JOURNAL OF CHEMOMETRICS
J. Chemometrics 2002; **16**: 408–418
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.750

Dealing with missing data in MSPC: several methods, different interpretations, some examples†

Francisco Arteaga[1]* and Alberto Ferrer[2]

[1]Facultad de Estudios de la Empresa, Dpto Métodos Cuantitativos, Guillém de Castro 175, E-46008 Valencia, Spain
[2]Universidad Politécnica de Valencia, Dpto Estadística e IO, Camino de Vera s/n, Edificio I-3, E-46022 Valencia, Spain

Received 30 September 2001; Revised 25 April 2002; Accepted 24 May 2002

This paper addresses the problem of using future multivariate observations with missing data to estimate latent variable scores from an existing principal component analysis (PCA) model. This is a critical issue in multivariate statistical process control (MSPC) schemes where the process is continuously interrogated based on an underlying PCA model. We present several methods for estimating the scores of new individuals with missing data: a so-called trimmed score method (TRI), a single-component projection method (SCP), a method of projection to the model plane (PMP), a method based on the iterative imputation of missing data, a method based on the minimization of the squared prediction error (SPE), a conditional mean replacement method (CMR) and various least squared-based methods: one based on a regression on known data (KDR) and the other based on a regression on trimmed scores (TSR). The basis for each method and the expressions for the score estimators, their covariance matrices and the estimation errors are developed. Some of the methods discussed have already been proposed in the literature (SCP, PMP and CMR), some are original (TRI and TSR) and others are shown to be equivalent to methods already developed by other authors: iterative imputation and SPE methods are equivalent to PMP; KDR is equivalent to CMR. These methods can be seen as different ways to impute values for the missing variables. The efficiency of the methods is studied through simulations based on an industrial data set. The KDR method is shown to be statistically superior to the other methods, except the TSR method in which the matrix to be inverted is of a much smaller size. Copyright © 2002 John Wiley & Sons, Ltd.

KEYWORDS: principal component analysis (PCA); missing data, sensor failure; NIPALS; multivariate statistical process control (MSPC)

1. INTRODUCTION

In the modern process industry, principal component analysis (PCA) [1,2] is widely used to develop models from data sets with large numbers of highly correlated variables, registered on-line by means of sensors hooked up to continuous and batch processes. Once a PCA model has been built, it can be applied in multivariate statistical process control (MSPC) schemes [3] to monitor and diagnose future process operating performance

In this context, missing measurements are a common occurrence [4] owing to several causes: sensor failure, sensor routine maintenance, samples not collected at the required times, data discarded by gross measurement errors, and sensors with different sampling periods. In batch process monitoring, at each time t some method is needed to fill in the future (unknown) data corresponding to the period between time t and the end of the batch.

In MSPC, two problems related to missing data appear: building PCA models from data sets with missing measurements [5,6], and using PCA models for monitoring future observations with missing information, assuming the estimated model to be fixed and known. This paper faces the second problem, in particular that of using future multivariate observations with missing data to estimate latent variable scores from an existing PCA model [7].

In the present paper we discuss and analyse the properties of several methods for handling missing data, some of which have been analysed by Nelson *et al.* [7]: a so-called trimmed score method (TRI), a single-component projection method (SCP) [7], a method based on the iterative imputation of missing data, a method based on the minimization of the squared prediction error (SPE), a conditional mean replacement method (CMR) [7] and various least squared-based methods.

*Correspondence to: F. Arteaga, Facultad de Estudios de la Empresa, Dpto Métodos Cuantitativos, Guillém de Castro 175, E-46008 Valencia, Spain.
E-mail: fjarteaga@fee.edu
†Paper presented at the 7th Scandinavian Symposium on Chemometrics, Copenhagen, Denmark, 19–23 August 2001.
Contract/grant sponsor: Spanish Government (CICYT)/European Union; Contract/grant number: 1FD1997-2159.

Section 2 introduces the notation. The basis for each method and the equivalence between some of them are discussed in Section 3. Expressions for the score estimation error arising from the missing data and the estimated score covariance matrix are developed for each method in Section 3. In Section 4 the properties of the estimated scores from the several methods studied are compared through simulation based on an industrial data set [8]. Finally, Section 5 presents the conclusions of the paper.

Some of the mathematical derivations have been moved to Appendices I–IV.

2. NOTATION [7]

Matrices are written as upper-case bold letters, while lower-case bold letters correspond to column vectors.

Consider a data matrix $\mathbf{X}$ with N rows and K columns, each row corresponding to an individual and each column to a variable. From $\mathbf{X}$ a PCA model can be expressed as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{1}$$

where $\mathbf{T}$ is an $N \times K$ matrix of scores and $\mathbf{P}$ is a $K \times K$ matrix of loadings.

The data matrix $\mathbf{X}$ can then be considered as a collection of row vectors $\mathbf{z}_i^{\mathrm{T}}$ (observations) or column vectors $\mathbf{x}_j$ (variables). The K columns of the loading matrix $\mathbf{P}$ are the loading vectors $\mathbf{p}_j$. The score matrix $\mathbf{T}$ can be considered as a collection of row vectors $\boldsymbol{\tau}_i^{\mathrm{T}}$ (scores of the ith observation) or column vectors $\mathbf{t}_j$ (latent variables).

For any new object $\mathbf{z}$ not used in model building, and assuming that it belongs to the same population as the N individuals of $\mathbf{X}$, the score vector $\boldsymbol{\tau}$ can be calculated as

$$\boldsymbol{\tau} = \mathbf{P}^{\mathrm{T}}\mathbf{z} \tag{2}$$

As $\mathbf{P}$ is an orthonormal matrix, the multivariate vector of new measurements can be expressed as

$$\mathbf{z} = \mathbf{P}\boldsymbol{\tau} \tag{3}$$

Consider that the new observation $\mathbf{z}$ has some unmeasured variables and that these can be taken to be the first R elements of the data vector without loss of generality. Thus the vector can be partitioned as

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}^{\#} \\ \mathbf{z}^{*} \end{bmatrix}$$

where $\mathbf{z}^{\#}$ denotes the missing measurements and $\mathbf{z}^{*}$ the observed variables. This induces the following partition in $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{\#} & \mathbf{X}^{*} \end{bmatrix}$$

where $\mathbf{X}^{\#}$ is the submatrix containing the first R columns of $\mathbf{X}$, and $\mathbf{X}^{*}$ accommodates the remaining $K - R$ columns.

Correspondingly, the $\mathbf{P}$ matrix can be partitioned as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}$$

where $\mathbf{P}^{\#}$ is the submatrix made up of the first R rows of $\mathbf{P}$, and matrix $\mathbf{P}^{*}$ contains the remaining $K - R$ rows.

Assuming that matrix $\mathbf{X}$ is of rank K and that only A out of the K components ($A \leq K$) are significant, we are only interested in working out the first A elements of the score vector for the new individual, $\boldsymbol{\tau}_{1:A}$. In this situation the $\mathbf{P}$ matrix can be expressed as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{1:A} & \mathbf{P}_{A+1:K} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{1:A}^{\#} & \mathbf{P}_{A+1:K}^{\#} \\ \mathbf{P}_{1:A}^{*} & \mathbf{P}_{A+1:K}^{*} \end{bmatrix}$$

where $\mathbf{P}_{1:A}$ contains the first A loadings and $\mathbf{P}_{A+1:K}$ the remaining $K - A$ loadings of the PCA model.

From the previous expressions, Equation (3) can be written as

$$\mathbf{z} = \mathbf{P}\boldsymbol{\tau} = \begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}\boldsymbol{\tau} = \begin{bmatrix} \mathbf{P}_{1:A}^{\#} & \mathbf{P}_{A+1:K}^{\#} \\ \mathbf{P}_{1:A}^{*} & \mathbf{P}_{A+1:K}^{*} \end{bmatrix}\begin{bmatrix} \boldsymbol{\tau}_{1:A} \\ \boldsymbol{\tau}_{A+1:K} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{A+1:K}^{\#}\boldsymbol{\tau}_{A+1:K} \\ \mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K} \end{bmatrix} \tag{4}$$

and the residuals

$$\mathbf{e} = \mathbf{P}_{A+1:K}\boldsymbol{\tau}_{A+1:K} = \begin{bmatrix} \mathbf{e}^{\#} \\ \mathbf{e}^{*} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{A+1:K}^{\#}\boldsymbol{\tau}_{A+1:K} \\ \mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K} \end{bmatrix}$$

$$\equiv \left\{ \begin{array}{l} \mathbf{e}^{\#} = \mathbf{P}_{A+1:K}^{\#}\boldsymbol{\tau}_{A+1:K} \\ \mathbf{e}^{*} = \mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K} \end{array} \right\} \tag{5}$$

Therefore

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}^{\#} \\ \mathbf{z}^{*} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} + \mathbf{e}^{\#} \\ \mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} + \mathbf{e}^{*} \end{bmatrix} = \mathbf{P}_{1:A}\boldsymbol{\tau}_{1:A} + \mathbf{e} \tag{6}$$

From the PCA model, score matrix $\mathbf{T}$ contains the new coordinates of the N individuals in the new K-dimensional space obtained from the orthornormal basis of the columns of $\mathbf{P}$. Therefore the K new variables $\{\mathbf{t}_1, \ldots, \mathbf{t}_k\}$ are independent, with covariance matrix $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a $K \times K$ diagonal matrix of eigenvalues $\{\lambda_1, \ldots, \lambda_K\}$ in decreasing order along the diagonal. If matrix $\mathbf{X}$ does not have full rank, some of the λ values will be zero.

In general, the objective is to obtain estimates of the first A elements of the $K \times 1$ score vector $\boldsymbol{\tau}$, $\boldsymbol{\tau}_{1:A}$, using the new incomplete multivariate observation $\mathbf{z}$.

In the rest of the paper, without loss of generality, we are considering data matrix $\mathbf{X}$ being mean-centred and scaled to unit variance.

3. METHODS

3.1. Trimmed score method (TRI)

In order to estimate the scores for the new incomplete individual $\mathbf{z}$, the first choice could be to impute their unconditional mean values to missing data, i.e. zero value. By substituting $\mathbf{z}^{\#} = \mathbf{0}$ in the expression

$$\boldsymbol{\tau}_{1:A} = \mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z} = \begin{bmatrix} \mathbf{P}_{1:A}^{\#\mathrm{T}} & \mathbf{P}_{1:A}^{*\mathrm{T}} \end{bmatrix}\begin{bmatrix} \mathbf{z}^{\#} \\ \mathbf{z}^{*} \end{bmatrix} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \tag{7}$$

the so-called trimmed score estimator is obtained:

$$\hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \quad \Rightarrow \quad \hat{\boldsymbol{\tau}}_{a} = \mathbf{p}_{a}^{*\mathrm{T}}\mathbf{z}^{*} \tag{8}$$

From the general model $\mathbf{z} = \mathbf{P}_{1:A}\boldsymbol{\tau}_{1:A} + \mathbf{e}$ it follows that the

trimmed score is the ordinary least squares estimator assuming $\mathbf{z}^{\#} = \mathbf{0}$.

As discussed later on, this choice is quite simple and can be efficient to estimate scores corresponding to loadings with low weights in the missing variables. Nevertheless, if influential variables (those with large weights in the loading vectors) are missing, this method can yield important estimation errors.

From Equations (5)–(8) the estimation error can be expressed as

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} &= \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#} = \mathbf{P}_{1:A}^{\#\mathrm{T}}(\mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{A+1:K}^{\#}\boldsymbol{\tau}_{A+1:K}) \\ &= \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{A+1:K}^{\#}\boldsymbol{\tau}_{A+1:K}\end{aligned}$$

or, equivalently, $\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{e}^{\#}$.

Given that $\mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{e} = \mathbf{0} \Rightarrow \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{e}^{\#} = -\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{e}^{*}$, the expression for the trimmed score estimation error is

$$\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} = (\mathbf{I}_A - \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})\boldsymbol{\tau}_{1:A} - \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{e}^{*} \tag{9}$$

the score error for each component being

$$\tau_a - \hat{\tau}_a = (1 - \mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_a^{*})\tau_a - \sum_{\substack{j=1 \\ j\neq a}}^{k} \mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_j^{*}\tau_j \tag{10}$$

The covariance matrices for the trimmed vector score estimator and for the estimation error are

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\tau}}_{1:A}) &= \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*} \\ \mathrm{Var}(\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A}) &= \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*} + \boldsymbol{\Theta}_{1:A} \\ &\quad - \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A} - \boldsymbol{\Theta}_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\end{aligned} \tag{11}$$

the variance for the ath element of the estimation error being

$$\mathrm{Var}(\tau_a - \hat{\tau}_a) = (1 - \mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_a^{*})^2\lambda_a + \left(\sum_{\substack{j=1 \\ j\neq a}}^{k}(\mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_j^{*})^2\lambda_j\right) \tag{12}$$

The first term in Equation (10) will tend to be large (in absolute value) whenever there exist influential variables with missing values for that component and the actual score value is high. The second term represents the loss of orthogonality between loading vectors due to the missing data ($\mathbf{p}_a^{*}$ and $\mathbf{p}_j^{*}$ are no longer orthogonal for $j \neq a$).

The larger the weights of the missing variables in loading vector $\mathbf{p}_a$, the lower is the squared norm of $\mathbf{p}_a^{*}$, which implies larger values for the first term of the estimator's variance. The second term in Equation (12) shows that this variance also increases with the collinearity between loading vectors induced by missing data. Both terms in Equation (12) are affected by the actual eigenvalues for the K principal components.

3.2. Single-component projection method (SCP) [7]

Nelson *et al.* [7] propose this method based on the NIPALS algorithm [2,5,9]. It is a non-iterative approach where the score calculation step of the NIPALS missing data model-building algorithm is applied to each dimension sequentially.

Let $\mathbf{z}$ be a new incomplete individual with only the last $K - R$ variables measured, $\mathbf{z}^{*}$. Consider $\mathbf{z}(0) = \mathbf{z}$ and let $\mathbf{z}^{*}(a-1)$ be the portion of $\mathbf{z}^{*}(0)$ not explained by the first $a-1$ larger components. To estimate the ath element of the vector score, τ_a (co-ordinate of the new observation in the ath component), the SCP algorithm is based on the simple regression model

$$\mathbf{z}^{*}(a-1) = \tau_a\mathbf{p}_a^{*} + \mathbf{e}^{*}(a) \tag{13}$$

The SCP algorithm minimizes the sum of the squared prediction errors, $\mathbf{e}^{*\mathrm{T}}(a)\mathbf{e}^{*}(a)$, which yields

$$\hat{\tau}_a = \frac{\mathbf{p}_a^{*\mathrm{T}}\mathbf{z}^{*}(a-1)}{\mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_a^{*}} \tag{14}$$

as the least squares estimate of τ_a based on the observed variables. The portion of $\mathbf{z}^{*}(a-1)$ explained by the ath component is then subtracted to yield the deflated object, $\mathbf{e}^{*}(a) = \mathbf{z}^{*}(a)$, and the next component $\hat{\tau}_{a+1}$ is then calculated analogously.

The expression for the estimation error in the first score can be written as

$$\tau_1 - \hat{\tau}_1 = -(\mathbf{p}_1^{*\mathrm{T}}\mathbf{p}_1^{*})^{-1}\mathbf{p}_1^{*\mathrm{T}}\sum_{j=2}^{K}\mathbf{p}_j^{*}\tau_j \tag{15}$$

and, in general, for the ath component ($a = 2, 3, \ldots, A$) as

$$\tau_a - \hat{\tau}_a = -(\mathbf{p}_a^{*\mathrm{T}}\mathbf{p}_a^{*})^{-1}\mathbf{p}_a^{*\mathrm{T}}\left(\sum_{j=1}^{a-1}\mathbf{p}_j^{*}(\tau_j - \hat{\tau}_j) + \sum_{j=a+1}^{K}\mathbf{p}_j^{*}\tau_j\right) \tag{16}$$

When there are no missing measurements, the error is zero. We consider that using all the principal components is equivalent to expressing the new object in a new space spanned by the loading vectors, so with complete data there is no room for error (Equation (3)).

However, in the presence of missing data, Equations (15) and (16) show that the same factors affecting the trimmed score estimation error also influence the SCP estimation error: loss of orthogonality between loading vectors and reduction in the length of the $\mathbf{p}_a$ loading vector. Moreover, in this case the errors made in estimating the earlier scores ($j < a$) also affect the error of the ath score estimate.

3.3. Projection to the model plane method (PMP)

This is a projection method for obtaining all the score estimates at once and has been advocated by Wold *et al.* [10] and Martens and Naes [2]. Nelson *et al.* [7] study its properties.

From Equation (6), if we do not impute any value to missing variables $\mathbf{z}^{\#}$, but only consider the measured variables for the new individual, $\mathbf{z}^{*}$, the model can be expressed as $\mathbf{z}^{*} = \mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} + \mathbf{e}^{*}$, yielding the PMP estimator

$$\hat{\boldsymbol{\tau}}_{1:A} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \tag{17}$$

as the least squares estimator based on the observed variables.

We can work out the score estimation error arising from

Equations (5), (6) and (17) as

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} &= \boldsymbol{\tau}_{1:A} - \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \\ &= \boldsymbol{\tau}_{1:A} - \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} + \mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K}\right) \\ &= -\left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{e}^{*} \qquad (18)\end{aligned}$$

By comparing Equation (18) with the error expression for the SCP method (Equation (16)), it is clear that the propagation of errors from preceding scores into the current estimate (first term in Equation (16)) is absent in the PMP algorithm. This is due to all scores being estimated simultaneously in the projection algorithm. Moreover, the error in PMP does not depend on the actual value of the first A large scores. Equation (18) is analogous to the second term in Equation (16) and represents the errors arising from incorrectly attributing some of the variance in $\mathbf{z}^{*}$ to the scores that are being estimated [7]. This variance is in the direction of the latent variables that have been ignored ($a = A + 1, \ldots, K$).

The PMP method, as with the trimmed score and SCP methods, is also affected by the loss of orthogonality between loading vectors induced by missing data. With columns of $\mathbf{P}_{1:A}^{*}$ nearly collinear, $\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}$ becomes ill-conditioned and the PMP vector score estimation error can become large. In this case a biased regression method such as principal component regression (PCR) [1], ridge regression (RR) [11] or projection to latent structures (PLS) [2] can be used.

The covariance matrix of the vector PMP estimator can be expressed as

$$\mathrm{Var}(\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A}) =$$

$$\left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{A+1:K}^{*}\Theta_{A+1:K}\mathbf{P}_{A+1:K}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}$$

$$\mathrm{Var}(\hat{\boldsymbol{\tau}}_{1:A}) = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1} \qquad (19)$$

where $\Theta_{A+1:K}$ is a $(K-A)\times(K-A)$ diagonal matrix of the lower $K-A$ eigenvalues in decreasing order along the diagonal.

Another difference between the SCP and the PMP algorithm is that SCP is a sequential method, and the value of the ath estimated score does not depend on the total number (A) of estimated scores. Nevertheless, the PMP method estimates the scores simultaneously, the number of scores being decided in advance. Therefore, for example, the first two PMP scores are not the same if one assumes one component, two components, etc. It is important to remark that this difference is not relevant in the context of estimating scores in future individuals, given the fact that the underlying model is assumed to be fixed and known, so the number of significant components (A) is not a matter of controversy.

3.4. Iterative imputation method

Equation (6) can be expressed as

$$\begin{bmatrix}\mathbf{z}^{\#} \\ \mathbf{z}^{*}\end{bmatrix} = \begin{bmatrix}\mathbf{P}_{1:A}^{\#} \\ \mathbf{P}_{1:A}^{*}\end{bmatrix}\boldsymbol{\tau}_{1:A} + \begin{bmatrix}\mathbf{e}^{\#} \\ \mathbf{e}^{*}\end{bmatrix} \qquad (20)$$

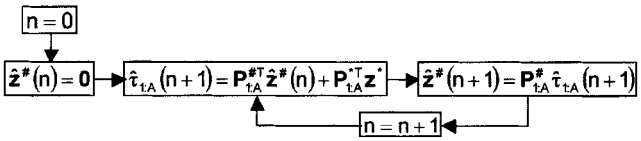

Figure 1. Iterative process for estimating scores for new observations with missing data.

If the missing variables $\mathbf{z}^{\#}$ were known for the new individual, the least squares estimator of $\boldsymbol{\tau}_{1:A}$ would be

$$\hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \qquad (21)$$

Substituting this estimation into Equation (20) yields a new estimation of $\mathbf{z}^{\#}$:

$$\hat{\mathbf{z}}^{\#} = \mathbf{P}_{1:A}^{\#}\boldsymbol{\tau}_{1:A} \qquad (22)$$

which, reincorporated into Equation (21), yields a new estimation of $\boldsymbol{\tau}_{1:A}$. This process, repeated until convergence, constitutes the iterative imputation method. Figure 1 shows a schema of this algorithm. The algorithm initially assumes that $\mathbf{z}^{\#}(0) = \mathbf{0}$, yielding the trimmed score in the first step.

By applying the iterative scheme n times, an estimation of $\mathbf{z}^{\#}$ is obtained (see Appendix I). At convergence the score vector estimator can be expressed as (see Appendix I)

$$\hat{\boldsymbol{\tau}}_{1:A} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \qquad (23)$$

showing that this algorithm is equivalent to the PMP method (Equation (17)).

3.5. Minimization of the squared prediction error method (SPE)

Given a new individual $\mathbf{z}$, its first A co-ordinates in the projected space would be $\boldsymbol{\tau}_{1:A} = \mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z}$, and from Equation (6) these co-ordinates could be reconstructed as $\hat{\mathbf{z}} = \mathbf{P}_{1:A}\boldsymbol{\tau}_{1:A} = \mathbf{P}_{1:A}(\mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z})$, so that $\mathbf{e} = \mathbf{z} - \hat{\mathbf{z}} = \mathbf{z} - (\mathbf{P}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}})\mathbf{z} = (\mathbf{I}_K - \mathbf{P}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}})\mathbf{z}$ is the residual (prediction error). Note that matrix $\mathbf{I}_K - \mathbf{P}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}}$ is idempotent and symmetric. From this it follows that the sum of the squared prediction errors can be expressed as

$$\mathrm{SPE}_{\mathbf{z}} = \mathbf{e}^{\mathrm{T}}\mathbf{e} = \mathbf{z}^{\mathrm{T}}\left(\mathbf{I}_K - \mathbf{P}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}}\right)\mathbf{z} \qquad (24)$$

Equation (24) can be expressed in blocks as a function of $\mathbf{z}^{\#}$ (unknown) and $\mathbf{z}^{*}$ (known). A reasonable estimation criterion would be to choose $\mathbf{z}^{\#}$ so that the sum of the squared prediction errors is minimized for the new observation $\mathbf{z}$. Wise and Ricker [12] adopt the same approach.

Appendix II shows that, by minimizing Equation (24), an estimator of $\mathbf{z}^{\#}$ is obtained. Substituting this into the expression $\hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{\mathrm{T}}\hat{\mathbf{z}} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$, it follows that (see Appendix II)

$$\hat{\boldsymbol{\tau}}_{1:A} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \qquad (25)$$

Therefore, like the iterative method, this algorithm is equivalent to the PMP method (Equation (17)).

Although Nelson *et al.* [7] state that, compared to the PMP algorithm, Wise and Ricker's algorithm is more difficult to implement and does not lend itself readily to error analysis,

we have shown that both algorithms, PMP and SPE, yield the same score vector estimator.

3.6. Known data regression method (KDR)

Once a PCA model (Equation (1)) has been built, and selecting only the A significant components, the model can be expressed as $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} = \mathbf{T}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}} + \mathbf{E}$, where $\mathbf{E} = \mathbf{T}_{A+1:K}\mathbf{P}_{A+1:K}^{\mathrm{T}}$ is the residual matrix of the fitted model. The score matrix for the A significant components is

$$\mathbf{T}_{1:A} = \mathbf{X}\mathbf{P}_{1:A} = \mathbf{X}^{\#}\mathbf{P}_{1:A}^{\#} + \mathbf{X}^{*}\mathbf{P}_{1:A}^{*} \tag{26}$$

The idea of this method is to estimate the scores of a new individual from the training data set $\mathbf{X}$, assuming the same variables to be missing in each row of data matrix $\mathbf{X}$.

From the model

$$\mathbf{T}_{1:A} = \mathbf{X}^{*}\mathbf{B} + \mathbf{U} \tag{27}$$

the least squares estimator of matrix $\mathbf{B}$ is $\hat{\mathbf{B}} = (\mathbf{X}^{\mathbf{*}\mathrm{T}}\mathbf{X}^{\mathbf{*}})^{-1}\mathbf{X}^{\mathbf{*}\mathrm{T}}\mathbf{T}_{1:A}$, where $\mathbf{X}^{\mathbf{*}}$ contains the $K-R$ columns of data matrix $\mathbf{X}$ corresponding to the measured variables for the new incomplete individual $\mathbf{z}$. The score vector for the new individual can be estimated as $\hat{\boldsymbol{\tau}}_{1:A} = \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{z}^{*}$, which, as shown in Appendix III, yields

$$\hat{\boldsymbol{\tau}}_{1:A} = \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{z}^{*} \tag{28}$$

where $\mathbf{P}^{\mathbf{*}}\Theta\mathbf{P}^{\mathbf{*}\mathrm{T}}$ is a $(K-R)\times(K-R)$ square matrix and may be very ill-conditioned with highly correlated data. (Note that $\mathbf{P}^{\mathbf{*}}\Theta\mathbf{P}^{\mathbf{*}\mathrm{T}} = \mathbf{X}^{\mathbf{*}\mathrm{T}}\mathbf{X}^{\mathbf{*}}/(N-1)$ is the sample covariance matrix of $\mathbf{X}$.) In this case, as in the PMP method, a biased regression method such as principal component regression (PCR) [1], ridge regression (RR) [11] or projection to latent structures (PLS) [2] can be used. The difference is that $K-R$ is expected to be much greater than A, and thus the PMP algorithm has the advantage of only needing the inverse of a much smaller $A\times A$ matrix.

The score vector error is $\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} = \boldsymbol{\tau}_{1:A} - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}})^{-1}(\mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} + \mathbf{e}^{*})$, and, by substituting Equation (5),

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} = {} & \left[\mathbf{I}_A - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{*}\right]\boldsymbol{\tau}_{1:A} \\ & - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K}\end{aligned} \tag{29}$$

the covariance matrix of the vector score estimator being

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A}) &= \left[\mathbf{I}_A - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{*}\right]\Theta_{1:A} \\ \mathrm{Var}(\hat{\boldsymbol{\tau}}_{1:A}) &= \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{*}\Theta_{1:A}\end{aligned} \tag{30}$$

3.7. Conditional mean replacement method (CMR)

This method has been proposed by Nelson *et al.* [7] and assumes that $\mathbf{z}$ follows a multivariate normal distribution with mean vector $E(\mathbf{z}) = \mathbf{0}$ and estimated covariance matrix $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/(N-1)$. Instead of imputing their unconditional mean values (as in the trimmed score method) to missing data, this algorithm replaces the missing variable values $\mathbf{z}^{\#}$ with the expected values from the conditional normal distribution, given the present data and the current estimate of the means and covariance, i.e.

$$\hat{\mathbf{z}}^{\#} = E\left[\mathbf{z}^{\#}/\mathbf{z}^{*}, \mathbf{S}\right] \tag{31}$$

Substituting the expression of the estimator of $\mathbf{z}^{\#}$ into $\hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$ yields

$$\hat{\boldsymbol{\tau}}_{1:A} = \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}(\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}})^{-1}\mathbf{z}^{*} \tag{32}$$

This is the same expression as Equation (28), showing the equivalence between the KDR and CMR methods. Nelson *et al.* [7] comment on this equivalence, although they only develop the CMR method.

This method is equivalent to the expectation step of the expectation–maximization (EM) algorithm [4], since we assume that we have already built a PCA model and are only interested in handling missing data in future multivariate observations. We consider that the addition of the information in any new individual will not change the estimates of the mean vector and covariance matrix. Then we use only the expectation step to calculate replacement values for missing data.

3.8. Trimmed score regression method (TSR)

As explained in Section 3.1, the trimmed score method is a simple method to estimate the scores of a new incomplete individual. The problem is that it can lead to large score estimation errors in those individuals where influential variables are missing.

The present method is a modification of the KDR method in the sense that we try to estimate the scores of a new individual not from matrix $\mathbf{X}^{\mathbf{*}}$ (as in the KDR algorithm), but from matrix $\mathbf{T}_{1:A}^{*} = \mathbf{X}^{*}\mathbf{P}_{1:A}^{*}$, assuming the same variables to be missing in each row of data matrix $\mathbf{X}$. Rows of matrix $\mathbf{T}_{1:A}^{*}$ can be interpreted as the trimmed scores corresponding to the N individuals of the training data matrix $\mathbf{X}$.

The idea is to reconstruct $\mathbf{T}_{1:A}$ from the trimmed scores $\mathbf{T}_{1:A}^{*}$ using the model

$$\mathbf{T}_{1:A} = \mathbf{T}_{1:A}^{*}\mathbf{B} + \mathbf{U} \tag{33}$$

This method constitutes an improvement of the trimmed score method through a regression model. As shown in Appendix IV, the estimator obtained in this case is

$$\hat{\boldsymbol{\tau}}_{1:A} = \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \tag{34}$$

and the expression for the error of the score vector is

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} = {} & \\ & [\mathbf{I}_A - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}]\boldsymbol{\tau}_{1:A} \\ & - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K}\end{aligned} \tag{35}$$

the covariance matrix of the score vector estimator being

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A}) = {} & \\ & [\mathbf{I}_A - \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}]\Theta_{1:A} \\ \mathrm{Var}(\hat{\boldsymbol{\tau}}_{1:A}) = {} & \Theta_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^{*}\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\Theta_{1:A}\end{aligned} \tag{36}$$

In this method, as in the KDR algorithm, all the principal components need to be extracted to obtain matrices $\mathbf{P}^*$ and Θ. Nevertheless, the advantage of this method is that the matrix to be inverted, $(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}^*\Theta\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^*)^{-1}$, is $A \times A$, much smaller than in the KDR method. In any case this matrix may be very ill-conditioned with highly correlated data, and then biased regression methods can be used.

It can be shown that this method is equivalent to the PMP method only if data matrix $\mathbf{X}$ is of rank A, which means that $\lambda_{A+1} = \lambda_{A+2} = \ldots = \lambda_K = 0$, and then, when extracting the A principal components, there is no error left.

4. INDUSTRIAL EXAMPLES

The properties of each of the methods for handling missing data for new individuals presented in this paper are illustrated through simulation using three industrial data sets. The first data set is studied in detail and the other two are employed for underlining and checking the results obtained.

For every data set, using cross-validation [11,13], a PCA model with three components ($A = 3$) has been built.

4.1. Mineral sorting plant (SOVR)

This data set is part of a known industrial data set from SIMCA-P 8.0 software [8]. The data come from a mineral sorting plant at LKAB in Sweden. In this process, raw iron ore is divided into finer material by passing through several grinders. For illustrative purposes, out of the 12 process variables, only eight have been selected for this study. From the 230 observations available, we have employed 150 for building the PCA model (calibration set), the remaining 80 being used as new individuals (test set). The first three larger PCA components jointly explain 94% of the total variance of the process variables ($R^2 = 0.94$), the remaining 6% being due to noise. The predictive ability is high ($Q^2 = 0.78$).

Table I shows the weights of each of the eight process variables in the three loading vectors. Variables X_1, X_2, X_3, X_7 and X_8 influence the first principal direction. The second principal direction is strongly dependent upon variables X_4 and X_6. Finally, variable X_5 creates the third principal direction. When the number of latent variables is small, inspecting this kind of weight table (or its corresponding loading plot) for several combinations of latent variables can give some insight into the impact of missing measurements in certain variables or sets of variables. This can also be used to determine which missing measurements will have the most impact on score calculation [7]. In this case, from Table I it is clear that whenever variable X_5 is missing, this will cause a large error in the estimate of τ_3. On the other hand, if the new individual is missing both variables X_4 and X_6, most of the information explaining the second component is lost, leading to large errors in the estimation of the second score τ_2. In this example, given that component 1 is related to five of the original process variables, it is expected that even in the extreme case of having three missing variables (representing 38% of missing data) the estimation of τ_1 may not be seriously affected.

Despite the fact that loading plots or tables of loading weights are useful tools to anticipate the effect of missing data combinations on score errors, they may fail to indicate critical combinations that can only be detected by simulation. In this example, all 92 combinations of one, two and three groups of missing variables (13%–38% of the total number of variables) have been simulated. In each case, each of the 80 observations in the validation set has been considered as if it were a new incomplete individual with missing variables corresponding to the particular combination. For each observation and for each combination of one, two and three missing variables the squared errors (SE) for the score vector estimation and for each of the three score values are worked out. SE measures the difference between the score estimates from incomplete data and those calculated with the complete data set.

Table I. SOVR. Loading vector weights for the three significant principal components

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
$\mathbf{p}_1$	0.4457	0.4410	0.4465	0.1841	0.0200	0.0509	0.4319	0.4290
$\mathbf{p}_2$	0.0844	0.0758	0.0635	−0.6642	−0.2166	−0.6949	0.1003	0.0450
$\mathbf{p}_3$	0.0170	0.0859	−0.0308	0.0024	0.9450	−0.2916	0.0275	−0.1121

Given the equivalences between several of the methods presented in Section 3, the methods used in the simulation are trimmed scores (TRI), single-component projection (SCP), projection to the model plane (PMP), regression on known data (KDR) and trimmed score regression (TSR).

To ease the interpretation of the results, for each missing data combination the following characteristics are calculated:

- $\left\{m_{ij} = \mathbf{p}_i^{*\mathrm{T}}\mathbf{p}_j^* | i,j = 1,2,3\right\}$, six values;
- $\left\{a_{ij} = \angle(\mathbf{p}_i^*, \mathbf{p}_j^*) | i,j = 1,2,3; i \neq j\right\}$, three values.

m_{ii} represents the squared length of the ith loading vector after removing the co-ordinates corresponding to missing data. If missing variables have low weight in $\mathbf{p}_i$, then m_{ii} will be close to 1. On the other hand, had influential variables for the ith component been missing, m_{ii} would be close to 0.

m_{ij} (for $i \neq j$) is the inner product of two loading vectors. In PCA, loading vectors are orthogonal and so this product is zero. Nevertheless, with missing data, loading vectors become collinear, angles a_{ij} ($i \neq j$) being related to m_{ij} through the expression

$$a_{ij} = \cos^{-1}\left(\frac{m_{ij}}{\sqrt{m_{ii}m_{jj}}}\right)$$

When there are no missing data, $a_{ij} = 90°$.

The loss of orthogonality due to missing data may affect the PMP, KDR and TSR methods, making the matrices that need to be inverted in each method ill-conditioned and yielding large score errors. Nevertheless, in all the simulations run, only the PMP method required a biased regression method to overcome this problem. Even in the highest-impact missing value sets neither the KDR nor the TSR method suffered from the ill-conditioning problem, and there was no need to use any biased regression method.

The results obtained from the simulations match the

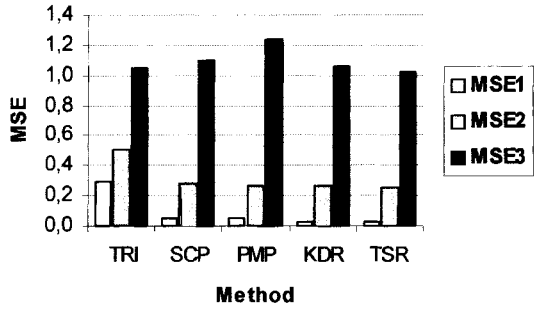

Figure 2. SOVR. Mean squared error (MSE) for every score estimation in a high-impact case (variables X_2, X_4 and X_5 are missing) for the different methods.

expected results drawn from Table I. The outcome in one of the highest-impact missing value sets is shown in Figure 2 and Table II. This table shows the squared lengths (m_{ii}) of and the angles (a_{ij}) between the loading vectors $\mathbf{p}^*$ together with the mean squared errors for every score estimation (MSE_a; $a = 1, 2, 3$) and for the score vector estimation (MSE_{TOTAL}) in the methods studied. Figure 2 displays the MSE_a values. In this case, three out of the eight process variables (38% missing data) are missing, yielding a dramatic reduction in the length of the loading vectors and inducing collinearity between them. In this case the missing variables are X_2, X_4 and X_5, being influential variables for the first, second and third dimensions respectively. Nevertheless, given the structure of the model in Table I, their impact on the score estimates is not expected to be the same. From Table I we can realize that the amount of information lost is about 20% (one out of five influential variables missing) in the first dimension, 50% in the second dimension (one out of two) and practically 100% in the third dimension (the only influential variable in the third dimension, X_5, is missing). Therefore in this case, the second and mainly third dimensions are expected to be largely affected by missing data. This is exactly what the simulation shows as presented in Table II and Figure 2, where in all the methods the largest error is produced in the estimate of τ_3, followed by τ_2. Note the dramatic reduction in the squared norm of $\mathbf{p}_3$ (m_{33}) and the collinearity between the $\mathbf{p}_2$ and $\mathbf{p}_3$ loading vectors (measured by the low value of the inner product m_{23} or, equivalently, by their small angle a_{23}).

In order to compare the mean value of the SE in the five estimation methods, an analysis of variance (ANOVA) was run using missing data combination and observation as block factors. This allows comparison of the five methods to be made in similar conditions, increasing the discriminant ability of the comparisons. Given the positive skewness of the SE, a logarithmic transformation was applied to the SE variable. Figure 3 displays the least significance difference (LSD) intervals for the average SE for each of the five methods. The inferiority of the trimmed score method (TRI) and the single-component projection method (SCP) with respect to the other three methods (PMP, KDR and TSR) is statistically significant, the TRI method being the worst (from the SE point of view). The trimmed score regression method (TSR) is shown to be slightly superior to the projection to the model plane method (PMP), with the known data regression method (KDR) being the best (from the SE point of view). In any case the behaviour of these three methods (PMP, KDR and TSR) is quite similar.

Table II. SOVR. High-impact case (variables X_2, X_4 and X_5 are missing) for the different methods. Mean squared error (MSE_a) for every score estimation and for the score vector estimation (MSE_{TOTAL}). Squared lengths (m_{ii}) of and angles (a_{ij}) between the loading vectors $\mathbf{p}^*$

				TRI	SCP	PMP	KDR	TSR
m_{11}	m_{22}	m_{33}	MSE_1	0.2952	0.0550	0.0563	0.0296	0.0281
−0.7712	0.5062	0.0996	MSE_2	0.5107	0.2800	0.2663	0.2601	0.2472
a_{12}	a_{13}	a_{23}	MSE_3	1.0482	1.1033	1.2316	1.0589	1.0252
81.4	78.1	27.1	MSE_{TOTAL}	1.8541	1.4383	1.5542	1.3486	1.3004

In order to visualize the impact that the different estimation methods have on the score estimation with missing data, Figure 4 displays the SE of the vector estimation for the first 10 observations from the validation set in the five methods assuming that the X_1 (linked to the first score) and X_4 (linked to the second score) variables are missing. As expected from the results obtained in the ANOVA, the TRI method leads to large score errors, with PMP, KDR and TSR behaving in a very similar way.

4.2. Polyurethane seats data set (SEATS)

The data come from a factory manufacturing seats in Spain. Polyalcohol and isocyanate are injected in a mould to produce polyurethane seats. In this case we have 61 observations with nine process variables ($K = 9$); 41 observations are used for building the PCA model (calibration set) and the remaining 20 observations for simulating new individuals (test set) ($R^2 = 0.81$ and $Q^2 = 0.41$). Following the same methodology as in the SOVR data set, very similar results are obtained. Figure 5 shows the LSD intervals for the average SE from the ANOVA for each of the five estimation methods under study. In this case, logarithmic transformation of the SE was not needed.

4.3. High-density polyethylene data set (HDPE)

This data set comes from a petrochemical company is Spain. A commercial-scale polymerization process produces large

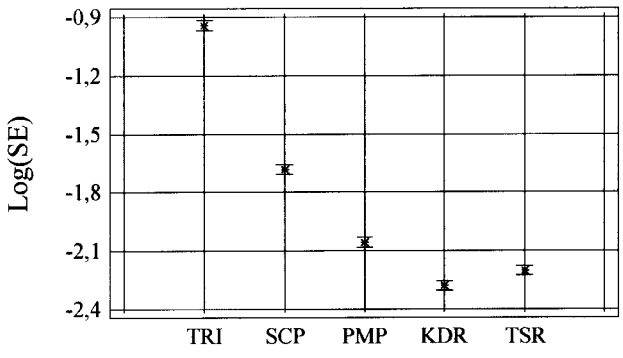

Figure 3. SOVR. Least significance difference (LSD) intervals for the average of the logarithm of the squared error (SE) for the score vector estimation in the methods under study.

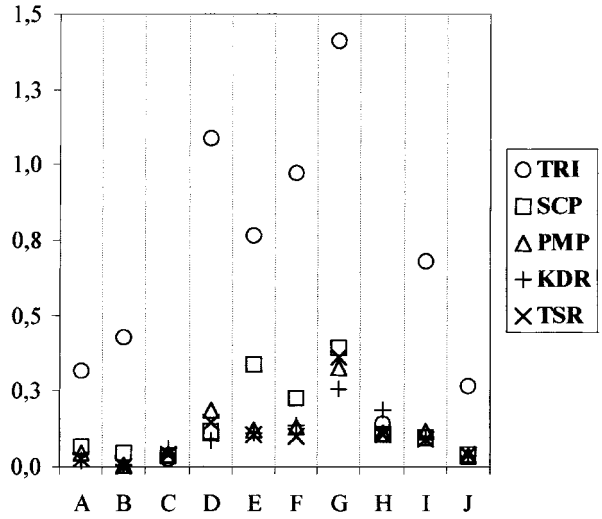

Figure 4. SOVR. Squared error (SE) for the score vector estimation in the five methods, assuming variables X_1 and X_4 are missing, for the first 10 observations (A, B, …, J) from the validation set.

volumes of a polymer (high-density polyethylene) used in many familiar consumer products. From the 91 observations with 13 process variables ($K = 13$), 61 have been used for building the PCA model (calibration set) and the remaining 30 for simulating new individuals (test set) ($R^2 = 0.85$ and $Q^2 = 0.71$). The results obtained from this data set match the previous results from the SOVR and SEATS data sets. This is shown in Figure 6, where the LSD intervals for the average SE from the ANOVA for each of the five estimation methods under study are displayed. In this case, logarithmic transformation of the SE was not needed.

5. CONCLUSIONS

This paper addresses the problem of using future multivariate observations with missing data to estimate latent variable scores from an existing PCA model. This is a critical issue in multivariate statistical process control (MSPC) schemes where the process is continuously interrogated based on an underlying PCA model. If PCA models required complete data sets (with no missing data), this would

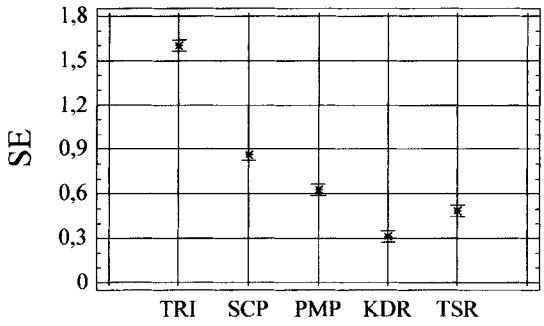

Figure 5. SEATS. Least significance difference (LSD) intervals for the average of the squared error (SE) for the score vector estimation in the methods under study.

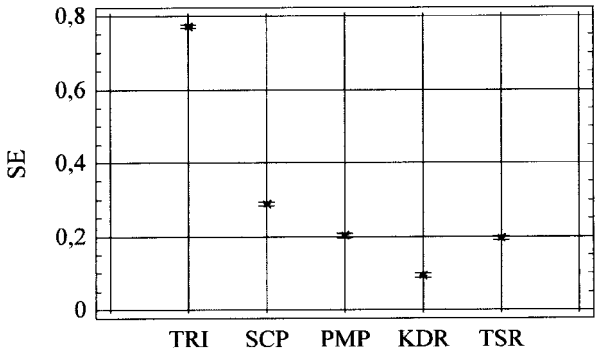

Figure 6. HDPE. Least significance difference (LSD) intervals for the average of the squared error (SE) for the score vector estimation in the methods under study.

involve throwing away large amounts of data, given that missing measurements are a common occurrence both in continuous and batch processes.

Several methods for estimating the scores of new individuals with missing data are presented: a so-called trimmed score method (TRI), a single-component projection method (SCP), a method of projection to the model plane (PMP), a method based on the iterative imputation of missing data, a method based on the minimization of the squared prediction error (SPE), a conditional mean replacement method (CMR) and various least squared-based methods: one based on a regression on known data (KDR) and the other based on a regression on trimmed scores (TSR).

The basis for each method and the expressions for the score estimators, their covariance matrices and the estimation errors are developed. Some of the methods discussed have already been proposed in the literature (SCP, PMP and CMR), some are original (TRI and TSR) and others are shown to be equivalent to methods already developed by other authors: iterative imputation and SPE methods are equivalent to PMP; KDR is equivalent to CMR.

Several methods can be seen as different ways to impute values for the missing variables, vector $\mathbf{z}^{\#}$. If $\mathbf{z}^{\#} = \mathbf{0}$ (unconditional mean value), this yields the TRI method. The PMP method assumes that no imputation value is given to missing variables $\mathbf{z}^{\#}$, and only the measured variables for the new individual $\mathbf{z}^{*}$ are considered to estimate the scores. PMP is shown to be equivalent to the iterative imputation method where $\mathbf{z}^{\#}$ is re-estimated until convergence. If vector $\mathbf{z}^{\#}$ is chosen so that the sum of the squared prediction errors is minimized for the new individual (SPE method), this also yields the PMP method. Finally, if missing values are replaced by their conditional means, given the measured variables and the estimate of the means and covariance matrix, $\hat{\mathbf{z}}^{\#} = E[\mathbf{z}^{\#}/\mathbf{z}^{*}, \mathbf{S}]$ this yields the CMR method. This method is shown to be equivalent to the least squares estimate of the scores based on the training data set, assuming that the same variables are missing in each row of $\mathbf{X}$ (KDR method). TSR uses the trimmed score estimators corresponding to the training data set to improve the score estimations by least squares regression.

The expressions for the score estimation errors from all the methods show that the main factors affecting the score

estimation errors under missing data are the loss of orthogonality between loading vectors and the reduction in the squared length of the $\mathbf{p}_a$ loading vectors. Moreover, in the SCP method, errors are shown to propagate from estimated scores to subsequently calculated ones through deflation. This is the only method which extracts the scores sequentially, and not at once like the rest of the methods studied.

The efficiency of the methods is studied through simulations based on three industrial data sets. Squared estimation errors are used as the basis for comparisons.

In the PMP, KDR and TSR methods, some matrices need to be inverted. In the KDR method the matrix to be inverted is $(K-R)\times(K-R)$ (K is the number of process variables and R is the number of missing variables). TSR and PMP have the advantage of only needing the inverse of a much smaller $A\times A$ matrix. These matrices may be ill-conditioned owing to missing data. This can lead to large score errors, and in such cases, biased regression methods such as ridge regression should be used. Nevertheless, in all the simulations run, only the PMP method required a biased regression method to overcome this problem. Even in the highest-impact missing value sets neither the KDR nor the TSR method suffered from the ill-conditioning problem, and there was no need to use any biased regression method.

KDR is shown to be statistically superior to the other methods (from the SE point of view). Nevertheless, the TSR method is practically equivalent to the KDR method and has the advantage of having to invert a matrix of a much smaller size. Additionally, TSR is statistically superior to the PMP method.

Expressions for the covariance matrix of the score vector estimators may be used to modify the classical statistical control limits of the scores to monitor new individuals with missing data. This can be used for on-line MSPC and to determine if the monitoring scheme (based on the PCA model) can continue to operate successfully in the presence of missing data.

Given a particular PCA model built on a stable industrial process, several tools such as loading plots, tables of loading weights, and simulations of various missing data combinations can be useful to evaluate off-line the future impact of combination of missing data on score estimation. Such studies could assist the design of sensor maintenance schedules.

Acknowledgements

This research was partially supported by the Spanish Government (CICYT) and the European Union (RDE funds) under grant 1FD1997-2159. We would like to thank Professor Rafael Romero for providing us with the SEATS data set.

APPENDIX I. DERIVATION OF THE ITERATIVE IMPUTATION METHOD

As shown in Figure 1, $\boldsymbol{\tau}_{1:A}$ is estimated using the value of $\mathbf{z}^{\#}$ given in the previous step:

$$\hat{\boldsymbol{\tau}}_{1:A}(n)=\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#}(n-1)+\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$$

With this score vector estimation of $\boldsymbol{\tau}_{1:A}$, $\mathbf{z}^{\#}$ is re-estimated through iteration:

$$\begin{aligned}
\hat{\mathbf{z}}^{\#}(n)&=\mathbf{P}_{1:A}^{\#}\hat{\boldsymbol{\tau}}_{1:A}(n)\\
&=\mathbf{P}_{1:A}^{\#}\left[\mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#}(n-1)+\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\right]\\
&=\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#}(n-1)+\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\\
&=\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\left[\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#}(n-2)+\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\right]+\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\\
&=\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{2}\hat{\mathbf{z}}^{\#}(n-2)+\left(\mathbf{I}_R+\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\\
&=\cdots\\
&=\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{n}\hat{\mathbf{z}}^{\#}(0)+[\mathbf{I}_R+\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\\
&\quad+\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{2}+\cdots+\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{n-1}]\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}
\end{aligned}$$

The second term is the sum of the first n terms of a geometric progression of matrices whose first term is $\mathbf{I}_R$ and whose common ratio is $\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}$.

Let $\mathbf{S}_n$ be the sum of the first n terms of a geometric progression of matrices whose first term is $\mathbf{B}$ and whose common ratio is $\mathbf{Q}$. We can write the expressions

$$\mathbf{S}_n=\mathbf{B}+\mathbf{BQ}+\mathbf{BQ}^2+\cdots+\mathbf{BQ}^{n-1}$$

$$\mathbf{S}_n\mathbf{Q}=\mathbf{BQ}+\mathbf{BQ}^2+\mathbf{BQ}^3+\cdots+\mathbf{BQ}^{n}$$

If the second term is subtracted from the first, we get $\mathbf{S}_n(\mathbf{Q}-\mathbf{I})=\mathbf{B}(\mathbf{Q}^n-\mathbf{I})$ and, resolving for $\mathbf{Q}-\mathbf{I}$, $\mathbf{S}_n=\mathbf{B}(\mathbf{Q}^n-\mathbf{I})(\mathbf{Q}-\mathbf{I})^{-1}$ is obtained.

In our case, $\mathbf{B}=\mathbf{I}_R$ and $\mathbf{Q}=\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}$, leading to

$$\hat{\mathbf{z}}^{\#}(n)=\mathbf{0}+\left[\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{n}-\mathbf{I}_R\right]\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}-\mathbf{I}_R\right)^{-1}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$$

Thus iteration gives

$$\begin{aligned}
\hat{\mathbf{z}}^{\#}&=\lim_{n\to\infty}\left(\hat{\mathbf{z}}^{\#}(n)\right)=-\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}-\mathbf{I}_R\right)^{-1}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\\
&=\left(\mathbf{I}_R-\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}
\end{aligned}$$

as $\left(\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{n}\overset{n\to\infty}{\longrightarrow}\bar{\bar{\mathbf{0}}}_{R\times R}$, provided that $0<R\leq K-A$ (to guarantee that all singular values of $\mathbf{P}_{1:A}^{\#}$ are less than 1). Replacing $\mathbf{z}^{\#}$ with the estimator found in the expression of $\boldsymbol{\tau}_{1:A}$ gives

$$\begin{aligned}
\hat{\boldsymbol{\tau}}_{1:A}&=\mathbf{P}_{1:A}^{\mathrm{T}}\hat{\mathbf{z}}=\mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#}+\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}\\
&=\left[\mathbf{I}_A+\mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R-\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\right]\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}
\end{aligned}$$

but

$$\mathbf{I}_A+\mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R-\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}=\left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\qquad(37)$$

since

$$\left[\mathbf{I}_A + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\right](\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*})$$

$$= \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}$$

and $\mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{P}_{1:A} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*} = \mathbf{I}_A$, so that the previous expression can be written as

$$\left(\mathbf{I}_A - \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\right) + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\left(\mathbf{I}_A - \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\right)$$

$$= \mathbf{I}_A - \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\left(\mathbf{P}_{1:A}^{\#} - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\right)$$

$$= \mathbf{I}_A - \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)\mathbf{P}_{1:A}^{\#}$$

$$= \mathbf{I}_A - \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} = \mathbf{I}_A$$

yielding

$$\hat{\boldsymbol{\tau}}_{1:A} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \tag{38}$$

APPENDIX II. MINIMIZATION OF THE SPE IN TERMS OF $\mathbf{z}^{\#}$

Given a new individual with missing data, because of the notation used, we can rewrite Equation (24) in blocks as

$$\mathrm{SPE}_{\mathbf{z}} = \begin{bmatrix}\mathbf{z}^{\#\mathrm{T}} & \mathbf{z}^{\#\mathrm{T}}\end{bmatrix}\left(\begin{bmatrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{bmatrix} - \begin{bmatrix}\mathbf{P}_{1:A}^{\#}\\ \mathbf{P}_{1:A}^{*}\end{bmatrix}\begin{bmatrix}\mathbf{P}_{1:A}^{\#\mathrm{T}} & \mathbf{P}_{1:A}^{*\mathrm{T}}\end{bmatrix}\right)\begin{bmatrix}\mathbf{z}^{\#}\\ \mathbf{z}^{*}\end{bmatrix}$$

From the previous expression it follows that

$$\mathrm{SPE}_{\mathbf{z}} = \mathbf{z}^{\mathrm{T}}\mathbf{z} - \mathbf{z}^{\mathrm{T}}\mathbf{P}_{1:A}\mathbf{P}_{1:A}^{\mathrm{T}}\mathbf{z}$$

$$= \mathbf{z}^{\#\mathrm{T}}\mathbf{z}^{\#} + \mathbf{z}^{*\mathrm{T}}\mathbf{z}^{*} - (\mathbf{z}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#}$$

$$+ 2\mathbf{z}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{*}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{*} + \mathbf{z}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*})$$

The objective is to minimize this expression in terms of $\mathbf{z}^{\#}$:

$$\frac{\partial(\mathrm{SPE}_{\mathbf{z}})}{\partial\mathbf{z}^{\#}} = 2\mathbf{z}^{\#} - \left(2\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{\#} + 2\mathbf{P}_{1:A}^{*}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{z}^{*}\right) = \mathbf{0}$$

resulting in

$$\hat{\mathbf{z}}^{\#} = \left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$$

From the estimation of $\mathbf{z}^{\#}$ we can estimate the score vector as

$$\hat{\boldsymbol{\tau}}_{1:A} = \mathbf{P}_{1:A}^{\#\mathrm{T}}\hat{\mathbf{z}}^{\#} + \mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$$

$$= \left[\mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#} + \mathbf{I}_A\right]\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*}$$

However, according to Equation (37),

$$\mathbf{I}_A + \mathbf{P}_{1:A}^{\#\mathrm{T}}\left(\mathbf{I}_R - \mathbf{P}_{1:A}^{\#}\mathbf{P}_{1:A}^{\#\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{\#} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}$$

is fulfilled and thus we can write

$$\hat{\boldsymbol{\tau}}_{1:A} = \left(\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right)^{-1}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{z}^{*} \tag{39}$$

APPENDIX III. DERIVATION OF THE KNOWN DATA REGRESSION METHOD

From the linear model $\mathbf{T}_{1:A} = \mathbf{X}^{*}\mathbf{B} + \mathbf{U}$ seen in Equation (27), the least squares estimator matrix is

$$\hat{\mathbf{B}} = \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{T}_{1:A} \tag{40}$$

The scores of the new individual $\mathbf{z}$, with missing data, can be estimated through the expression

$$\hat{\boldsymbol{\tau}}_{1:A} = \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{z}^{*} \tag{41}$$

We derive estimator $\hat{\mathbf{B}}$ in Equation (40) by considering the following expressions:

$$\mathbf{T}_{1:A} = \mathbf{X}\mathbf{P}_{1:A} = \mathbf{X}^{\#}\mathbf{P}_{1:A}^{\#} + \mathbf{X}^{*}\mathbf{P}_{1:A}^{*} \tag{42}$$

$$\mathbf{X}^{\#} = \mathbf{T}\mathbf{P}^{\#\mathrm{T}}, \qquad \mathbf{X}^{*} = \mathbf{T}\mathbf{P}^{*\mathrm{T}} \tag{43}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{T}_{1:A}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\left(\mathbf{X}^{\#}\mathbf{P}_{1:A}^{\#} + \mathbf{X}^{*}\mathbf{P}_{1:A}^{*}\right)$$

$$= \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{\#}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{*}$$

$$\hat{\mathbf{B}} = \left(\mathbf{P}^{*}\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}^{*}\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{P}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{*}$$

$$\hat{\mathbf{B}} = \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{*}$$

$$\hat{\mathbf{B}} = \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\begin{bmatrix}\mathbf{P}_{1:A}^{*} & \mathbf{P}_{A+1:K}^{*}\end{bmatrix}$$

$$\begin{bmatrix}\boldsymbol{\Theta}_{1:A} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Theta}_{A+1:K}\end{bmatrix}\begin{bmatrix}\mathbf{P}_{1:A}^{\#\mathrm{T}}\\ \mathbf{P}_{A+1:K}^{\#\mathrm{T}}\end{bmatrix}\mathbf{P}_{1:A}^{\#} + \mathbf{P}_{1:A}^{*}$$

$$= \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\left(\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A}\mathbf{P}_{1:A}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\right.$$

$$\left. + \mathbf{P}_{A+1:K}^{*}\boldsymbol{\Theta}_{A+1:K}\mathbf{P}_{A+1:K}^{\#\mathrm{T}}\mathbf{P}_{1:A}^{\#}\right) + \mathbf{P}_{1:A}^{*}$$

$$= \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\left(\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A} - \mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right.$$

$$\left. - \mathbf{P}_{A+1:K}^{*}\boldsymbol{\Theta}_{A+1:K}\mathbf{P}_{A+1:K}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right) + \mathbf{P}_{1:A}^{*}$$

$$= \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\left(\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A} - \mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}_{1:A}^{*}\right) + \mathbf{P}_{1:A}^{*}$$

$$= \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A} - \mathbf{P}_{1:A}^{*} + \mathbf{P}_{1:A}^{*}$$

Thus $\hat{\mathbf{B}} = \left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{P}_{1:A}^{*}\boldsymbol{\Theta}_{1:A}$. From this, Equation (41) is

$$\hat{\boldsymbol{\tau}}_{1:A} = \boldsymbol{\Theta}_{1:A}\mathbf{P}_{1:A}^{*\mathrm{T}}\left(\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\right)^{-1}\mathbf{z}^{*} \tag{44}$$

APPENDIX IV. DERIVATION OF THE TRIMMED SCORE REGRESSION METHOD

For a new individual $\mathbf{z}$, with missing data, $\boldsymbol{\tau}^*_{1:A}$ will be the corresponding estimator to the trimmed score method, i.e. $\boldsymbol{\tau}^*_{1:A} = \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{z}^*$.

We know the scores of each of the N individuals used in the construction of the model, and we can calculate their scores using the trimmed score method, assuming that the first R variables are unknown:

$$\boldsymbol{\tau}^*_{1:A}(i) = \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{z}^*_i \quad (i = 1, \cdots, N) \tag{45}$$

The scores are summarized in matrix $\mathbf{T}_{1:A}$ of N rows and A columns, thereby constructing matrix $\mathbf{T}^*_{1:A}$ with the same dimensions as $\mathbf{T}_{1:A}$, so that each row contains the A coordinates of the estimation of Equation (45). Hence $\mathbf{T}^*_{1:A} = \mathbf{X}^*\mathbf{P}^*_{1:A}$.

The linear model used to reconstruct $\mathbf{T}_{1:A}$ from $\mathbf{T}^*_{1:A}$ is

$$\mathbf{T}_{1:A} = \mathbf{T}^*_{1:A}\mathbf{B} + \mathbf{U} \tag{46}$$

the least squares estimator of the matrix of coefficients $\mathbf{B}$ being

$$\hat{\mathbf{B}} = \left(\mathbf{T}^{*\mathrm{T}}_{1:A}\mathbf{T}^*_{1:A}\right)^{-1}\mathbf{T}^{*\mathrm{T}}_{1:A}\mathbf{T}_{1:A}$$

In the same way as $\mathbf{T}^*_{1:A}$ was defined, we can express $\mathbf{T}^\#_{1:A} = \mathbf{X}^\#\mathbf{P}^\#_{1:A}$, proving that

$$\mathbf{T}_{1:A} = \mathbf{T}^\#_{1:A} + \mathbf{T}^*_{1:A}$$

Hence we derive the expression of the estimator as

$$\begin{aligned}
\hat{\mathbf{B}} &= \left(\mathbf{T}^{*\mathrm{T}}_{1:A}\mathbf{T}^*_{1:A}\right)^{-1}\mathbf{T}^{*\mathrm{T}}_{1:A}\left(\mathbf{T}^\#_{1:A} + \mathbf{T}^*_{1:A}\right) \\
&= \mathbf{I}_A + \left(\mathbf{T}^{*\mathrm{T}}_{1:A}\mathbf{T}^*_{1:A}\right)^{-1}\mathbf{T}^{*\mathrm{T}}_{1:A}\mathbf{T}^\#_{1:A} \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\mathbf{P}^*_{1:A}\right)^{-1}\left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^\#\mathbf{P}^\#_{1:A}\right)
\end{aligned}$$

By using the expressions in (43), we can write

$$\begin{aligned}
\hat{\mathbf{B}} &= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{P}^{\#\mathrm{T}}\mathbf{P}^\#_{1:A}\right) \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{\#\mathrm{T}}\mathbf{P}^\#_{1:A}\right) \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\Big(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A}\mathbf{P}^{\#\mathrm{T}}_{1:A}\mathbf{P}^\#_{1:A} \\
&\quad + \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{A+1:K}\boldsymbol{\Theta}_{A+1:K}\mathbf{P}^{\#\mathrm{T}}_{A+1:K}\mathbf{P}^\#_{1:A}\Big) \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\big(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A} \\
&\quad - \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A} - \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{A+1:K}\boldsymbol{\Theta}_{A+1:K}\mathbf{P}^{*\mathrm{T}}_{A+1:K}\mathbf{P}^*_{1:A}\big) \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^{**\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A} - \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right) \\
&= \mathbf{I}_A + \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A} - \mathbf{I}_A
\end{aligned}$$

concluding that

$$\hat{\mathbf{B}} = \left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\boldsymbol{\Theta}_{1:A} \tag{47}$$

By applying Equation (46) to the new individual $\mathbf{z}$, considering that $\boldsymbol{\tau}^*_{1:A} = \mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{z}^*$, then

$$\begin{aligned}
\hat{\boldsymbol{\tau}}_{1:A} &= \hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\tau}^*_{1:A} = \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{z}^* \\
&= \boldsymbol{\Theta}_{1:A}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*_{1:A}\left(\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\mathrm{T}}\mathbf{P}^*_{1:A}\right)^{-1}\mathbf{P}^{*\mathrm{T}}_{1:A}\mathbf{z}^*
\end{aligned} \tag{48}$$

REFERENCES

1. Jackson JE. *A User Guide to Principal Components*. Wiley: New York, 1991.
2. Martens H and Naes T. *Multivariate Calibration*. Wiley: New York, 1989.
3. Kourti T and MacGregor JF. Multivariate SPC methods for process and product monitoring. *J. Qual. Technol.* 1996; **28**: 409–428.
4. Little RJA and Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
5. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemometrics Intell. Lab. Syst.* 1987; **2**: 37–52.
6. Grung B and Manne R. Missing values in principal component analysis. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 125–139.
7. Nelson PPC, Taylor PA and MacGregor JF. Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemometrics Intell. Lab. Syst.* 1996; **35**: 45–65.
8. *SIMCA-P 8.0: User Guide and Tutorial*. Umetrics AB: Umeå, 1999.
9. Geladi P and Kowalski BR. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 1986; **85**: 1–17.
10. Wold S, Albano C, Dunn WJ, Esbensen K, Hellberg S, Johansson E and Sjostrom M. Pattern recognition: finding and using regularities in multivariate data. In *Food Research and Data Analysis*, Martens H, Russwurm Jr H (eds). Applied Science Publishers: London and New York, 1983; 183–185.
11. Draper N and Smith H. *Applied Regression Analysis* (2nd edn). Wiley: New York, 1981.
12. Wise BM and Ricker NL. Recent advances in multivariate statistical process control: improving robustness and sensitivity. IFAC Int. Symp., ADCHEM '91, Toulouse, 1991; 125–130.
13. Wold S. Cross validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.