-1,-i}\right)^2, \tag{7}$$

where $y_{h-1,i}$ is the residual of observation i when $h-1$ components are fitted, and $\hat{y}_{h-1,-i}$ is the predicted y_i obtained when the $i-th$ observation is removed.

A new $\mathbf{t_h}$ component is kept for the prediction of $\mathbf{y}$ if (see Tenenhaus, 1998):

$$\sqrt{PRESS_h} \leq 0.95\sqrt{RSS_{h-1}} \Leftrightarrow Q_h^2 \geq 0.0975, \tag{8}$$

where (0.95) is arbitrary value.

For components selection by information criteria, the number of degrees of freedom (DoF) has been computed using the methods of Krämer and Sugiyama (2012) and implemented in the `plsdof` package (Krämer & Braun, 2015). The PLS routines in the `plsRglm` package are based on these DoF except in the case of incomplete data for which only naive DoF are currently implemented.

Krämer and Braun (2015) defined *AIC* and *BIC* as :

$$AIC = \frac{RSS}{n} + 2\frac{DoF}{n}\sigma^2 \tag{9}$$

$$BIC = \frac{RSS}{n} + log(n)\frac{DoF}{n}\sigma^2 \tag{10}$$

where RSS represents the Residual Sum of Squares as associated to the PLS regression, n is the number of observations, σ^2 is the unknown variance of the error variables and *DoF* is the degrees of freedom for the PLS regression (Krämer & Sugiyama, 2012).

2.4 Imputation methods

2.4.1 Multiple imputation

Multiple imputation is a general statistical method for the analysis of incomplete data sets (Rubin, 1987; Royston, 2004; Van Buuren, 2012). This method has become a conventional approach for dealing with missing data in

numerous analyses from different domains. Multiple imputation aims to provide unbiased and valid estimates of associations based on information from the available data.

The idea underlying multiple imputation is to use the observed data distribution to generate plausible values for the missing data, replacing them several times over several runs, then combining the results. The multiple imputation algorithm has three steps (Rubin, 1996). The first involves specifying and generating plausible values for missing values in the data. This stage, called imputation, creates multiple imputed data sets (m of them). In the second step, a statistical analysis is performed on each of the m imputed data set to estimate quantities of interest. The results of the m analyses will differ because the m imputations differ. There is variability both within and between the imputed data set because of the uncertainty related to missing values. The third step pools the m estimates into one, combining both within- and between- imputation variation.

Several authors have addressed the question of the optimal number of imputations. Rubin recommended 2–5 imputations in (Rubin, 1987). He argued that even with 50% missing data, five imputed data sets would produce point estimates that were 91% as efficient as those based on an infinite number of imputations. In 1998, Graham, Olchowski, and Gilreath (2007) suggested 20 or more imputations. Later Bodner (2008) and White, Royston, and Wood (2011) suggested the rule of thumb that m, the number of imputations, should be at least equal to the percentage of missing entries, which is what we do in this paper.

Multiple imputation by chained equations (MICE) is a practical approach to generating imputation in the first step of multiple imputation (Van Buuren & Groothuis-Oudshoorn, 2011). A more detailed description of the theory involved is provided by Van Buuren (2007), Van Buuren and Groothuis-Oudshoorn (2011), and Azur et al. (2011). In this study, we used the `mice` package (Van Buuren, 2018).

2.4.2 *K*-Nearest Neighbours imputation

The method of *K*-Nearest Neighbours imputation estimates a missing data point using values calculated from its *K* nearest neighbours, defined in terms of similarity (Dixon, 1979). In particular, Nearest Neighbours imputation can be with respect to some distance function.

Types of distances that can be used include the Pearson correlation, Euclidean, Mahalanobis, Chebyshev, and Gower distances. Typically, two far apart vectors are less likely than close together ones to have similar values. For a given missing data point, KNNimpute searches the whole data sets for its nearest neighbours. The missing value is then replaced by averaging the (non-missing) values of these neighbours. The method's accuracy depends on the number of neighbours taken into account. The Gower distance is coded by Kowarik and Templ (2016) in the `VIM` R package (Templ et al., 2017).

2.4.3 Singular Value Decomposition imputation

Troyanskaya et al. (2001) proposed The Singular Value Decomposition imputation algorithm. This algorithm estimates missing values as linear combinations of the k most significant eigenvectors, where the most significant eigenvector is the one with the largest, in absolute value, eigenvalue. In this study, we used the `bcv` package (Perry, 2015) to run this.

3 Simulation procedure

3.1 Reference data set construction

Complete data set with a defined number of components were generated using the method described in Li, Morris, and Martin (2002). The actual number of components was chosen to be 2, 4 and 6. The univariate response $\mathbf{y}$ was distributed according to a Gaussian distribution $\mathcal{N}(0, 1)$. Simulations were performed by adapting the `simul_data_UniYX` function available in the `plsRglm` package (Bertrand, Meyer & Maumy-Bertrand, 2014).

3.2 Data dimensions

PLS regression is particularly pertinent for data matrices $\mathbf{X}$ in which $n < p$, but the behaviour of the NIPALS algorithm can depend on whether $\mathbf{X}$ in which $n < p$, or vice versa. Its properties have thus been studied on vertical data matrices, i.e. those for which $n > p$ (e.g. $n = 100$ and $p = 20$ in our study) and horizontal data

matrices, i.e. those for which $n < p$ (e.g. $n = 20$ and $p = 100$ here). The range of scenarios we consider is shown in Section 3.4.

3.3 Missing data and missingness mechanism

Missing data were generated under MCAR and MAR. Missing data are simulated only on matrix **X**. The percentage of missing data d took values in $d \in \{5, 10, \ldots, 50\}\%$. Smaller proportions than 5% of missing data could have been used for $n = 100$, but preliminary runs showed that the results were very close to those for $d = 5\%$. Moreover, it was decided not to include missing data rates larger than 50% in our study since it is more than questionable to run a model on a data set in which more than half of the data are missing.

3.4 Simulation study design

The simulation study was designed as displayed on Figure 1.

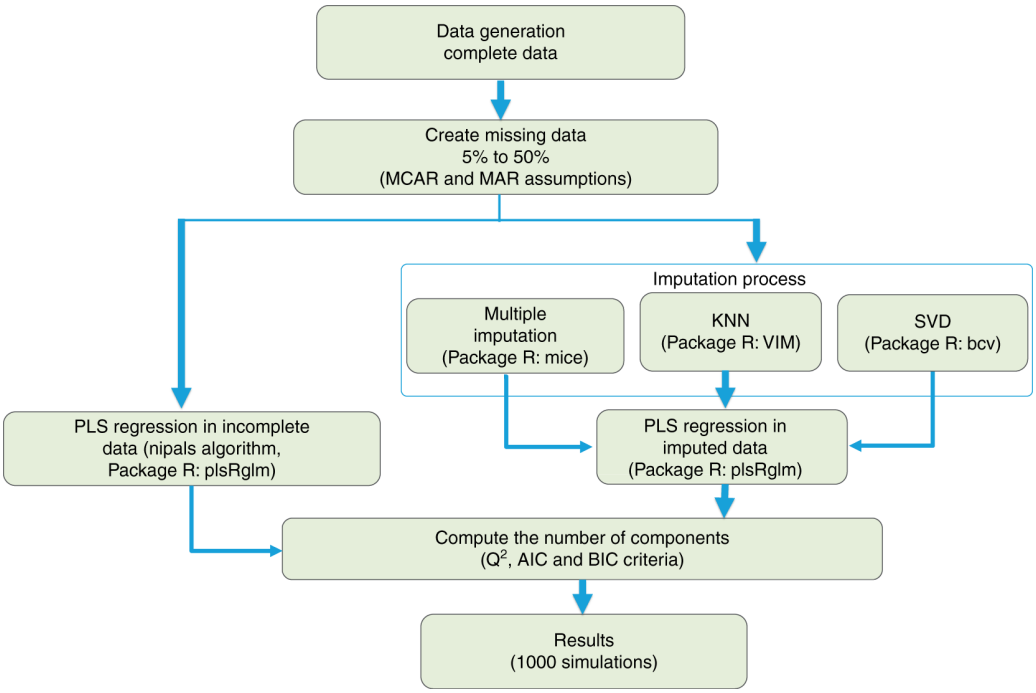

Figure 1: Simulation design.

1. Data were simulated as in Li, Morris, and Martin (2002) on the univariate response **y** and on outcome variable **X** with n (number of observations) and p (number of variables) and t^* (the actual number of components) set to 2, 4 and 6 with each of the following six dimensions set-ups:
   - $n = 500$ and $p = 20$,
   - $n = 100$ and $p = 20$,
   - $n = 80$ and $p = 25$,
   - $n = 60$ and $p = 33$,
   - $n = 40$ and $p = 50$,
   - $n = 20$ and $p = 100$.
2. Missing data were created in the **X** matrices under the MCAR and MAR assumptions, with the proportion of missing data going from 5% to 50% in steps of 5%.
3. Missing data were imputed using MICE from the `mice` package and the `norm` imputation method, KNNimpute with the `VIM` package, and SVDimpute with the `bcv` package.

4. The number of components was computed using Q^2 leave-one-out cross-validation and Q^2 10-fold cross-validation computed on the incomplete data according to the *standard* or *adaptative* methods in the `PlsRglm` package. In the multiple imputation, the number of components was calculated – by Q^2 cross-validation – as the mode of the computed number of components across all m imputations, where m was equal to the percentage of missing data (White, Royston & Wood, 2011).

For each combination of (i) proportion of missing data, (ii) matrix dimensions, and (iii) type of missingness, 1000 replicate data set were drawn.

4 Simulation results

4.1 Complete data set

First, the complete data set were simulated related to each data set both MCAR and MAR assumptions (Table 1–Table 2). We can see that the Q^2-10-fold criterion is the best criterion as it selects the largest true number of components in every dimension either MCAR or MAR assumptions. When $t^* = 2$, the correct model dimension was selected in 97% cases. Furthermore, the performances of both Q^2-LOO and Q^2-10-fold criteria for selecting the correct number of components decrease when the sample size decreases, the number of variables increases and the number of components increases. Similar conclusions can be drawn for AIC-DoF and BIC-DoF. These methods, however, perform less well and, overall, selects the correct number of components less frequently than Q^2 does.

On the other hand, the AIC and BIC criteria are the less efficient ones to determine the actual number of components. It generally yields a larger number of components than expected. The AIC and BIC criteria selected eight components in almost every run of the 1000 simulations (results not shown here). It must be reminded that the number of computed components was set to a maximum of eight. Thus it cannot be excluded that the observed number of components may have been larger in some of the simulations.

Table 1: The evaluation of complete data in MCAR assumption.

n	p	$t^* = 2$						$t^* = 4$						$t^* = 6$					
		Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF
20	100	469	970	0	0	0	0	190	724	0	0	0	0	49	134	0	0	0	0
40	50	896	986	0	161	0	259	817	925	0	297	0	377	770	786	0	392	0	515
60	33	979	995	0	483	0	786	945	958	0	508	0	688	958	944	0	466	0	644
80	25	991	998	0	689	0	886	977	977	0	630	0	792	987	987	0	484	0	662
100	20	996	999	0	767	0	905	992	994	0	708	1	846	995	997	0	389	4	572
500	20	1000	1000	0	884	7	951	1000	1000	0	840	12	920	1000	1000	0	128	29	391

The results are expressed as number of simulations for which the selected components number (t^*) equals to 2, 4 and 6 (the actual value) over 1000 simulations, n is the number of observations, p is the number of variables.

Table 2: The evaluation of complete data in MAR assumption.

n	p	$t^* = 2$						$t^* = 4$						$t^* = 6$					
		Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2-LOO	Q^2-10-fold	AIC	AIC-DoF	BIC	BIC-DoF
20	100	440	972	0	0	0	0	182	718	0	0	0	0	62	128	0	0	0	0
40	50	896	983	0	170	0	262	810	915	0	298	0	380	778	796	0	387	0	504
60	33	973	994	0	481	0	786	943	955	0	496	0	684	952	935	0	474	0	649
80	25	992	998	0	682	0	880	976	977	0	634	0	791	989	987	0	476	0	654
100	20	996	999	0	759	1	903	994	995	0	693	2	832	997	997	0	386	3	568
500	20	1000	1000	0	893	6	953	1000	1000	0	838	13	919	1000	1000	0	130	32	399

The results are expressed as number of simulations for which the selected components number (t^*) equals to 2, 4 and 6 (the true value) over 1000 simulations, n is the number of observations, p is the number of variables.

4.2 Comparison of the different algorithms

The framework for our simulations was based on those in Li, Morris, and Martin (2002) with the actual number of components is set to 2, 4 and 6. In detail, Figure 3–Figure 14 plot the performance of each method as a function of the proportion of missing data with various shaped matrices under both MCAR and MAR. A summary of these results for all criteria and methods are shown in Table 3.

These results show that the performance of Q^2-LOO in selecting the correct number of components increases generally as the sample size does, and as expected, decreases as the proportion of missing data increases for both MCAR and MAR. The NIPALS-PLSR with Q^2-LOO generally provides a satisfactory performance with any dimension under MCAR for $t^* = 2$. In comparison, the MICE with Q^2-LOO, Q^2-10-fold perform well under both MCAR and MAR assumptions for $t^* = 4$ and $t^* = 6$. They give results closest to the correct number of components on the incomplete data set when the proportion of missing data was small (< 30%).

The SVDimpute with Q^2-LOO performs the worst when $t^* = 4$ and $t^* = 6$. The actual number of components was correctly selected in only around one-third of the simulations. On the contrary, it works well when $t^* = 2$ for horizontal data matrices setting under MAR assumption and the proportion of missing data equals 5%.

We see that the Q^2-10-fold performs less well and, overall, selects the correct number of components less frequently than Q^2-LOO does. In the vast majority of situations (i.e. combinations of matrix size, proportion, and pattern of missing data), the number of components selected is on average larger than the actual number of components.

We also found that AIC, AIC-DoF, and BIC systematically select a larger number of components than Q^2. This difference can sometimes be as large as three or four for each of the methods and both the MCAR and MAR cases.

The true number of components selected by either MICE, KNNimpute, or SVDimpute with the BIC-DoF criterion are systematically larger than the number of components selected by BIC, AIC, and AIC-DoF. BIC-DoF's performance increases and then decreases as the proportion of missing data increases, instead of regularly decreasing over the whole range of missing data proportions.

Table 3: The evaluation of NIPALS-PLSR, MICE, KNNimpute and SVDimpute.

n	p	Assumption	$t^* = 2$						$t^* = 4$						$t^* = 6$					
			Q^2 -LOO	Q^2 -10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2 -LOO	Q^2 -10-fold	AIC	AIC-DoF	BIC	BIC-DoF	Q^2 -LOO	Q^2 -10-fold	AIC	AIC-DoF	BIC	BIC-DoF
20	100	MCAR	NIPALS-PLSR SVDimpute	KNNimpute SVDimpute	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
		MAR	SVDimpute	SVDimpute KNNimpute	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
40	50	MCAR	NIPALS-PLSR KNNimpute SVDimpute	MICE KNNimpute SVDimpute	–	–	–	–	MICE	MICE	–	–	–	–	–	–	–	–	–	–
		MAR	KNNimpute SVDimpute	KNNimpute SVDimpute	–	–	–	–	MICE	MICE	–	–	–	–	–	–	–	–	–	–
60	33	MCAR	NIPALS-PLSR KNNimpute	SVDimpute KNNimpute	–	–	–	–	MICE	MICE	–	–	–	–	–	–	–	–	–	–
		MAR	KNNimpute SVDimpute	KNNimpute SVDimpute	–	–	–	–	MICE	MICE	–	–	–	–	–	–	–	–	–	–
80	25	MCAR	NIPALS-PLSR KNNimpute	SVDimpute KNNimpute	–	–	–	–	MICE	MICE	–	MICE	–	MICE	MICE	MICE	–	–	–	–
		MAR	KNNimpute SVDimpute	KNNimpute SVDimpute	–	–	–	–	MICE	MICE	–	–	–	MICE	MICE	MICE	–	–	–	–
100	20	MCAR	NIPALS-PLSR KNNimpute	MICE KNNimpute	–	–	–	MICE	MICE	MICE	–	MICE	–	MICE	MICE	MICE	–	–	–	–
		MAR	MICE KNNimpute	MICE KNNimpute SVDimpute	–	–	–	–	MICE	MICE	–	–	–	MICE	MICE	MICE	–	–	–	–
500	20	MCAR	NIPALS-PLSR KNNimpute	MICE KNNimpute	–	MICE	–	MICE	NIPALS-PLSR MICE KNNimpute	KNNimpute MICE	–	MICE	–	MICE	MICE	MICE	–	–	–	–
		MAR	NIPALS-PLSR KNNimpute	KNNimpute MICE	–	MICE	–	MICE	NIPALS-PLSR MICE	KNNimpute MICE	–	MICE	–	MICE	MICE	MICE	–	–	–	–

MICE	KNNimpute

The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 2, 4 and 6 (the true value), n is the number of observations, p is the number of variables.

5 Real data

We applied PLS regression to four type data: (i) The bromhexine data in a pharmaceutical syrup (Goicoechea & Olivieri, 1999a); (ii) The tetracycline data in serum (Goicoechea & Olivieri, 1999b); (iii) The Los Angeles ozone pollution data 1976 which is provided by the `mlbench` package (Leisch & Dimitriadou, 2010); (iv) The octane data in gasolines from NIR data (Goicoechea & Olivieri, 2003).

The focus of this paper is to determine the number of components of a PLS regression fitted with the NIPALS algorithm. The real data analysis aims to extract components to build representative models of the process. We repeatedly selected one hundred times the number of components in complete data using the $q = 10$ CV-criterion. The larger number of components that could be selected was twelve. The results of this selection process are displayed in Figure 2 with a summary in Table 4. The selected significant number of components in complete data real (t^{**}) in Bromhexine, Tetracycline, Los Angeles ozone pollution, and octane data set is 3-, 5-, 2- and 2-components, respectively. Besides, in Bromhexine data, $t^{**} = 2$ can be selected based on this number of components being selected in 42 out of 100 runs.

The procedure for real data analysis is the same as a simulation study. Missing data were first created on complete real data, both MAR and MCAR assumptions from 5% to 50% in steps 5%. PLS regression was applied then on incomplete data with NIPALS-PLSR and PLS regression on imputed data set which used three methods of imputation: multiple imputation by chained equations (MICE), k-nearest neighbour imputation (KNNimpute) and a singular value decomposition imputation (SVDimpute). Finally, the number of components is computed using Q^2, AIC and BIC and their performance are compared with t^{**}. For addition, we also compare the performance of real data results with the simulation results. The real data evaluation of the criteria for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR and MAR assumptions with different proportions of missing values are shown in Table 6–Table 9. A summary of the real data results for all criteria is shown in Table 5.

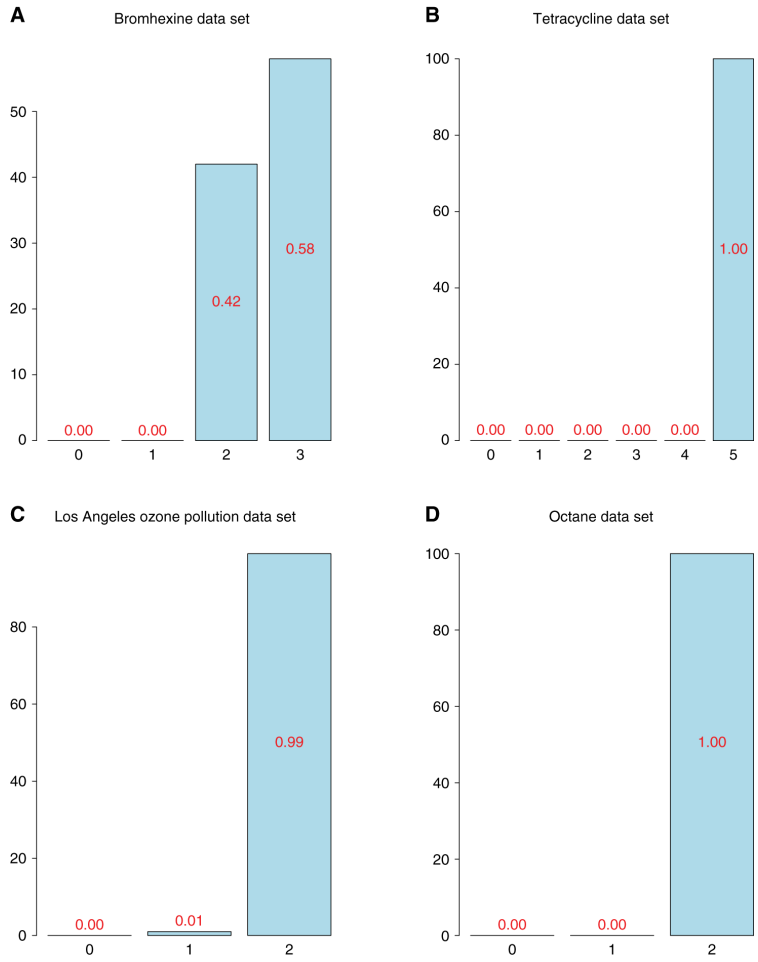

Figure 2: Plot of extracted significant numbers of components in complete data real using Q^2-10-fold with 100 times. **A**. the bromhexine data. **B**. the tetracycline data in serum. **C**. the Los Angeles ozone pollution data 1976. **D**. the octane data in gasolines from NIR data.

Table 4: Selected significant number of components in complete data real (t^{**}).

Data set	Bromhexine	Tetracycline	Los Angeles ozone pollution	Octane
t^{**}	3(58/100) 2(42/100)	5(100/100)	2 (100/100)	2(100/100)

Table 5: The evaluation of NIPALS-PLSR, MICE, KNNimpute and SVDimpute in incomplete data real.

Data set	Dimension ($n \times p$)	t^{**}	Assumption	Criteria					
				Q^2 -LOO	Q^2 -10-fold	AIC	AIC-DoF	BIC	BIC-DoF
Bromhexine	23 x 64	3 / 2	MCAR	NIPALS-PLSR SVDimpute	KNNimpute SVDimpute	–	–	–	–
			MAR	NIPALS-PLSR SVDimpute	SVDimpute	–	–	–	–
Tetracycline	107 x 101	5	MCAR	SVDimpute	–	–	–	–	–
			MAR	SVDimpute	–	–	–	–	–
Los Angeles ozone pollution	203 x 12	2	MCAR	MICE KNNimpute	MICE KNNimpute	–	–	–	–
			MAR	MICE KNNimpute	MICE KNNimpute	–	–	–	–
Octane	68 x 493	2	MCAR	KNNimpute NIPALS-PLSR SVDimpute	KNNimpute MICE SVDimpute	–	–	–	–
			MAR	KNNimpute NIPALS-PLSR SVDimpute	KNNimpute – –	–	–	–	–

The results are expressed as the selected combination between the criteria and the methods that the number of components of a PLS regression is close to the selected significant number of components (t^{**}).

5.1 Bromhexine data

In this data set, the author discussed the possibility of determining bromhexine in syrups by applying electronic absorption measurements together with robust multivariate calibration analyses. The authors used PLS1, among other methods. A data set of 23 samples were prepared with 64 concentrations of bromhexine ($n = 23 \times p = 64$).

The performance of criteria for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR and MAR assumptions with different proportions of missing values are shown in Table 6. For comparison, we select the dimension $n = 20 \times p = 100$ with $t^* = 2$ in the simulation results.

The SVD impute, with either Q^2-LOO or Q^2-10-fold criterion generally performs well to select the correct number of components ($t^{**} = 2$ and 3) in this data set, under both MCAR and MAR assumptions. These situations also correspond with the simulation results.

In contrast, the performance of the AIC, AIC-DoF, BIC, and BIC-DoF include too many components, clearly overfitting the data in this case. In most of the results, the selected number of components is almost five times the selected significant number of components (10 components). This finding is also supported by the results that we obtained with our previous simulation study. We found that the numbers of components selected by those criteria were systematically larger than the correct number of components.

5.2 Tetracycline data set

Tetracycline has been determined in human serum samples. It has 107 observations on 101 explanatory variables ($n = 107 \times p = 101$) with spectral range (in nm) between 0 and 600. The tetracycline data set was obtained from Richet Laboratories, and its purity was checked according to Pharmacopeia recommendations. This research discussed the possibility of quantitating tetracycline and its derivatives in blood-serum samples by applying synchronous spectrofluorometric measurements together.

The overall performance of the criteria for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR and MAR assumptions with different missing data proportions are shown in Table 7. We did not compare this data set with the simulation results because we could not correctly match the data sets dimensions.

Either for the MCAR or MAR situation, the performance of SVDimpute with Q^2-LOO criterion is the best combination between the methods and the criteria to determine the selected significant number of components (Table 4). As with the case of Bromhexine data set, the AIC, AIC-DoF, BIC, and BIC-DoF criteria have less good performances overall. They select a larger number of component than the Q^2-LOO and Q^2-10-fold criteria. The selected dimension can sometimes be twice the correct dimension.

5.3 Los Angeles ozone pollution data set

Los Angeles ozone pollution data set concerns 12 predictor variables which contain the measurement dates and, among others, information on wind speed, humidity, temperature. This data set contains 366 daily observations used to predict the daily maximum one-hour-average ozone reading. The original data contains missing values. Among the 366 samples, we used the 203 complete ones ($n = 203 \times p = 12$).

The evaluation of the criteria for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR and MAR assumptions with different proportions of missing values are shown in Table 8. We compared the real data set results ($t^{**} = 2$) with the simulation results on $n = 100 \times p = 20$ and $t^{*} = 2$ that almost matched the real data set's dimensions.

In the Los Angeles ozone pollution data set analysis, the number of selected components using Q^2-LOO and Q^2-10-fold criteria on MICE and KNNimpute are the closest, for any proportion of missing data, to the number of components that we selected using complete data. They perform well to determine $t^{**} = 2$. These results correspond with the simulation results obtained under the MAR assumption but are somewhat different from those obtained under the MCAR assumption. The NIPALS-PLSR and KNNImpute perform well in the simulation.

Generally, AIC, AIC-DoF, BIC, and BIC-DoF have less good performance. These criteria include too many components. The selected number of components are almost more than twice as large as the selected significant number of components. This finding is also supported by the simulation results obtained.

5.4 Octane data set

The experiment of Octane data set studied the concentration of glucuronic acid in complex mixtures studied by Fourier transform mid-infrared spectroscopy and the octane number in types of gasoline monitored by near-infrared spectroscopy. Determination of octane in types of gasoline from NIR data on this study used 68 NIR spectra of gasoline samples collected in a local distillery, in the range 4020–9996 cm^{-1} which is categorized in 493 explanatory variables ($n = 68 \times p = 493$).

The evaluation of the criteria for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute both MCAR and MAR assumptions with different proportions of missing values are shown in (Table 9).

All methods except MICE with Q^2-LOO criterion exhibit good and similar performance in terms of determination of the number of components, whatever the proportion and mechanism of missing data. The AIC, AIC-DoF, BIC, and BIC-DoF include too many components. The difference can sometimes be as large as eight components.

6 Discussion and conclusion

PLS regression is a multivariate method for which two algorithms (SIMPLS and NIPALS) can be used to provide the model's parameter estimates. The NIPALS algorithm has the interesting property of being able to provide estimates on incomplete data; this has been extensively studied in the case of PCA – for which NIPALS was originally devised.

Here, we have studied the behavior of the NIPALS algorithm when used to fit PLS regression models for various proportions of missing data and different types of missingness. Comparisons with MICE, KNNimpute, and SVDimpute were performed under the MCAR and MAR assumptions. In our simulations, the model dimension (i.e. the optimal number of components) was computed according to different criteria, including Q^2, AIC, and BIC.

The number of selected components, be it for complete, incomplete, or imputed data set, depends on the criterion used. The fact that AIC and BIC select a larger number of components than Q^2 has been already observed in another context by the present authors (Meyer, Maumy-Bertrand & Bertrand, 2010) and by others (Li, Morris & Martin, 2002).

In the simulation results of Q^2-LOO, the number of components selected using NIPALS-PLSR (two-components) under MCAR assumption and MICE (two- and four-components) under MCAR and MAR assumptions are much closer to the correct number of components when the proportion of missing data is small (< 30%) and for vertical matrices. However, the MICE computation time was long, and depended on the proportion of missing data, increasing as the proportion of missing data did. For instance, the average MICE run time for $n = 100 \times p = 20$, was about 11 times longer than that of NIPALS-PLSR when $d = 10\%$, and around 40 times longer when $d = 50\%$ under MCAR. In contrast, NIPALS-PLSR, KNNimpute, and SVDimpute had swift run times: 0.5–1.5 seconds on average. Generally, the run time under MAR was longer than under MCAR for both vertical and horizontal matrices. Consequently, though MICE may be the method of choice, its run time may prohibit its use in practice.

BIC-DoF, a criterion derived from BIC, gives a slightly better estimation of the number of components in the simulation when the proportion of missing data is small, particularly for MICE. This finding shows that taking into account a modified number of DoF can substantially improve the likelihood of selecting the correct number of components (Krämer & Sugiyama, 2012). Further research would nevertheless be useful to extend this version of BIC to other settings like, for instance, GLM or adapt it to specific cases of incomplete data set that require further *DoF* adjustments. In contrast to the real data results, the performance of BIC-DoF includes too many components.

For smaller sample sizes n, the multivariate structure of the data was not taken into account in the imputations due to high levels of collinearity. Indeed, the smaller the sample size, the more difficult it was for the MICE algorithm to converge. Thus, though it would have been possible to run the imputation, the PLS regression estimates would have been biased. This result implies that our conclusions for tiny sample sizes may be misleading. Such biased parameter estimates could also bias the comparisons between the methods but also hint at the fact that even a small proportion of missing data can make it difficult to estimate the correct number of components in PLS regression.

In the vast majority of situations, either the simulation results or the practical examples discussed (that is any combination of size, proportion and pattern of missing data), It is clear that the Q^2-LOO criterion has the best performance. Theoretically, the leave-one-out method can extract the maximum possible information (Eastment & Krzanowski, 1982). The simulation results in this study support this result. Furthermore, our real data analysis shows similar results and our simulation results back up our real data analysis.

In conclusion, our simulations show that whatever the criterion used, the type of missingness and proportion of missing data must also be taken into consideration since they both influence the number of components selected. These results match our real data studies. The actual number of components of a PLS regression was challenging to determine, especially for small sample sizes, and when the proportion of missing data was larger than 30%. Moreover, under MCAR, the number of selected components using these methods was generally closer to the actual number of components than in the MAR setting.

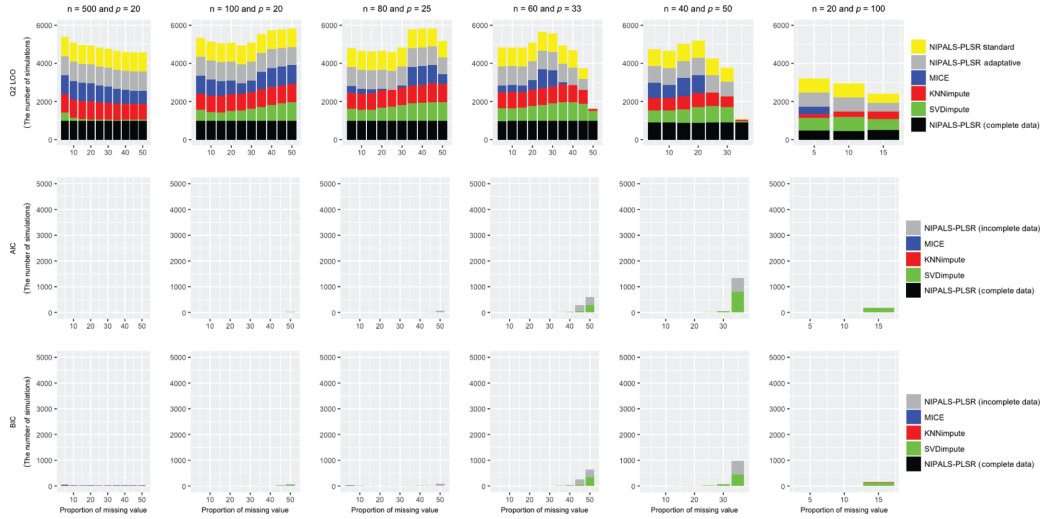

Figure 3: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 2 (the true value).

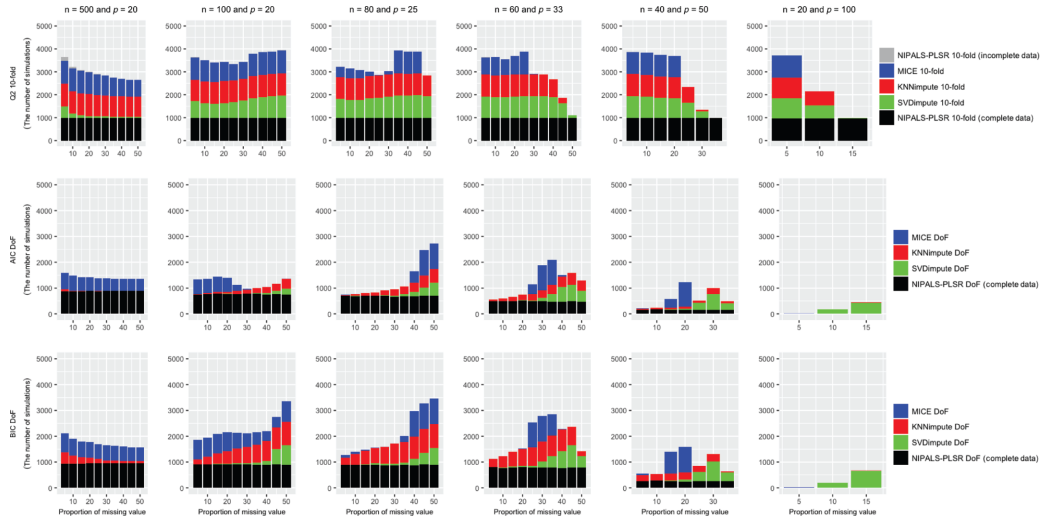

Figure 4: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 2 (the true value).

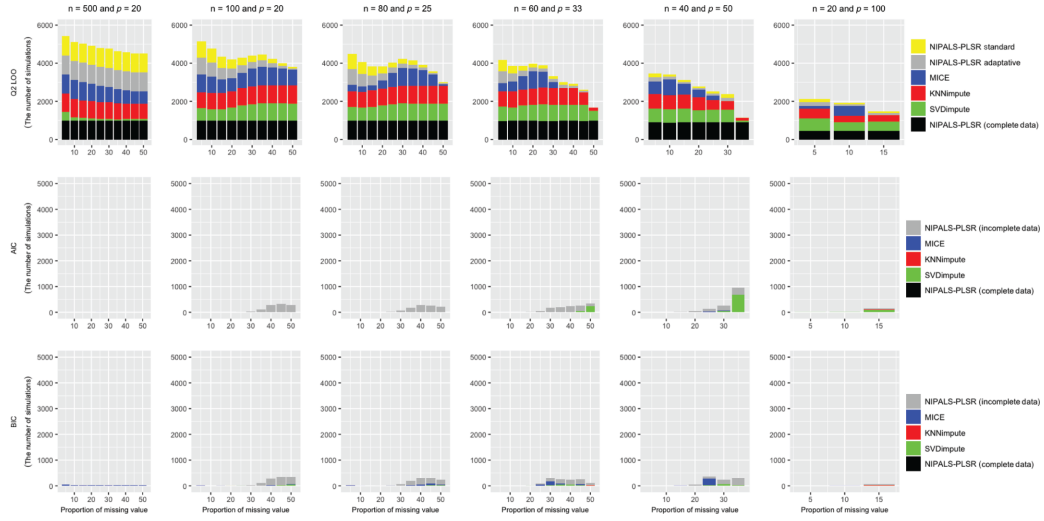

Figure 5: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 2 (the true value).

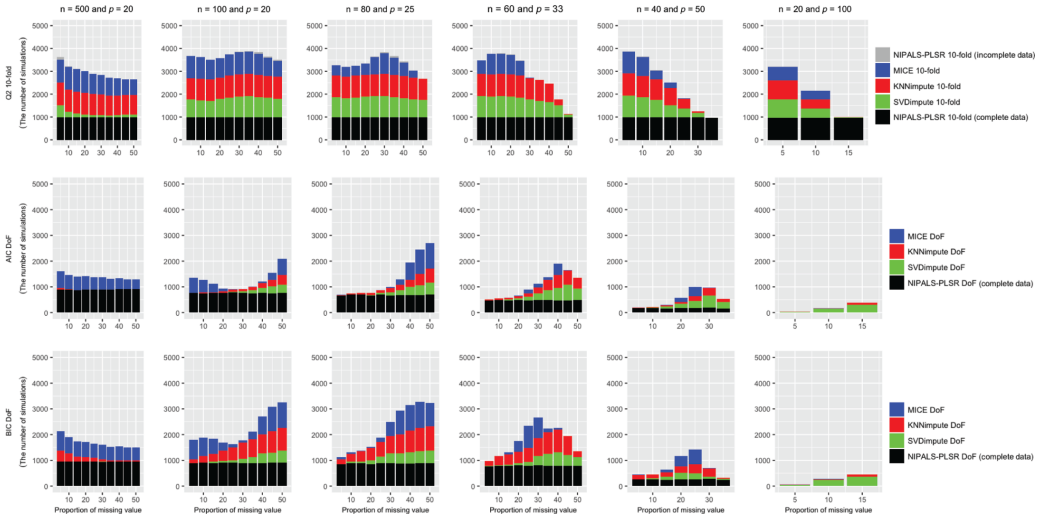

Figure 6: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 2 (the true value).

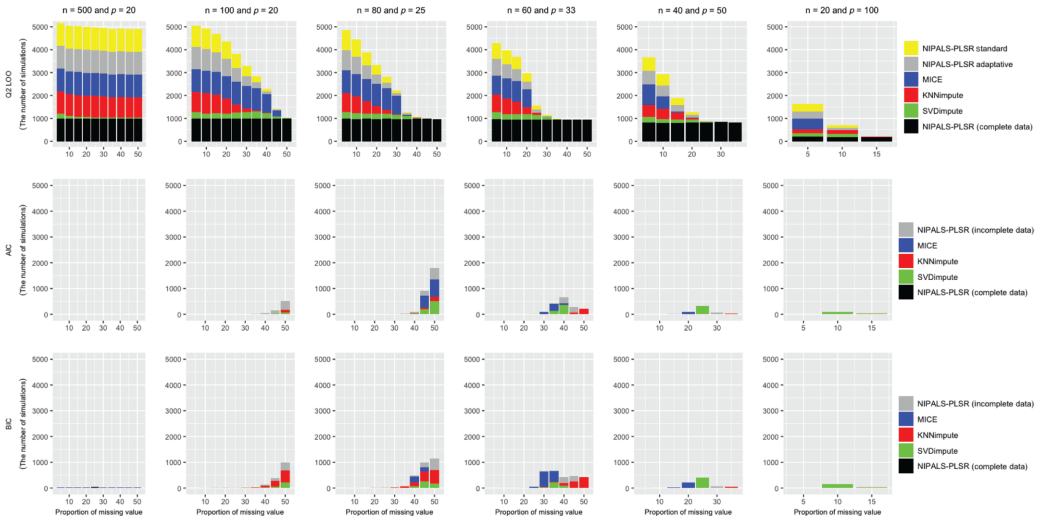

Figure 7: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 4 (the true value).

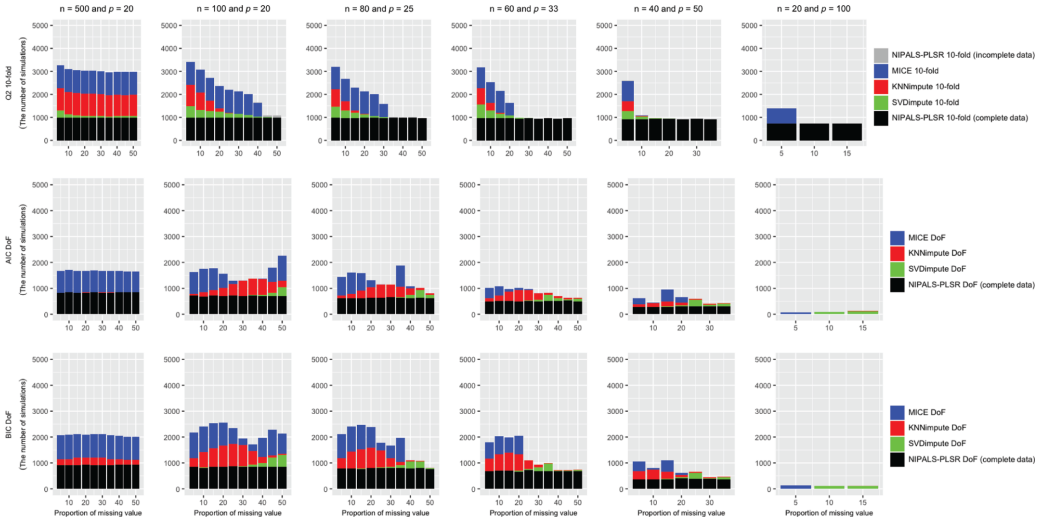

Figure 8: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 4 (the true value).

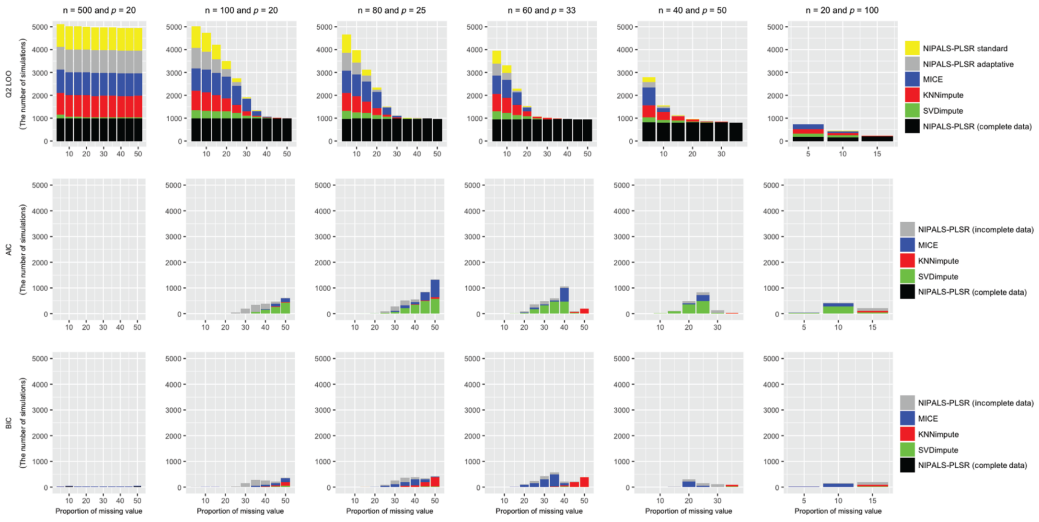

Figure 9: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 4 (the true value).

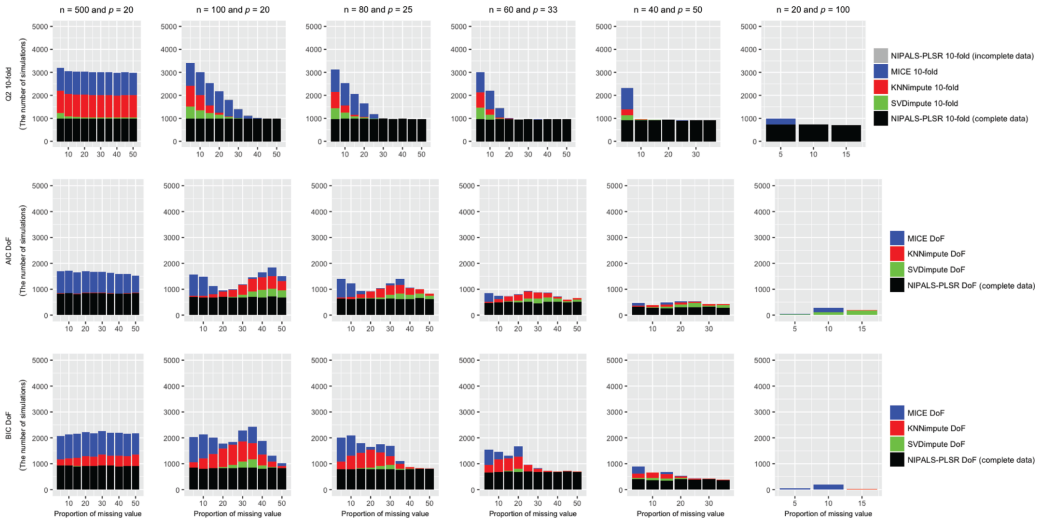

Figure 10: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 4 (the true value).

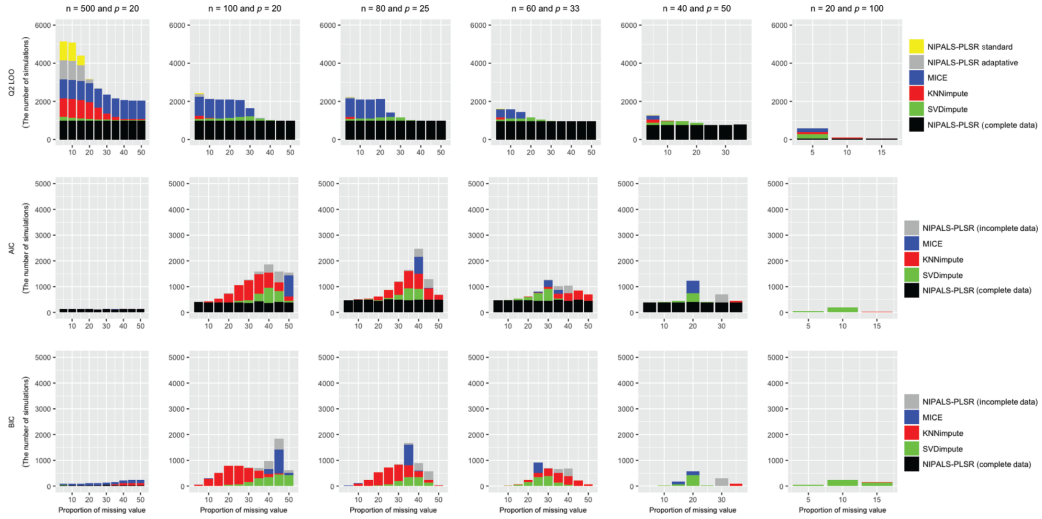

Figure 11: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 6 (the true value).

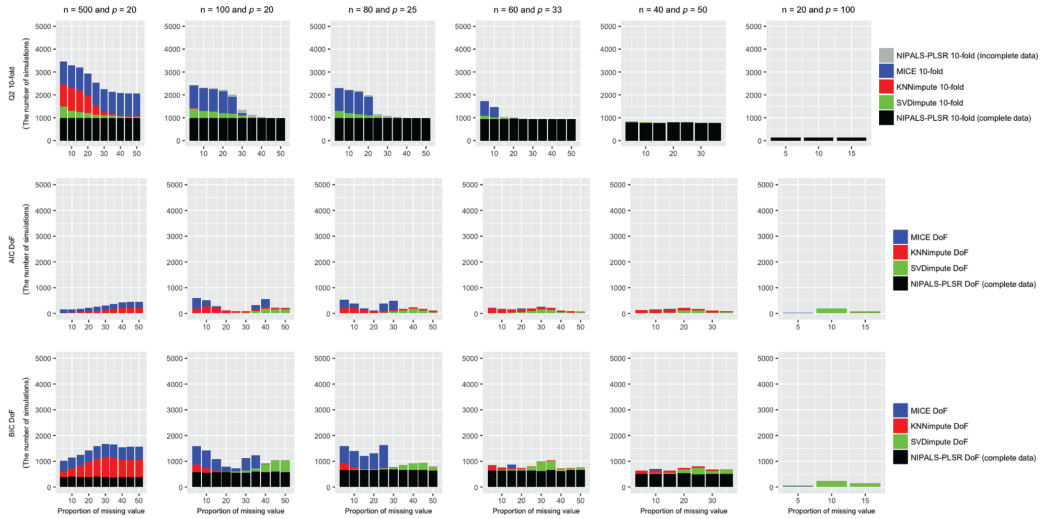

Figure 12: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MCAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 6 (the true value).

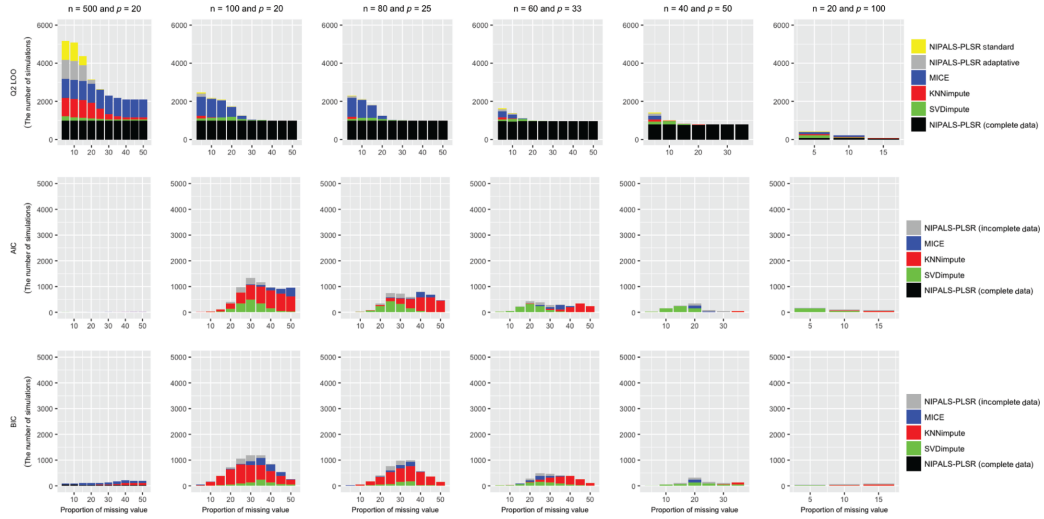

Figure 13: Evaluation of Q^2-LOO, AIC, and BIC for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 6 (the true value).

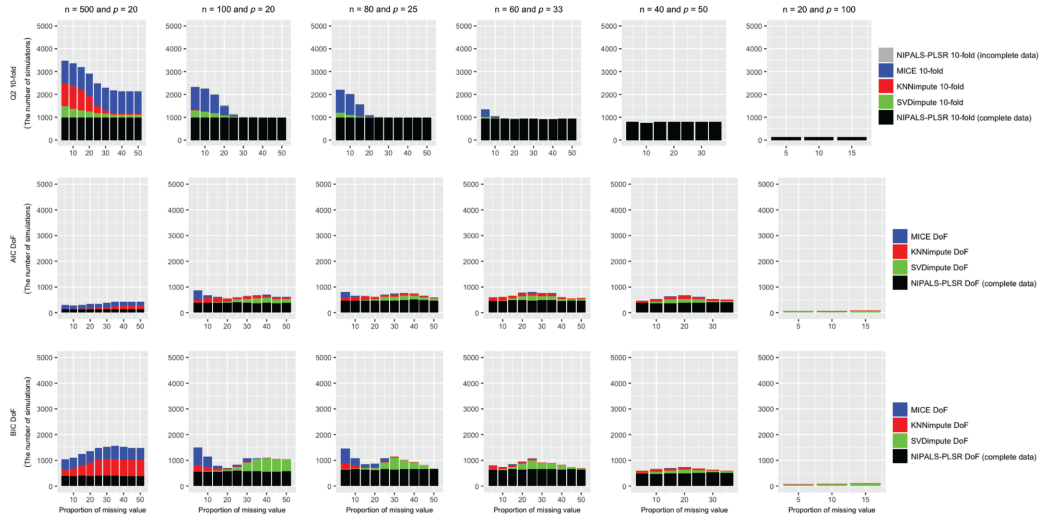

Figure 14: Evaluation of Q^2-10-fold, AIC-DoF, and BIC-DoF for NIPALS-PLSR, MICE, KNNimpute, and SVDimpute under MAR. The results are expressed as the selected combination of criteria and methods with the number of components of a PLS regression which are close to the correct number of components over 1000 simulations for which the selected component number (t^*) equals to 6 (the true value).

Table 6: Bromhexine data set: the results are the number of components selected for each criteria, d is the percentage of missing value.

Assumption	d	Q^2 -LOO					Q^2-10-fold				AIC				AIC-DoF			BIC				BIC-DoF		
		NIPALS-PLSR (standard)	NIPALS-PLSR (adaptative)	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute
MCAR	5	3	3	4	4	2	NA	1	2	2	10	7	10	8	9	10	10	10	7	10	5	9	10	10
	10	2	2	1	4	2	NA	1	2	2	10	6	10	5	1	10	9	10	5	10	5	1	10	9
	15	2	2	1	1	2	NA	1	1	1	10	2	10	3	1	10	4	10	2	10	3	1	10	4
MAR	5	4	4	4	2	2	NA	1	2	2	10	10	10	5	10	10	6	10	10	10	5	10	10	6
	10	3	3	1	1	2	NA	1	1	2	10	7	10	3	9	10	8	10	7	10	3	9	10	8
	15	3	3	1	1	1	NA	1	1	1	10	3	10	2	4	10	1	10	3	10	2	1	10	1

Table 7: Tetracycline data set: the results are the number of components selected for each criteria, *d* is the percentage of missing value.

Assumption	d	Q^2 -LOO					Q^2-10-fold				AIC				AIC-DoF			BIC				BIC-DoF		
		NIPALS-PLSR (standard)	NIPALS-PLSR (adaptative)	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute
MCAR	5	5	5	5	5	5	3	5	4	5	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	10	5	5	5	5	5	3	5	4	5	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	15	3	3	5	5	5	3	5	4	5	10	10	10	10	5	10	10	10	10	10	10	5	10	10
	20	5	5	5	5	5	3	5	4	5	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	25	5	5	5	4	5	3	5	3	4	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	30	5	5	3	4	5	3	3	4	4	10	10	10	10	3	7	10	10	10	10	10	3	6	10
	35	3	3	3	4	5	3	3	3	4	10	10	10	10	3	10	10	10	10	10	10	3	10	10
	40	5	5	3	4	5	3	2	3	5	10	10	10	10	8	10	10	10	10	10	10	8	4	10
	45	5	5	3	3	5	3	2	3	4	10	10	10	7	3	10	10	10	10	10	7	2	10	10
	50	4	4	2	3	5	3	2	3	4	10	10	10	10	2	10	10	10	10	10	10	2	10	10
MAR	5	5	5	5	5	5	3	5	5	5	10	10	10	10	10	10	10	10	10	10	10	10	10	5
	10	5	5	5	5	5	3	5	4	5	10	10	10	10	10	10	9	10	10	10	10	10	10	8
	15	5	5	5	5	5	3	5	4	5	10	10	10	10	5	9	10	10	10	10	10	5	9	10
	20	5	5	5	5	5	3	5	4	5	10	10	10	10	10	6	10	10	10	10	10	10	5	10
	25	5	5	5	4	5	3	4	4	5	10	10	10	10	10	6	10	10	10	10	10	5	4	10
	30	5	5	3	4	5	3	3	3	4	10	10	10	10	10	10	10	10	10	10	10	3	4	10
	35	3	3	3	3	5	3	2	3	4	10	10	10	10	5	10	10	10	10	10	10	2	10	10
	40	5	5	3	3	5	3	2	3	4	10	10	10	8	10	10	10	10	10	10	8	2	3	10
	45	4	4	1	3	4	4	1	3	4	10	10	10	10	2	10	10	10	10	10	10	1	10	10
	50	4	4	1	3	5	4	1	3	4	10	10	10	10	8	4	10	10	10	10	6	1	3	10

Table 8: Los Angeles ozone pollution data set: the results are the number of components selected for each criteria, d is the percentage of missing value.

Assumption	d	Q^2 -LOO					Q^2-10-fold				AIC				AIC-DoF			BIC				BIC-DoF		
		NIPALS-PLSR (standard)	NIPALS-PLSR (adaptative)	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute
MCAR	5	2	2	2	2	2	1	2	2	2	5	5	5	4	9	10	7	4	4	4	4	7	4	7
	10	2	2	2	2	2	1	2	2	2	7	7	5	5	10	9	6	4	4	5	4	7	6	7
	15	2	2	2	2	2	1	2	2	1	7	5	7	4	7	7	4	4	4	3	4	7	2	6
	20	2	2	2	2	2	1	2	2	2	7	5	5	4	9	7	5	7	4	4	4	7	5	7
	25	2	2	2	2	1	1	2	2	1	6	5	6	4	9	9	5	4	4	3	4	7	5	5
	30	2	2	2	2	2	1	2	2	1	7	4	5	4	7	8	10	4	4	4	4	8	5	4
	35	2	2	2	2	2	1	2	2	1	7	5	5	3	7	7	4	5	4	3	4	4	4	7
	40	1	1	2	2	1	1	2	2	1	6	4	5	3	4	7	4	4	4	3	3	7	2	4
	45	2	2	2	2	1	1	2	2	1	5	6	4	3	9	9	7	4	4	2	5	4	2	9
	50	2	2	2	2	1	1	2	2	1	5	5	4	3	7	6	5	4	4	3	4	4	2	4
MAR	5	2	2	2	2	2	1	2	2	2	5	5	5	4	9	9	6	4	4	4	4	7	7	4
	10	2	2	2	2	2	1	2	2	1	7	5	5	4	10	9	4	4	4	4	4	8	7	4
	15	2	2	2	2	2	1	2	2	2	8	5	5	4	7	7	4	4	4	4	4	7	7	4
	20	2	2	2	2	2	1	2	2	2	7	5	5	4	10	7	4	4	4	4	3	4	4	4
	25	2	2	2	2	2	1	2	2	2	6	5	5	5	10	7	10	4	4	4	4	6	6	9
	30	1	1	2	2	2	1	2	2	2	7	4	4	3	9	6	4	4	4	4	3	4	4	4
	35	2	2	2	2	2	1	2	2	1	7	6	5	3	10	7	4	5	4	4	3	7	4	2
	40	2	2	2	2	1	1	2	2	1	5	4	4	3	6	9	3	4	4	4	3	4	4	2
	45	2	2	2	2	1	1	2	2	1	5	5	5	4	7	8	7	5	4	4	3	5	8	2
	50	2	2	2	2	1	1	2	2	1	4	6	6	3	9	7	4	4	4	5	3	2	7	2

Table 9: Octane data set: the results are the number of components selected for each criteria, *d* is the percentage of missing value.

Assumption	d	Q^2 -LOO					Q^2-10-fold				AIC				AIC-DoF			BIC				BIC-DoF		
		NIPALS-PLSR (standard)	NIPALS-PLSR (adaptative)	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute	NIPALS-PLSR	MICE	KNNimpute	SVDimpute	MICE	KNNimpute	SVDimpute
MCAR	5	2	2	2	2	2	2	1	2	2	10	10	10	10	9	7	8	10	10	10	10	9	6	8
	10	2	2	4	2	2	2	2	2	1	10	10	10	10	10	10	10	10	10	10	10	10	10	3
	15	2	2	3	2	2	2	2	2	2	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	20	2	2	2	2	2	2	2	2	2	10	10	10	10	10	10	10	10	10	10	10	10	10	10
	25	2	2	2	2	2	NA	2	2	2	10	10	10	10	10	10	10	10	10	10	10	10	10	10
MAR	5	2	2	2	2	2	2	2	2	2	10	10	10	10	10	8	10	10	10	10	10	10	8	10
	10	2	2	2	2	2	2	2	2	2	10	10	10	10	10	7	8	10	10	10	10	10	7	8
	15	2	2	2	2	2	2	2	2	2	10	10	10	10	10	5	9	10	10	10	10	10	5	9
	20	2	2	4	2	2	2	2	2	2	10	10	10	10	10	9	10	10	10	10	10	10	9	10
	25	2	2	2	2	2	NA	2	2	2	10	10	10	10	10	10	10	10	10	10	10	10	10	10

References

Akaike, H. (1969): "Fitting autoregressive models for prediction," Ann. Ins. Stat. Math., 21, 243–247.

Arteaga, F. and A. Ferrer (2002): "Dealing with missing data in MSPC: Several methods, different interpretations, some examples," J. Chemom., 16, 408–418.

Azur, M. J., E. A. Stuart, C. Frangakis and P. J. Leaf (2011): "Multiple imputation by chained equations: what is it and how does it work?" Int. J. Methods Psychiatr. Res., 20, 40–49.

Bastien, P. and M. Tenenhaus (2003): "PLS regression and multiple imputation." In: Proceedings of the PLS'03 International Symposium, Vilares, M, Tenenhaus, M, Coelho, P & Esposito Vinzi, V editors CISIA Paris. pp. 497–498.

Bertrand, F., N. Meyer and M. Maumy-Bertrand (2014): plsRglm: partial least squares regression for generalized linear models, book of abstracts, User2014!, Los Angeles. R package version 1.2.5.

Bodner, T. E. (2008): "What improves with increased missing data imputations?" Structur. Equ. Modeling, 15, 651–675.

Burnham, A. J., R. Viveros and J. F. Macgregor (1996): "Frameworks for latent variable multivariate regression," J. Chemom., 10, 31–45.

Burnham, A. J., J. F. Macgregor and R. Viveros (1999): "Latent variable multivariate regression modeling," Chemom. Intell. Lab. Syst., 48, 167–180.

De Jong, S. (1993): "SIMPLS: an alternative approach squares regression to partial least," Chemom. Intell. Lab. Syst., 18, 251–263.

Dixon, J. K. (1979): "Pattern recognition with partly missing data," IEEE Trans. Syst. Man Cybern., 10, 617–621.

Eastment, H. T. and W. J. Krzanowski (1982): "Cross-validatory choice of the number of components from a principal component analysis," Technometrics, 24, 73–77.

Eriksson, I., E. Johansson, N. Kettaneh-Wold and S. Wold (2002): "Multi- and megavariate data analysis, principles and applications," J. Chemom., 16, 261–262.

Folch-Fortuny, A., F. Arteaga and A. Ferrer (2016): "Missing data imputation toolbox for MATLAB," Chemom. Intell. Lab. Syst., 154, 93–100.

Goicoechea, H. C. and A. C. Olivieri (1999a): "Determination of bromhexine in cough-cold syrups by absorption spectrophotometry and multivariate calibration using partial least-squares and hybrid linear analyses. Application of a novel method of wavelength selection," Talanta, 49, 793–800.

Goicoechea, H. C. and A. C. Olivieri (1999b): "Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations," Anal. Chem., 71, 4361–4368.

Goicoechea, H. C. and A. C. Olivieri (2003): "A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy," J. Chemom., 17, 338–345.

Graham, J. W., A. E. Olchowski and T. D. Gilreath (2007): "How many imputations are really needed? Some practical clarifications of multiple imputation theory," Prev. Sci., 8, 206–213.

Grung, B. and R. Manne (1998): "Missing values in principal component analysis," Chemom. Intell. Lab. Syst., 42, 125–139.

Horton, N. J. and S. R. Lipsitz (2001): "Multiple imputation in practice: Comparison of software packages for regression models with missing variables," Am. Stat., 55, 244–254.

Höskuldsson, A. (1988): "PLS regression," J. Chemom., 2, 211–228.

Kowarik, A. and M. Templ (2016): "Imputation with the R package VIM," J. Stat. Softw., 74, 1–16.

Krämer, N. and M. L. Braun (2015): plsdof: degrees of freedom and statistical inference for partial least squares regression. R package version 0.2-9.

Krämer, N. and M. Sugiyama (2012): "The degrees of freedom of partial least squares regression," J. Am. Stat. Assoc., 106, 697–705.

Kvalheim, O. (1992): "The latent variable," Chemom. Intell. Lab. Syst., 14, 1–3.

Lazraq, A., R. Cléroux and J.-P. Gauchi (2003): "Selecting both latent and explanatory variables in the PLS1 regression model," Chemom. Intell. Lab. Syst., 66, 117–126.

Leisch, F. and E. Dimitriadou (2010): mlbench: Machine Learning Benchmark Problems. R package version 2.1-1.

Li, B., J. Morris and E. B. Martin (2002): "Model selection for partial least squares regression," Chemome. Intell. Lab. Syst., 64, 79–89.

Little, R. J. and D. B. Rubin (1987): Statistical analysis with missing data,Wiley, New York, Wiley Series in Probability and Statistics – Applied Probability and Statistics Series.

Little, R. J. and D. B. Rubin (2002): Statistical analysis with missing data, A John Wiley & Sons, Inc., New York, 2nd edition.

Meyer, N., M. Maumy-Bertrand and F. Bertrand (2010): "Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives: application aux données d'allélotypage," J. Soc. Stat. Paris., 151, 1–18.

Nelson, P. R., P. A. Taylor and J. F. MacGregor (1996): "Missing data methods in PCA and PLS: score calculations with incomplete observations," Chemom. Intell. Lab. Syst., 35, 45–65.

Nguyen, D. V. and D. M. Rocke (2004): "On partial least squares dimension reduction for microarray-based classification: a simulation study," Comput. Stat. Data An., 46, 407–425.

Oleszko, A., J. Hartwich, A. Wójtowicz, M. Gąsior-Głogowska, H. Huras and M. Komorowska (2017): "Comparison of FTIR-ATR and Raman spectroscopy in determination of VLDL triglycerides in blood serum with PLS regression," Spectrochim. Acta A Mol. Biomol. Spectrosc., 183, 239–246.

Pérez-Enciso, M. and M. Tenenhaus (2003): "Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach Received," Hum. Genet., 112, 581–592.

Perry, P. O. (2015): bcv: Cross-validation for the SVD (Bi-cross-validation): R package version 1.0.1.

Rännar, S., P. Geladi, F. Lindgren and S. Wold (1995): "A PLS Kernel algorithm for data sets with many variables and few objects. 2. Cross-validataion, missing data and examples," J. Chemom., 9, 459–470.

Rosipal, R. and N. Krämer (2005): "Overview and recent advances in partial least squares." In: Subspace, Latent Structure and Feature Selection, Statistical and Optimization, pp. 34–51.

Royston, P. (2004): "Multiple imputation of missing values," Stata J., 4, 227–241.

Rubin, D. B. (1987): Multiple imputation for nonresponse in surveys, John Wiley & Son, New York, New York.
Rubin, D. B. (1996): "Multiple imputation after 18+ years," J. Am. Stat. Assoc., 91, 473–489.
Sawatsky, M. L., M. Clyde and F. Meek (2015): "Partial least squares regression in the social sciences," Quant. Method Psychol., 11, 52–62.
Schwarz, G. (1978): "Estimating the dimension of a model," Ann. Stat., 6, 461–464.
Serneels, S. and T. Verdonck (2008): "Principal component regression for data containing outliers and missing elements," Comput. Stat. Data An., 52, 1712–1727.
Stone, M. (1974): "Cross-validatory choice and assessment of statistical predictions," J. R. Stat. Soc., 36, 111–147.
Templ, M., A. Alfons, A. Kowarik and B. Prantner (2017): VIM: visualization and imputation of missing values. R package version 4.8.0.
Tenenhaus, M. (1998): La Régression PLS: théorie et pratique, Editions Technip.
Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman. (2001): "Missing value estimation methods for DNA microarrays," Bioinformatics, 17, 520–525.
Van Buuren, S. (2007): "Multiple imputation of discrete and continuous data by fully conditional specification," Stat. Methods Med. Res., 16, 219–242.
Van Buuren, S. (2012): Flexible imputation of missing data, Chapman & Hall/CRC, Boca Raton.
Van Buuren, S. (2018): mice: Multivariate imputation by chained equations. R package version 3.3.0.
Van Buuren, S. and K. Groothuis-Oudshoorn (2011): mice: Multivariate imputation by chained equation in R," J. Stat. Softw., 45.
Wakeling, I. N. and J. J. Morris (1993): "A test of significance for partial least squares regression," J. Chemom., 7, 291–304.
White, I. R., P. Royston and A. M. Wood (2011): "Multiple imputation using chained equations: issues and guidance for practice," Stat. Med., 30, 377–399.
Wiklund, S., D. Nilsson, L. Eriksson, M. Sjöström, S. Wold and K. Faber (2007): "A randomization test for PLS component selection," J. Chemom., 21, 427–439.
Wold, H. (1966): Estimation of principal components and related models by iterative least squares, volume 1. Academic Press, New York.
Wold, S., K. Esbensen and P. Geladi (1987): "Principal component analysis," Chemom. Intell. Lab. Syst., 2, 37–52.
Wold, S., M. Sjöström and L. Eriksson (2001): "PLS-regression: a basic tool of chemometrics," Chemom. Intell. Lab. Syst., 58, 109–130.
Yang, T. C., L. S. Aucott, G. G. Duthie and H. M. Macdonald (2017): "An application of partial least squares for identifying dietary patterns in bone health," Arch. osteoporosis, 12, 63.