I&EC research
Industrial & Engineering Chemistry Research

Process Systems Engineering

Automated model selection in principal component analysis: A new approach based on the cross-validated ignorance score

Stefania Russo, Guangyu LI, and Kris Villez

Ind. Eng. Chem. Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.iecr.9b00642 • Publication Date (Web): 21 Jun 2019

Downloaded from http://pubs.acs.org on June 24, 2019

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Automated model selection in principal component analysis: A new approach based on the cross-validated ignorance score

Stefania Russo,[*,†] Guangyu Li,[*,†,‡] and Kris Villez[*,†,‡]

[†] *Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland*

[‡] *ETH Zürich, Institute of Environmental Engineering, 8093 Zürich, Switzerland*

E-mail: stefania.russo@eawag.ch; ligu@student.ethz.ch; kris.villez@eawag.ch

Abstract

Principal component analysis (PCA) is by far the most widespread tool for unsupervised learning with high-dimensional data sets. It is popularly studied for exploratory data analysis and online process monitoring. Unfortunately, fine-tuning PCA models and particularly the number of components remains a challenging task. Today, this selection is often based on a combination of guiding principles, experience, and process understanding. Unlike the case of regression, where cross-validation of the prediction error is a widespread and trusted approach for model selection, there are no tools for PCA model selection enjoying this level of acceptance. In this work, we address this challenge and evaluate the utility of the cross-validated ignorance score with both simulated and experimental data sets. Application of this model selection criterion is based on the interpretation of PCA as a density model, as in probabilistic principal component analysis. With simulation-based benchmarking, it is shown to be *(a)* the overall best performing criterion, *(b)* the preferred criterion at high noise levels, and *(c)* very robust to changes in noise level. Tests on experimental data sets suggest that the ignorance score is sensitive to deviations from the PCA model structure, which suggests the criterion is also useful to detect of model-reality mismatch.

Introduction

Principal component analysis (PCA) is one of the most popular models for data mining, machine learning, and process monitoring.[1–11] This is partially explained by the widespread availability of efficient algorithms for data decomposition via PCA. Despite its popularity, it remains difficult to optimize the single hyper-parameter of PCA models, i.e. the number of principal components (PCs) that are retained. An associated challenge is that PCA models can be hard to interpret. In addition, the use of PCA implies that retaining more variance means capturing more information, an assumption that is often hard to inspect carefully. The problem of selecting an optimal model hyper-parameter is shared with many unsupervised learning models.[12–14]

The adequate selection of the number of PCs is crucial for both applications relying on data compression (e.g., exploratory analysis) as well as for predictive tasks (e.g., multivariate classification and regression). As a result, a wide variety of criteria for determining the number of PCs have been proposed.[3] A first group of criteria, like the Akaike information criterion, Kaiser rule, parallel analysis, and the variance of reconstruction error (VRE) are grounded in theory.[15–19] Application of these criteria assumes that *(i)* the obtained measurements depend linearly on an unknown number of hidden variables and that *(ii)* the measurement errors are sampled independently from the same univariate normal distribution and are therefore homoskedastic and uncorrelated. These criteria differ in their definition of optimality.

When the model structure itself cannot be assumed correct but one still wants to use PCA for dimensionality reduction, it is recommended to use cross-validated model selection criteria.[16,20] Whereas the use of the cross-validated mean squared residual in linear regression[21,22] is almost universal, there are no cross-validation methods for PCA that reach this kind of widespread adoption. The sum-of-squares of prediction error (PRESS) computed in a row-wise K-fold cross-validation (RKF) pattern is a popular criterion to evaluate PCA model performance, yet always exhibits a monotonic profile as function of the number of PCs.[23] For this reason, it is inadequate for automated dimensionality selection. Fortunately, this can be addressed by evaluating the ability of a PCA model to compute another variation of the PRESS based on missing data imputation. To this end, one repeatedly removes measurements from the data available for model identification and imputes these with the PCA model[24,25]. This leads to model selection based on the element-wise K-fold cross-validation (EKF) of the squared imputation error.[26] At this time, the best-known imputation procedures associated with PCA include *iterative estimation* (ITR), *projection to the model plane* (PMP), *trimmed score imputation* (TRI), and *trimmed score regression* (TSR). More recently, improved versions of these imputation procedures have been obtained by means of data augmentation.[27] Their imputation performances have been compared for both simulation data and experimental data.[27,28] They have not been benchmarked for the purpose of PCA model selection. Optimal dimensionality selection for data compression with PCA models forms the focus of our work. Note also that we are particularly focused on finding methods that enable automation of this modelling step.

An important inconvenience of EKF is that the computational cost does not scale well with the number of variables in a data set. For this reason, we propose two new criteria for PCA model selection and compare it to the imputation-based criteria discussed above and the VRE criterion. The new criteria are based on the application of the ignorance score (IGN) as a criterion for model selection[29] and require that PCA is interpreted as a density model, as in probabilistic principal component analysis (PPCA).[30] Importantly, this approach can be used with EKF as well as with RKF, in turn leading to a computationally efficient model selection criterion. To evaluate the performance of both existing and the newly proposed criteria, we deploy an extensive simulation study of a scale that has not been used before.

By repeated simulation of benchmarking data sets, we obtain a distribution for the selected dimensionality, rather than a single selection, in turn making our conclusions stronger.

In what follows, we demonstrate that the overall accuracy of the proposed IGN-EKF and IGN-RKF criteria is higher than the overall accuracy for a wide range of preexisting model selection criteria. This is so despite the observed computational savings. We also report on tests of all considered criteria with two experimental data sets containing light absorbance spectra. Our tests point out that not a single criterion identifies the expected number of components. The proposed IGN-EKF and IGN-RKF criteria may however be suited well to detect this kind of model-reality mismatch. In the following sections, we describe the simulated and experimental data sets used in this work. We then proceed with a classical structure including results, discussion, and conclusions.

Materials and Methods

We first describe the simulated and experimental data sets used for this study. After, the applied methods for data analysis are described in detail. All mathematical symbols are listed in Table 2. Table 1 lists all acronyms.

Table 1: List of acronyms

Acronym	**Full wording**
EKF	Element-wise K-fold cross-validation
IGN	Ignorance score
ITR	Iterative estimation
PCA	Principal component analysis
PC (PCs)	Principal component(s)
PPCA	Probabilistic principal component analysis
PMP	Projection to the model plane
PRESS	Sum-of-squares of prediction error
RKF	Row-wise K-fold cross-validation
RMSR	Root mean squared residual
RKF	Row-wise K-fold cross-validation-wise K-fold cross-validation
TRI	Trimmed score imputation
TSR	Trimmed score regression
VRE	Variance of reconstruction error

Data sets

Simulated data sets

Simulated data sets are constructed with known numbers of samples (I), variables (J), and PCs (K). We define the known number of PCs as the number of eigenvalues of the simulated covariance matrix that are different from the smallest eigenvalue. Each data set type is indexed as $c.s$, where c indicates the data set class and s indicates the data set type in that class. We name the simulated data classes (c) B1, B2, C1, C2, C3, C4 in line with previous studies.[31] The B1 and B2 classes both contain 18 types indicated with s (e.g., $c.s =$ B1.1,. . ., B1.18). The B1 (B2) class simulates mean-centered data with 9 (18) variables and the known number of PCs varies between 0 and 7 (13).[32] The C1, C2, C3, and C4 classes simulate 10, 10, 27, and 50 variables with 4, 8, 12, and 15 PCs.[27] Each class is simulated with six different noise levels, which are indicated with s (e.g. $c.s =$ C1.1,. . ., C4.6). The applied noise levels are 5%, 10%, 20%, 30%, 40%, and 50%. Each simulation of a given data type is repeated 100 times to enable an accurate evaluation of the expected accuracy of the model selection criteria. Each repetition is indicated with r (e.g. $c.s.r =$ B1.1.1,. . ., C4.5.100) The use of repetitions is an important distinction between our benchmark simulations and earlier studies with this type of simulation data sets. The detailed simulation procedures are defined in the *Supporting Information* (Section B).

Experimental data sets

To test the quality of the studied model selection criteria under real-world conditions, we test their performance also with two experimental data sets. The first one was collected specifically for this study while the second one was obtained from prior work.[33]

Nitrogen species data set The first experimental data set was collected in a way that a PCA model with two PCs is expected to describe the obtained data well. To this end, stock solutions of nitrite ($NO_2{}^-$) and nitrate ($NO_3{}^-$) are prepared first. Each of these two stock media were prepared as a single batch with target concentrations of 5 g atomic nitrogen per liter (5 $gNO_2{}^{-}$-N/L, 5 $gNO_3{}^{-}$-N/L).

With the prepared stock solutions, diluted media were obtained adding 600 mL of nanofiltered water to a glass cylinder first. Then, well-measured amounts of the two stock solutions are added in steps of 2 mL with a minimum of 1 and a maximum of 16 steps, leading to a square two-dimensional grid of added volumes of the two stock solutions as shown in Figure 1. As a result, the expected concentrations of both species range from 15.8 mg N/L (1 step of one stock solution and 16 steps for the other) to 252.4 mg N/L (e.g., 1 step with $NO_2{}^{-}$ and 16 steps with $NO_3{}^{-}$ stock solution). The order in which the samples were prepared was randomized partially to avoid temporal correlation within the collected data sets. More details regarding the experimental procedure used to prepare the solutions are found in the *Supporting Information* (Section C).

Immediately after preparation of each diluted solution, five replicate absorbance spectra were collected by submerging an on-line ultraviolet-visible light absorbance spectrophotometer (spectro::lyser^{TM}, S::CAN Messtechnik, Vienna, Austria) into the diluted solution, thus producing 1280 spectra in total. The applied spectrophotometer has a light path length of 2 mm and produces measurements which are composed of 215 absorbance values taken at wavelengths between 200 nm and 735 nm with steps of 2.5 nm. PCA models are studied for a variety of variable selections for reasons explained below. We refer to the resulting modified data sets as nitrogen species data set 0 (no variable selection, 215 wavelengths), nitrogen species data set 1 (wavelengths 285-735 nm, 181 wavelengths), and nitrogen species data set 2 (wavelengths 285-385 nm, 41 wavelengths).

The design of the experiment includes two factors, the nitrite and nitrate concentration. According to the Beer-Lambert law, the absorbance measurements depend linearly on these concentrations. Thus, a model with two components should deliver a good representation of the collected data. In addition, the number of factors in the experimental design can be used as a gold standard for evaluation of automatic model selection criteria.

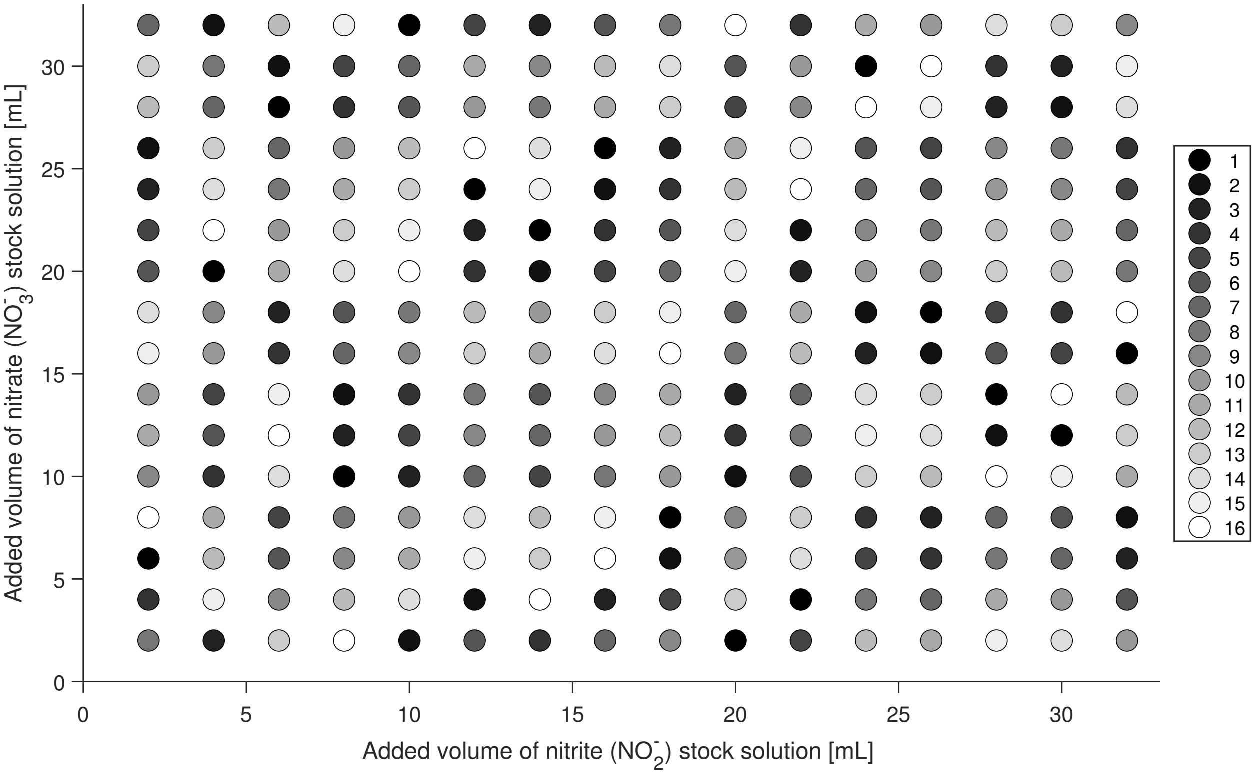


Figure 1: **Nitrogen species data – Experimental design and block assignment.** Five spectra are collected for each combination of added volume of the nitrite and nitrate stock solutions. The obtained spectra are divided in 16 blocks by means of a randomized Latin square for the purpose of cross-validation. Each block is indicated with a unique shade.

Metal ion data set The second experimental data set consists of UV-Vis spectra for 26 mixtures of three metal ions (Co(II), Cr(III), and Ni(II)) in water containing 4% nitric acid (HNO_3). These mixtures are obtained with a 3×3×3 full factorial design, while one sample was omitted during experimental data collection. For each of the mixtures, five spectra are recorded. The original spectra consist of light absorbance measurements at 176 wavelengths from 300 nm to 650 nm in steps of 2 nm. For more details on the data collection procedure we refer to the original publication.[33] As in this prior work, two of the spectra are considered outliers and removed prior to analysis, thus leading to a 128×176-dimensional data matrix. This data set is considered to have strongly correlated measurement errors.[33] In this work, these data are used primarily to demonstrate that the results obtained with the nitrogen species data set are likely for experimental absorbance data.

Methods

We now explain how the dimensionality for the considered PCA models is identified with 11 model selection criteria. In what follows, we assume that an $I \times J$-dimensional matrix $\mathbf{Y}$ without missing entries is available for model selection.

Cross-validation patterns

In this work, we make use of two cross-validation procedures for model selection, named RKF and EKF. These are explained next.

Row-wise K-fold cross-validation (RKF). In the row-wise K-fold cross-validation we split the data into B blocks along the row dimension ($b = 1, \ldots, B$). Each block is a $I^{(b)} \times J$-dimensional matrix $\mathbf{Y}^{(b)}$ with the row dimensions ($I^{(b)}$) distributed as evenly as possible. The vector $\mathbf{i}^{(b)}$ contains the indices of the rows in $\mathbf{Y}$ matching the rows of $\mathbf{Y}^{(b)}$. This is visualized in the left panel of Figure 2. Then, cross-validation with any model selection criterion computes an I-dimensional vector $\boldsymbol{q}$ of model performance measures as follows:

(a) Set $v \leftarrow 1$

(b) Select block v as the validation data matrix $\mathbf{Y}^{(v)}$ and compose the $I^{(c)} \times J$-dimensional calibration data matrix $\mathbf{Y}^{(c)}$ by combining the remaining blocks ($\{b = 1, \ldots, B | b \neq v\}$, $I^{(c)} = I - I^{(v)}$)

(c) Calibrate the model with $\mathbf{Y}^{(c)}$

(d) Evaluate the model performance for each row in $\mathbf{Y}^{(v)}$, leading to an $I^{(v)}$-dimensional vector of performance measures, named $\boldsymbol{q}^{(v)}$

(e) Place all elements of $\boldsymbol{q}^{(v)}$ into the positions $\mathbf{i}^{(v)}$ of $\boldsymbol{q}$

(f) If $v < B$, set $v \leftarrow v + 1$ and go back to step (b). Otherwise, terminate.

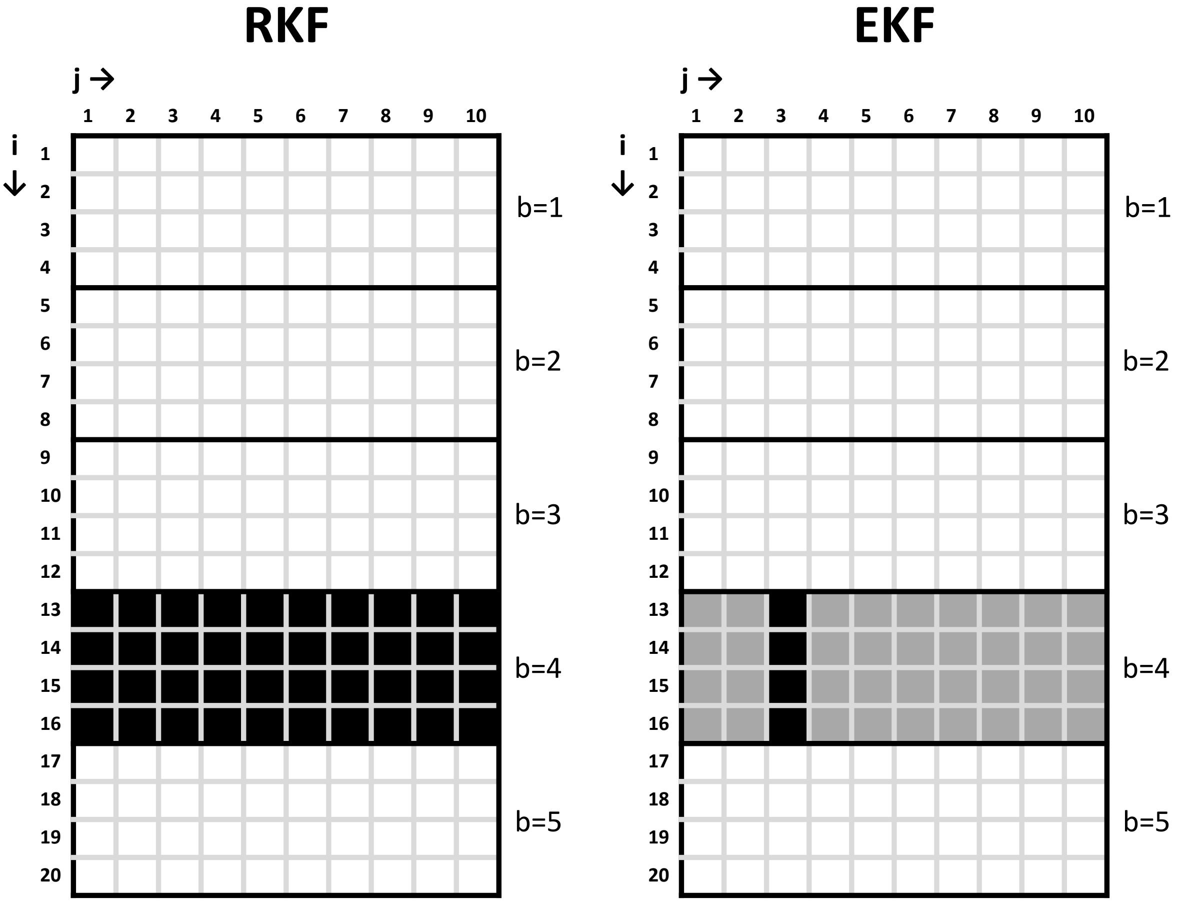


Figure 2: **Cross-validation patterns.** Example with $I = 20$ samples and $J = 10$ variables. Left: row-wise K-fold cross-validation (RKF) splits the data into $B = 5$ blocks in the row-wise direction. In the 4th iteration of RKF, block $b = 4$ is used for validation (black) while the other blocks are used for calibration (white). Right: element-wise K-fold cross-validation (EKF) splits the data into $B = 5$ blocks in the row-wise direction first. Then, in every iteration one column in one block (e.g., $b = 4$ & $j = 3$, black) is treated as missing data and imputed on the basis of the data in the other columns (grey) and a model estimated with the calibration data in the other blocks (white).

At the end of this procedure, all elements of $\boldsymbol{q}$ have been evaluated, meaning that every row in $\mathbf{Y}$ has been used once in the validation step (d). The way the elements of $\boldsymbol{q}$ are processed further depends on the model being used and is explained below.

In this study, the blocks are defined slight differently for every data set:

- *Simulation data sets:* The $I = 1024$ rows are randomly assigned to $B = 16$ blocks, each containing $I^b = 64$ rows of the matrix $\mathbf{Y}$.

- *Nitrogen species data sets:* The $I = 1280$ rows are assigned to $B = 16$ blocks, each containing $I^b = 80$ rows of the matrix $\mathbf{Y}$. The assignment is executed in such a way that *(a)* all replicates corresponding to a single set of added stock volumes are assigned to the same block and that *(b)* each block b contains data corresponding to each of the 16 added stock volumes for both stock solutions. Each of these blocks is identified with a unique shade in Figure 1.

- *Metal ion sets:* The $I = 128$ spectra are assigned into $B = 9$ blocks, with $I^{(1)} = 9$, $I^{(2)} = 14$, and $I^{(b)} = 15$ ($b \geq 3$). This assignment is executed so that *(a)* all replicates corresponding to a single set of concentration levels are assigned to the same block and that *(b)* every block contains data corresponding to every level for every metal ion concentration.

Element-wise K-fold cross-validation (EKF). In the EKF, the matrix $\mathbf{Y}$ is first divided into B blocks as with RKF discussed above. Cross-validation then computes an $I \times J$-dimensional matrix $\mathbf{Q}$ of model performance measures by the following procedure, which is visualized in the bottom panel of Figure 2. :

(a) Set $v \leftarrow 1$

(b) Select block v as the validation data matrix $\mathbf{Y}^{(v)}$ and compose the $I^{(c)} \times J$-dimensional calibration data matrix $\mathbf{Y}^{(c)}$ by combining the remaining blocks ($\{b = 1, \ldots, B | b \neq v\}$, $I^{(c)} = I - I^{(v)}$)

(c) Calibrate the model with $\mathbf{Y}^{(c)}$

(d) For every column j ($j = 1, \ldots, J$):

    (i) Select the $I^{(v)}$-dimensional vector $\mathbf{Y}^{(v)}_{\bullet,j}$ as the jth column of $\mathbf{Y}^{(v)}$ and treat is the missing data. Call the $I^{(v)} \times (J-1)$-dimensional matrix composed of the remaining $J - 1$ columns of $\mathbf{Y}^{(v)}$ the imputation data and note this matrix as $\mathbf{Y}^{(v)}_{\bullet,-j}$ ($-j = \{l = 1, \ldots, J | j \neq l\}$).

(ii) Compute estimates for the imputed data $\hat{\mathbf{Y}}^{(v)}_{\bullet,j}$ on the basis of the calibrated model obtained in step (c) and the imputation data defined in step (d.ii)

(iii) Compute the $I^{(v)}$-dimensional vector $\boldsymbol{q}^{(v,j)}$ of model imputation performance measures

(iv) Place all elements of $\boldsymbol{q}^{(v,j)}$ into the row positions $\mathbf{i}^{(v)}$ and jth column of the $I \times J$-dimensional matrix $\mathbf{Q}$

(e) If $v < B$, set $v \leftarrow v + 1$ and go back to step (b). Otherwise, terminate.

Common steps to all considered models

Standardization. We apply mean centering in all studied cases and do not apply any scaling. This means that we study covariance PCA exclusively. Additionally, this corresponds to the implicit assumption that all measurement errors are drawn from an isotropic multivariate normal distribution, i.e. that they are uncorrelated and homoskedastic. When this assumption is correct, it follows that the resulting PCA models are optimal in both the total least-squares and the maximum-likelihood sense.[33–36] In what follows, we refer to this as the spherical noise assumption.

The J-dimensional column-wise mean vector, $\mathbf{m}$, is always computed with the calibration data in step (c) of the cross-validation procedure and then applied to the validation data during step (d).

Singular value decomposition (SVD). The calibration data matrix, $\mathbf{Y}^{(c)}$, is decomposed with SVD so that:

$$\mathbf{Y}^{(c)} = \mathbf{1}\,\mathbf{m}^T + \mathbf{U}^{(c)}\,\mathbf{S}\,\mathbf{V}^T := \mathbf{1}\,\mathbf{m}^T + \mathbf{T}^{(c)}\,\mathbf{V}^T \tag{1}$$

with $\mathbf{T}^{(c)}$ the $I^{(c)} \times \overline{K}$-dimensional matrix of principal scores; $\mathbf{U}^{(c)}$ the $I^{(c)} \times \overline{K}$-dimensional matrix of standardized principal scores; $\mathbf{S}$ the $\overline{K} \times \overline{K}$-dimensional diagonal matrix containing all non-zero singular values, ordered from largest to smallest; and $\mathbf{V}$ the $J \times \overline{K}$-dimensional matrix containing the loading vectors as columns. $\overline{K}$ is the maximum number of PCs and equals $\overline{K} := \min(I,\ J)$.

Model 1: Principal component analysis

We now discuss *(i)* steps (c), (d.ii), and (d.iii) of the EKF procedure as applied to the PCA models, *(ii)* PCA model selection with the EKF procedure, and *(iii)* a second PCA model selection procedure based on the variance of reconstruction error (VRE) as proposed in.[16]

PCA calibration - Step (c) of the EKF procedure. After standardization and SVD, principal component analysis (PCA) proceeds by selecting a number K ($K \leq \overline{K}$) and choosing the first K columns in $\mathbf{U}^{(c)}$ and $\mathbf{V}$ and selecting the first K rows and columns of $\mathbf{S}$. This leads to the following least-squares optimal approximation of the calibration data:[37]

$$\mathbf{Y}^{(c)} \approx \hat{\mathbf{Y}}^{(c)} := \mathbf{1}\,\mathbf{m}^T + \mathbf{U}^{(c)}_{\bullet,1:K} \cdot \mathbf{S}_{1:K,1:K} \cdot \mathbf{V}_{\bullet,1:K}{}^T = \mathbf{1}\,\mathbf{m}^T + \mathbf{T}^{(c)}_{\bullet,1:K} \cdot (\mathbf{V}_{\bullet,1:K})^T\,. \qquad (2)$$

PCA validation - Step (d.ii) and (d.iii) of the EKF procedure. Every PCA model is validated by means of 8 distinct imputation-based model selection criteria. The imputation procedures are named ITR, PMP, TRI, TSR, cITR, cPMP, cTRI, and cTSR and have been described in detail in prior work.[24,25,27] In view of readability of our study, we provide the details of step (d.ii) for each of these imputation procedures in the *Supporting Information* (Section D).

After computation of the imputed estimates $\hat{\mathbf{Y}}^{(v)}_{\bullet,j}$ for a given validation block (v) and missing data column (j) (step (d.ii)), one evaluates the model performance as the imputation error (step (d.iii)), i.e. the deviations between the imputed values and the original data:

$$\boldsymbol{q}^{(i,j)} := \hat{\mathbf{Y}}^{(v)}_{\bullet,j} - \mathbf{Y}^{(v)}_{\bullet,j} \tag{3}$$

PCA model selection with EKF. After completing the complete EKF procedure with steps (c), (d.ii), and (d.iii) as detailed above, one computes the cross-validated PCA model performance as the root mean squared residual (RMSR), which equals the square root of the prediction error sum of squares (PRESS) divided by the number of estimated elements:

$$\text{RMSR} = \sqrt{\frac{PRESS}{I\,J}} := \sqrt{\frac{1}{I\,J}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\mathbf{Q}_{i,j}\right)^2} \tag{4}$$

The RMSR is evaluated with the EKF procedure for every feasible number of PCs (K). The number of PCs leading to the smallest RMSR is then selected. Note that above averaging over all samples and variables also implies that the reconstruction error in all variable is weighed equally. Combined with the fact that the reconstruction errors are squared, this means that the model selection criterion implies that the reconstruction errors are ideally homoskedastic and uncorrelated.

Model selection with VRE.

In this work, the VRE criterion is computed by considering faults in individual sensors and assuming the use of covariance PCA as above. This means we consider J unique fault directions defined as J-dimensional vectors ξ_j, each defined as unit vectors with $\xi_j(j) = 1$ and $\xi_j(l) = 0 (j \neq l)$. Then, one projects these fault directions in the residual space conditional to a PCA model with K component as follows:

$$\tilde{\xi}_j = \left(\mathbf{I}_J - \mathbf{V}_{\bullet,1:K} \cdot \left(\mathbf{V}_{\bullet,1:K}\right)^T\right) \cdot \xi_j \tag{5}$$

Then, the overall VRE criterion for the K-PC model is then computed as:

$$\mathrm{VRE} = \frac{1}{J}\sum_{j}^{J}\left(\frac{\tilde{\xi}_j{}^{T}\cdot\tilde{\Sigma}\cdot\tilde{\xi}_j}{\left(\tilde{\xi}_j{}^{T}\cdot\tilde{\xi}_j\right)^{2}\left(\xi_j{}^{T}\cdot\tilde{\Sigma}\cdot\xi_j\right)}\right) \tag{6}$$

with $\tilde{\Sigma}$ the empirical covariance matrix obtained with the complete data set:

$$\tilde{\Sigma} := \frac{1}{I}\left(\mathbf{Y}-\mathbf{1}\,\mathbf{m}^{T}\right)^{T}\left(\mathbf{Y}-\mathbf{1}\,\mathbf{m}^{T}\right) \tag{7}$$

where $\mathbf{m}$ is composed of the column means for the whole data set.

This criterion averages the expected variance of the estimated fault magnitude across all considered faults. The PCA model with minimum VRE is then selected as the model that has the best overall ability - in the minimum-variance sense - to estimate the magnitude of any of the considered (sensor) faults.

Model 2: Probabilistic principal component analysis

The ignorance score is a model performance criterion designed to select models that can accurately predict densities. PCA can be interpreted as a density model in the form of PPCA. The PPCA model is discussed first. Step (c) and (d) of PPCA-based model selection are described after.

PPCA was introduced with the goal of formulating a probabilistic version of PCA.[30] The PPCA model is identified as the maximum likelihood estimate of the following generative latent variable model:

$$\boldsymbol{x}_k \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\psi}_k\right) \tag{8}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_J\right) \tag{9}$$

$$\boldsymbol{y} := \mathbf{W}\,\boldsymbol{x} + \boldsymbol{\epsilon} \tag{10}$$

with $\boldsymbol{x}$ the K-dimensional vector of latent variables, $\boldsymbol{x}_k$, each one of them drawn from a univariate normal distribution ($k = 1 \ldots K$), $\boldsymbol{\epsilon}$ a J-dimensional vector of measurement errors drawn from a spherical multivariate normal density, $\mathbf{I}_J$ an identity matrix of appropriate dimensions, $\mathbf{W}$ a $J \times K$-dimensional matrix with full column rank, and $\boldsymbol{y}$ the J-dimensional vector of recorded measurements. During model calibration one has access to I vectors $\boldsymbol{y}$ which are organized as row vectors in a $I^{(c)} \times J$-dimensional data matrix. Note that each of the data simulations discussed above can be represented in this generative form.

PPCA calibration - Step (c) of the RKF and EKF procedures. As in PCA, model identification proceeds by choosing a number K for the number of PCs. Then, starting with the SVD result, the density model for the measured data is formulated as the following multivariate normal distribution:[30]

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(K)}\right) \tag{11}$$

where $\boldsymbol{\Sigma}^{(K)}$ is the maximum likelihood estimate of the covariance matrix given K components. To compute this matrix, one first obtains $\boldsymbol{\lambda}$, the vector of the first K eigenvalues of the empirical covariance matrix, as:

$$\boldsymbol{\lambda}_k := \frac{\left(\mathbf{S}_{k,k}\right)^2}{I^{(c)}} \qquad\qquad k = 1 \ldots K \tag{12}$$

$\mathbf{\Sigma}^{(K)}$ is then available via:

$$\mathbf{\Sigma}^{(K)} := \mathbf{V}_{\bullet,1:K}\,\mathbf{\Psi}\,(\mathbf{V}_{\bullet,1:K})^T + \sigma_\epsilon\,\mathbf{I}_J \tag{13}$$

$$\sigma_\epsilon := \frac{1}{J}\sum_{k=K+1}^{\overline{K}} (\mathbf{S}_{k,k})^2 \tag{14}$$

$$\boldsymbol{\psi}_k := \boldsymbol{\lambda}_k - \sigma_\epsilon,\; k = 1\,\ldots\,K \tag{15}$$

$$\mathbf{\Psi} := \mathbf{diag}(\boldsymbol{\psi}) \tag{16}$$

where $\mathbf{\Psi}$ is a diagonal matrix with $\boldsymbol{\psi}$, the vector of K deflated variances for the K selected components, on its diagonal, and σ_ϵ the estimated noise variance. $\mathbf{V}_{\bullet,1:K}$ functions as the estimate of $\mathbf{W}$.

One can compute the scaled scores for the kth principal component, $\mathbf{X}_{\bullet,k}$, with the variance equal to $\boldsymbol{\psi}_k$ as:

$$\mathbf{X}_{\bullet,k} := \mathbf{U}_{\bullet,k}\,\boldsymbol{\psi}_k = \mathbf{T}_{\bullet,k}\,\frac{\boldsymbol{\psi}_k}{\boldsymbol{\lambda}_k} \tag{17}$$

In what follows below, we aim to evaluate whether the assumption of spherical measurement noise affects the results with the experimental data sets. To this end, the row vectors in $\mathbf{X}$ are combined with (9) and (10) to generate new samples with spherical noise without changing the distribution of the first K latent variables.

A key observation is that PPCA explicitly accounts for the fact that the principal scores computed with PCA are subject to noise. Under the assumptions of spherical measurement noise and given a choice for K, the deflated variances reflect the magnitude of variation in the PCs that is not interpreted as noise. This also means that the fraction of the variance associated with a particular principal component that is interpreted as meaningful, i.e. non-noisy, *(a)* will be lower than the fraction of explained variance obtained with PCA ($\boldsymbol{\psi}_k < \boldsymbol{\lambda}_k$,

$k = 1, \dots K,\ K < \overline{K} - 1$) and *(b)* will increase with increasing values of K ($K_1 < K_2 : \boldsymbol{\psi}_k|_{K=K_1} \leq \boldsymbol{\psi}_k|_{K=K_2}, k = 1, \dots K_2$). Note also that $K = J - 1$ and $K = J$ deliver the same estimate for $\boldsymbol{\Sigma}^{(K)}$, the only distinction being the interpretation of the Jth loading vector as a noisy direction ($K = J - 1$) or an informative direction ($K = J$).

PPCA validation - Step (d.ii) and (d.iii) of the EKF. To apply EKF for cross-validation of the PPCA model, we make use of the ignorance score.[29] This criterion has been used in hydrological modeling for ensemble model calibration[38] and has been proposed in data mining to identify the number of clusters in density-based cluster models.[12] The ignorance score is defined as the negative logarithm of the likelihood of a scalar measurement, y, given a likelihood function $L(\bullet)$:

$$\mathcal{I}(y) = -\mathbf{ln}\left(\mathcal{L}(y, \boldsymbol{\theta})\right) \tag{18}$$

with $\boldsymbol{\theta}$ the vector of parameters. The ignorance score has the advantage of being a negatively oriented score,[29] so that a minimum value indicates the model which maximizes the predicted likelihood of a measurement.

The PPCA model is used to compute both the mean and the variance of the jth variable conditional to the other measurements. This is unlike the EKF for PCA model selection, which only computes a mean estimate. The computed mean and variance describe the conditional univariate normal density for each of the missing measurements. Practically, the imputed estimates $\hat{\mathbf{Y}^{(v)}}_{\bullet,j}$ are equal to the conditional means:

$$\hat{\mathbf{Y}}^{(v)}_{\bullet,j} := \mathbf{Y}^{(v)}_{\bullet,-j} \cdot \left(\boldsymbol{\Sigma}^{(K)}_{-j,-j}\right)^{-1} \cdot (\boldsymbol{\Sigma}^{(K)}_{-j,-j})\,. \tag{19}$$

The conditional variances is unique for each validation block v and each variable j and is equal to:

$$\phi^{(v,j)} := \mathbf{\Sigma}^{(K)}_{-j,-j} - (\mathbf{\Sigma}^{(K)}_{j,-j}) \cdot \left(\mathbf{\Sigma}^{(K)}_{-j,-j}\right)^{-1} \cdot (\mathbf{\Sigma}^{(K)}_{-j,-j}) \tag{20}$$

This completes step (d.ii). To execute step (d.iii), the univariate normal density is applied to evaluate the ignorance score for each row and column of the validation data matrix and record the model performance measure:

$$\begin{aligned} \boldsymbol{q}_i^{(v,j)} :=& \mathcal{I}(\mathbf{Y}^{(v)}_{i,j}) = -\mathbf{ln}\left(\mathcal{L}\left(\hat{\mathbf{Y}}_{i,j}, \mathbf{Y}^{(v)}_{i,j}, \phi^{(v,j)}\right)\right) \\ =& \frac{1}{2}\left(\mathbf{ln}\,(2\pi) + \mathbf{ln}\left(\left(\phi^{(v,j)}\right)\right) + \frac{\left(\hat{\mathbf{Y}}^{(v)}_{i,j} - \mathbf{Y}^{(v)}_{i,j}\right)^2}{\phi^{(v,j)}}\right), \; i = 1, \ldots, I^{(v)} \end{aligned} \tag{21}$$

with i and j indicating the row and column positions of the elements of the validation data in $\mathbf{Y}^{(v)}$.

PPCA model selection with EKF. At the end of the EKF procedure, the overall performance of the PPCA model to predict the data density is evaluated as the averaged ignorance score (IGN):

$$\text{IGN} = \frac{1}{I\,J}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\mathbf{Q}_{i,j}\right)^2 \tag{22}$$

The selected model is the one producing the minimal value for IGN. We refer to this model selection procedure as IGN-EKF.

PPCA validation - Step (d) of the RKF procedure. The ignorance score can easily be extended for application in the RKF procedure. In this case, the ignorance score simply follows from evaluating the density associated with the PPCA model for each row in the validation matrix.[29] The ignorance score for every row in the validation data matrix $\mathbf{Y}^{(v)}$ is computed as:

$$\boldsymbol{y}_i^{(v)} = \left(\mathbf{Y}_{i,\bullet}^{(v)}\right)^T \tag{23}$$

$$\begin{aligned}\boldsymbol{q}_i^{(v)} :=& \mathcal{I}\left(\boldsymbol{y}_i^{(v)}\right) = -\mathbf{ln}\left(\mathcal{L}\left(\boldsymbol{y}_i^{(v)}, \mathbf{0}, \boldsymbol{\Sigma}^{(K)}\right)\right) \\ =& \frac{1}{2\,J}\left(J\,\mathbf{ln}\,(2\pi) + \mathbf{ln}\left(\left|\Sigma^{(K)}\right|\right) + \left(\boldsymbol{y}_i^{(v)}\right)^T \left(\boldsymbol{\Sigma}^{(K)}\right)^{-1} \boldsymbol{y}_i^{(v)}\right),\ i = 1, \ldots, I^{(v)}\end{aligned} \tag{24}$$

with i indicating the row position in $\mathbf{Y}^{(v)}$. Note that we divide by J on the right hand side to produce values in the same scale as the IGN-EKF.

Note that the ignorance score in the RKF is equal to the Mahalanobis distance using the estimated covariance matrix $\Sigma^{(K)}$ plus a constant, which depends on the chosen K. The ignorance score is therefore a distance defined in a J-dimensional data space. This distance assumes equalization of the variances for the residual space as described above. This is unlike other PCA-based distances like the squared prediction error, which is defined for the $J - K$-dimensional residual space (without any equalization), and Hotelling's T^2 statistic, which is defined for K-dimensional principal component space. This is discussed further in the Discussion section.

PPCA model selection with RKF. At the end of the RKF procedure, the overall performance of the PPCA model to predict the data density is evaluated as the averaged ignorance score (IGN):

$$\text{IGN} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\boldsymbol{q}_i\right)^2 \tag{25}$$

Note that we divide by $I \cdot J$ to produce an ignorance score in the same scale as the EKF-IGN. The selected model is the one producing the minimal value for IGN. We refer to this model selection procedure as IGN-RKF.

Benchmarking of the model selection criteria

In the above, we have discussed 8 pre-existing criteria for PCA dimensionality selection based on the application of EKF and missing data imputation. We name these model selection criteria according to the applied imputation procedure: ITR, PMP, TRI, TSR, cITR, cPMP, cTRI, and cTSR. We also described two dimensionality selection criteria based on cross-validation of the PPCA model: EKF-IGN and RKF-IGN and include one model selection criterion based on the expected VRE.

The efficiency and accuracy of these selection criteria for model selection are evaluated with the simulated data sets described above. To quantify the performance of the model selection procedures, the following criteria are computed:

(a) The fraction of the number of instances of a data set for which the identified number of PCs matches the ground truth value exactly. This is reported as a percentage.

(b) The average run time, measured in seconds, for a single execution of the studied cross-validation procedure. All computations are executed on a single machine (Intel(R) Core(TM) i5-7200U CPU: 2.50GHz, RAM: 8.0 GB; Microsoft Windows 10 Enterprise; Matlab R2017b).

In addition, we will inspect the histograms of the identified number of PCs as a function of the data set type and noise level. Evaluating these histograms is possible thanks to repeated simulations. This kind of intensive simulation benchmarking sets a new standard in the study of latent variable modelling.

Results

The results with simulated data are discussed first. After that, results obtained with the experimental data sets are described.

Simulation data sets

Exemplary cross-validation results

Figure 3 shows the cross-validated criteria obtained with data set C.4.6.11 (class C, type 4, noise level 6, repetition 11). The results are shown in three distinct panels with each panel grouping a subset of the applied criteria. This figure arrangement will be repeated below. All criteria exhibit a unimodal profile (i.e., with a single minimum) which makes the automatic selection of a number of PCs (K) straightforward. The minimum is however observed at different locations. The profiles for the ITR, PMP, and VRE profiles are qualitatively similar with a rather narrow valley and have a minimum at 10 PCs. The TRI and TSR criteria select 12 PCs and exhibit a smoother profile. The cITR, cPMP, cTRI, and cTSR all exhibit monotonically decreasing profiles so that the maximum number of PCs is selected. The IGN-EKF and IGN-RKF profiles are generally smooth. In this case, IGN-EKF and IGN-RKF are the only criteria that select the known number of PCs (15) perfectly.

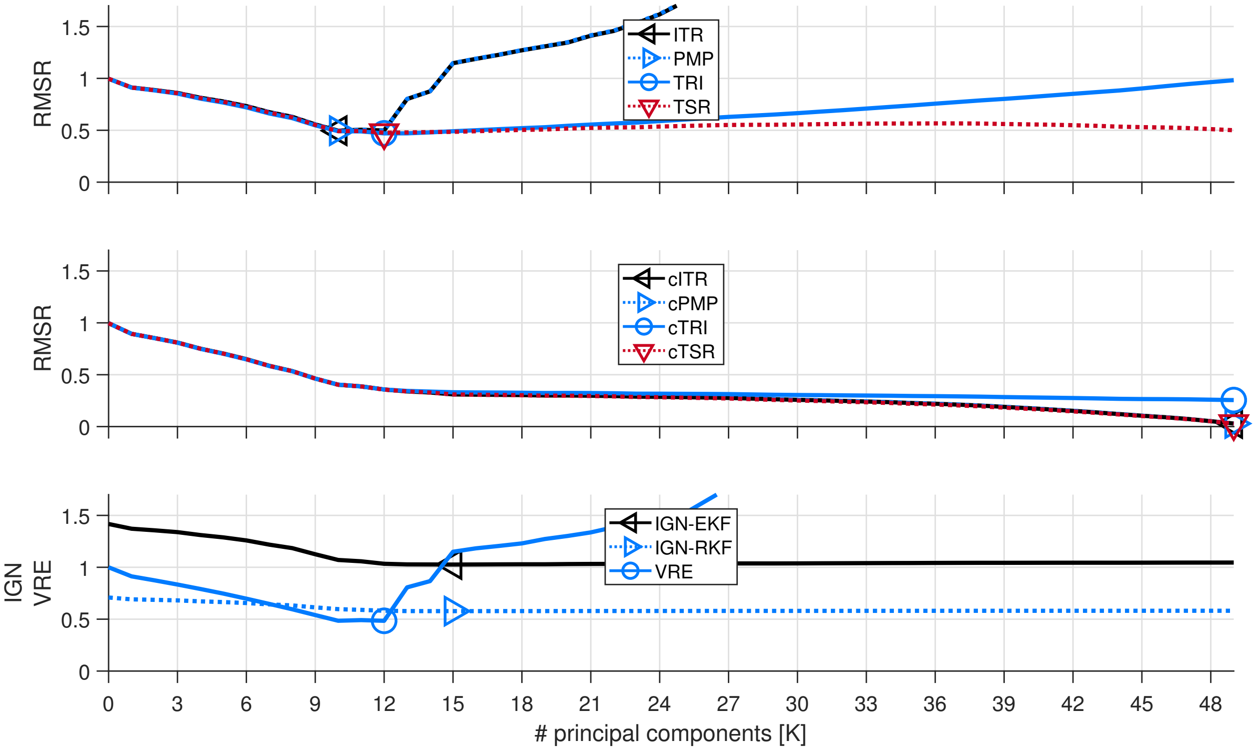


Figure 3: **Simulated data set C.4.6.11 – Model selection criteria profiles.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE. The model selection criteria are shown as a function of the number of retained PCs. The selected number of PCs is indicated with a triangle or circle.

Benchmarking with simulated data sets

We first discuss the accuracy for each model selection criteria averaged over all data sets in the simulated data classes (B1, B2, C1, C2, C3, C4). Inspections that are more detailed follow after.

Average performance. Figure 4 displays the obtained accuracy of the selected dimensionality as a function of the applied criterion and for every data class. Most importantly, one can see that the IGN-EKF and IGN-RKF criteria are the only ones leading to an average accuracy of 85% higher for each data class (B1, B2, C1, C2, C3, C4). In contrast, the cITR, cPMP, and cTSR criteria deliver 0% accuracy in all data classes. The performance of the remaining criteria ranges from 0% to 100% and is sensitive to the chosen data class. For example, the ITR, PMP, and VRE criteria deliver an accuracy around 42% and 55% for the data classes B1-B2, and 0% for the data classes C1-C4.

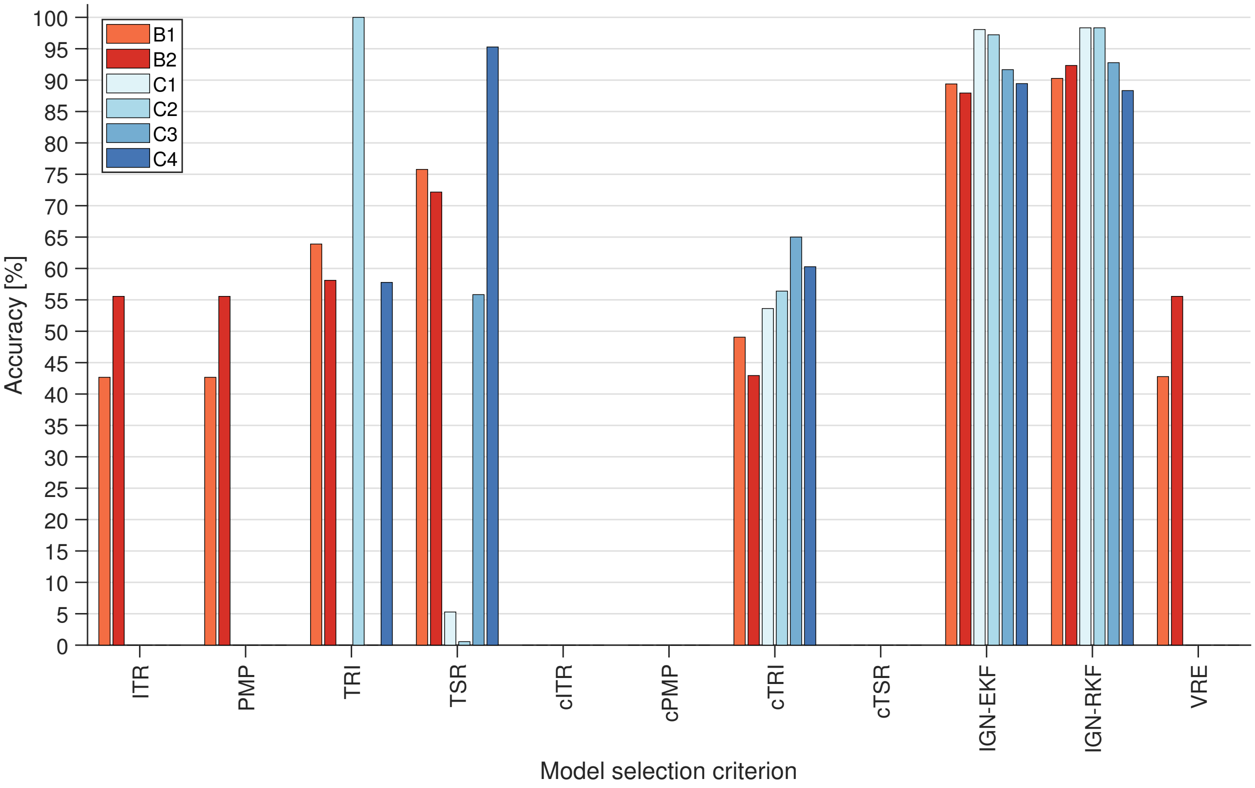


Figure 4: **Simulation data – Overall accuracy.** The fraction of correctly identified number of PCs is shown for each data class (B1-B2, C1-C4) and for each model selection criterion. The averaged accuracy with IGN-EKF and IGN-RKF is over 85% for all data classes. These are also the best performances for classes B1, B2, C2, and C3. TRI is the best criterion for class C1 while TSR is the best for class C4. Among the imputation-based criteria, TSR is best for classes B1, B2, C3, and C4 while TRI is best for class C1 and cTRI is best for class C2.

Detailed results for data class B1. Figure 5 displays the obtained accuracy of the selected dimensionality with all criteria as a function of the data type (B1.1 to B1.18). There are three important observations that could not be concluded from the average accuracy as discussed above. First, the performance of the imputation-based criteria ITR, PMP, TRI, TSR, and cTRI is highly variable, ranging from 0% to 100% accuracy. Among these criteria, a 0% accuracy is reported for at least 4 data types (criterion: TSR) and a 80% accuracy or better is reported for at least 7 data types (criteria: ITR, PMP) and at most 13 data types (criterion: TSR). Second, VRE delivers a performance similar to those obtained with

ITR and PMP. Third, and most importantly, the accuracy obtained with the IGN-EKF and IGN-RKF never drops below 85% for a single data type. Note also that IGN-EKF and IGN-RKF deliver the best accuracy for data types B1.13 and B1.14.

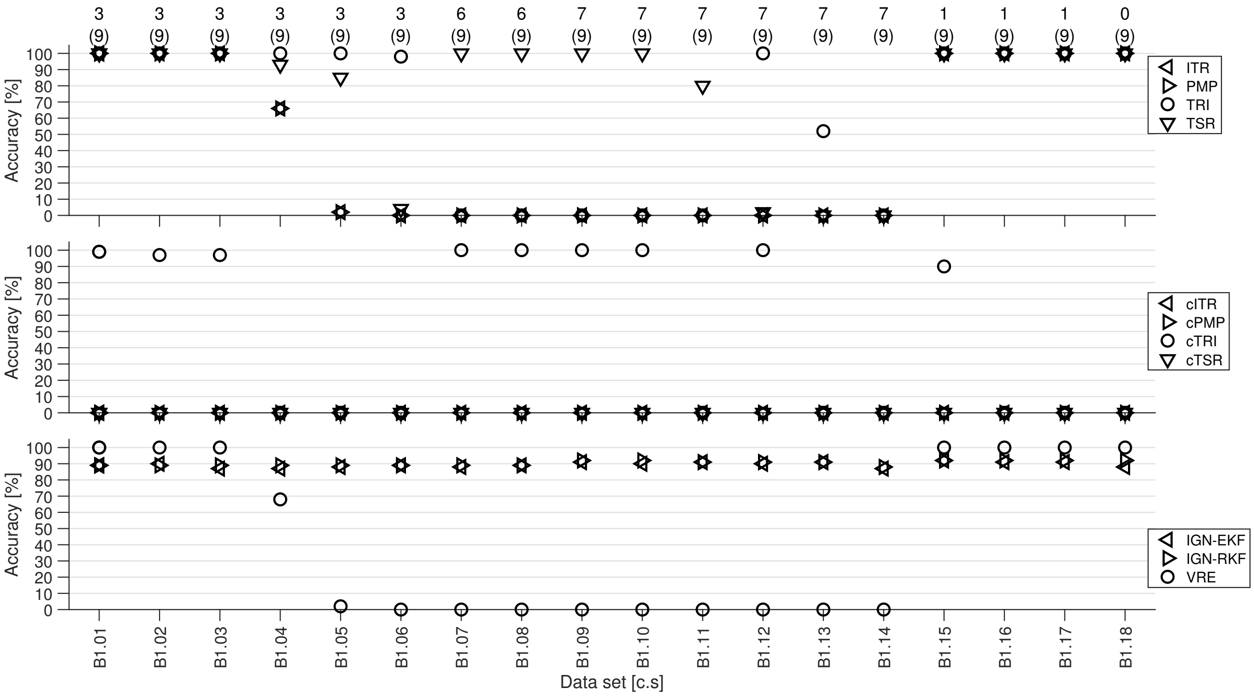


Figure 5: **Simulation data class B1 – Accuracy.** The fraction of correctly identified number of PCs is shown for every data set in the class B1 (B1.1 until B1.6) and model selection criterion. Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE. The numbers at the top of the figure are *(a)* the known number of PCs and *(b)* the number of variables (in parentheses). The reported accuracy for IGN-EKF and IGN-RKF is higher than 85% in all cases. These are the only criteria that achieve a non-zero accuracy for all data types.

It is worth noting that the data types B1.1 to B1.6 are simulated with a covariance matrix composed of blocks, each of which has the same value in all row-column positions. The variations of these covariance matrices lead to a gradual decrease in the fraction of non-noisy variance to the total variance ($\sum_k^K \boldsymbol{\psi}_k / \sum_k \boldsymbol{\psi}_k + \sigma_\epsilon$) from just above 45% to below 15%. This coincides with degraded accuracy for data types B1.4 to B1.6 for ITR, PMP, TSR, cITR, cPMP, cTSR, and VRE. This suggests that these model selection criteria are particularly sensitive to high levels of noise, in turn implying that the IGN-EKF and IGN-RKF are particularly robust. Figure 6(top) shows the fraction of non-noisy variance to the total variance. This fraction is also lower for data sets B1.17 and B1.18 where almost all criteria provide an excellent accuracy (above 90%). This means that the fraction of noise variance alone cannot completely explain our results. For this reason, Figure 6(bottom) shows the same fraction divided by the known number of PCs, i.e. the average non-noisy fraction of the total variance contained in a single PC. A comparison of this panel with the performances in Figure 5 suggests that this property of the covariance matrix correlates well with the obtained performance: decreasing the relative amount of non-noisy variance in a single component corresponds to a degraded performance of ITR, PMP, TSR, cITR, cPMP, cTSR, and VRE. Conversely, the accuracy of IGN-EKF and IGN-RKF appears largely insensitive to this property, even if the non-noisy fraction of the variance is as low as 2.3% per component (data set B1.14). To a lesser extent, this is also true for the TRI and TSR criteria.

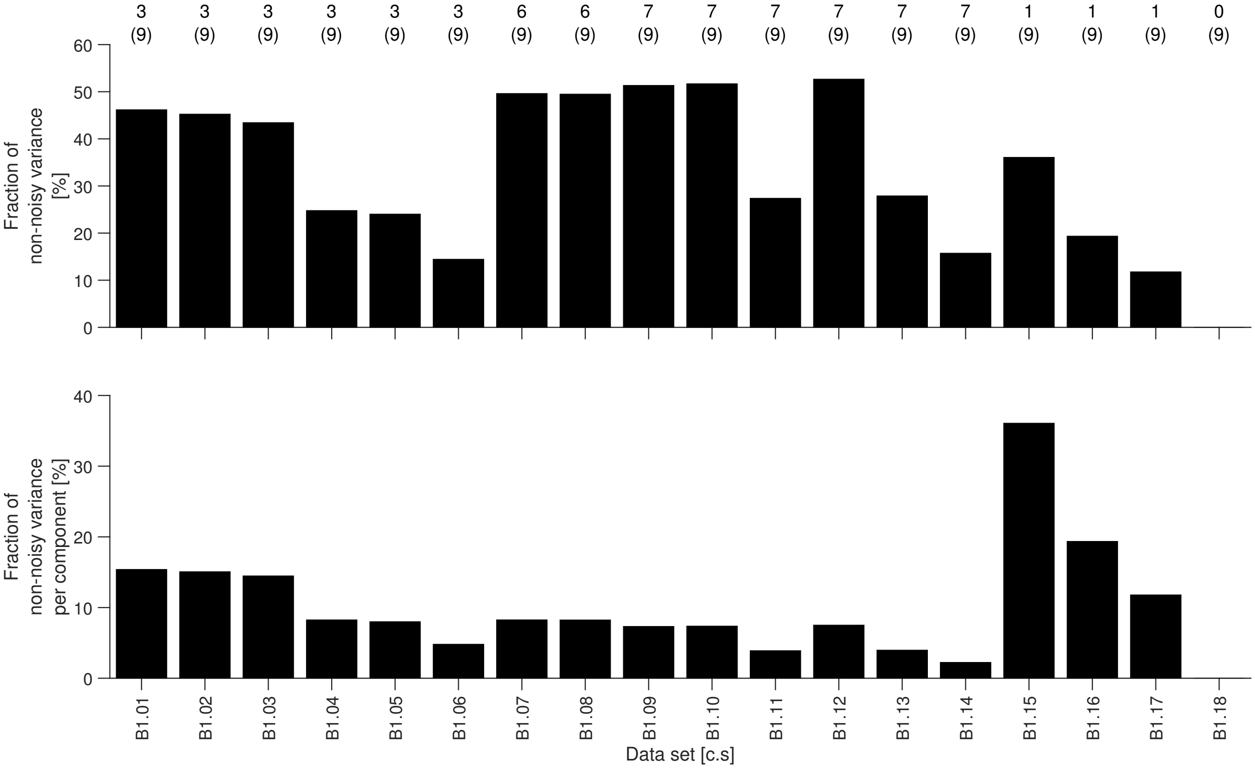


Figure 6: **Simulation data class B1 – Non-noisy variance.** Top: Fraction of non-noisy variance to total variance. Bottom: Fraction of non-noisy variance to total variance per principal component. The numbers at the top of the figure are *(a)* the known number of PCs and *(b)* the number of variables (in parentheses).

We inspect the distribution of the selected number of PCs for data type B1.6. This distribution is visualized in Figure 7. Most importantly, one can see that ITR, PMP, and VRE tend to underestimate the number of PCs while TSR, cITR, cPMP, cTRI, and cTSR overestimate the number of PCs. IGN-EKF and IGN-RKF are fairly accurate yet overestimate the number of PCs with at most 3 PC for less than 15% of the simulated data sets. The equivalent figures for all remaining data types in the B1 class are shown in the *Supporting Information* (Section E).

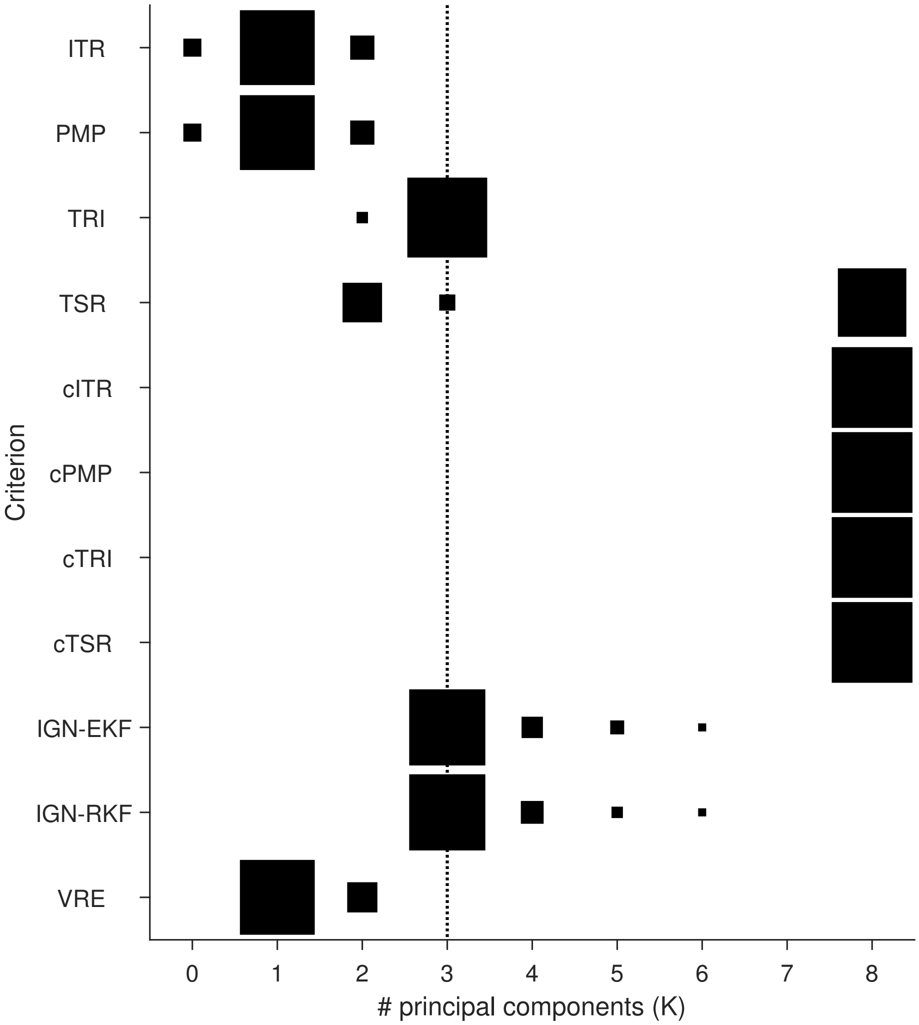


Figure 7: **Simulation data type B1.18 – Distribution of the identified number of principal components.** For each number of PCs (K, bottom to top), and for every noise level and every criterion (left to right), a black box is shown with a surface proportional to the number of data instances for which K component are selected. The vertical line indicates the simulated value of K (3).

The average computational requirements are shown per criterion and per data type in Figure 8. The computational effort appears insensitive to the data type within the B1 class for all criteria except ITR. The ITR criterion can demand computing times that are up to 10 times larger for the covariance structures with a large number of PCs (B1.7 to B1.14) relative to the requirements for the other data types (B1.1 to B1.6, B1.15 to B1.18). Secondly, one can see that the computational requirements are the lowest with VRE, requiring less than 5 ms per repetition of the model selection procedure. This is followed by the IGN-RKF, which requires under 50 ms per repetition, and TRI, PMP, and IGN-EKF, which require under 100 ms. For some imputation procedures, it appears computationally efficient to implement the corrected version (e.g., ITR vs. cITR) whereas the original version is most efficient for others (e.g., TRI vs. cTRI).

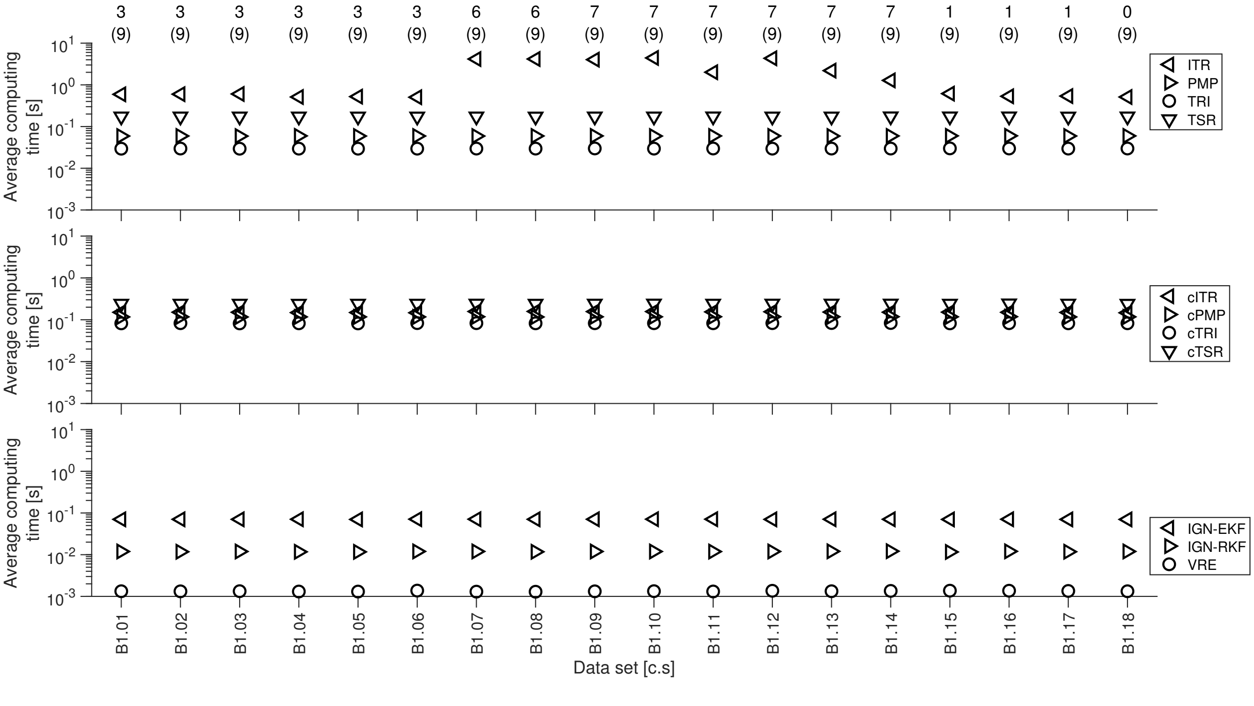


Figure 8: **Simulation data class B1 – Average computation time.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE. The numbers at the top of the figure are *(a)* the known number of PCs and *(b)* the number of variables (in parentheses).

Detailed results for data class B2.

The results for data class B2 lead to the same general observations as for data class

B1. A notable exception is that the performance of the IGN-EKF and IGN-RKF criteria drops below 80% (15% and 75%) for data set B2.14. This is good in comparison to all other criteria, which deliver 0% accuracy for this data set. The non-noisy fraction of the variance equals 1.2% per component in this case, which is about half of the next lowest non-noisy fraction within the B1 and B2 data classes (data type B2.11: 2%). More detailed results can be found in the *Supporting Information* (Section F).

Detailed results for data classes C1, C2, C3, and C4.

Figure 9 shows the fraction of the instances where the identified number of PCs matches the simulated ground truth.

Similarly to the B1 and B2 classes, there are four important observations. First, ITR, PMP, cITR, cPMP, and cTSR fail to identify the correct number of dimensions in the majority of the simulated data sets (accuracy below 5% for all data types). Second, the performance of the remaining imputation-based criteria (TRI, TSR, cTRI) is highly variable, ranging from 0% to 100% accuracy. Among these criteria, cTRI always delivers a 0% accuracy for the highest two noise levels while delivering excellent performance for the lowest three noise levels (above 95%). The TSR criterion performs well (accuracy above 80%) at the lowest three noise levels with data class C1 and C2 yet delivers an accuracy below 20% for data classes C3 and C4, regardless of the noise level. For the C3 data class, the TRI criterion appears well-suited (accuracy of 100%). However, this criterion is not robust to noise with data type C1 and leads to a low accuracy for the C2 and C4 data types (below 5%). Third, VRE delivers a low accuracy across the C classes with values similar to those obtained with ITR and PMP. Fourth, and most importantly, the accuracy obtained with the IGN-EKF and IGN-RKF never drops below 85% and is particularly insensitive to noise. In addition, these two criteria appear to do better with increasing dimensionality and the number of known PCs (both increasing from C1 to C4).

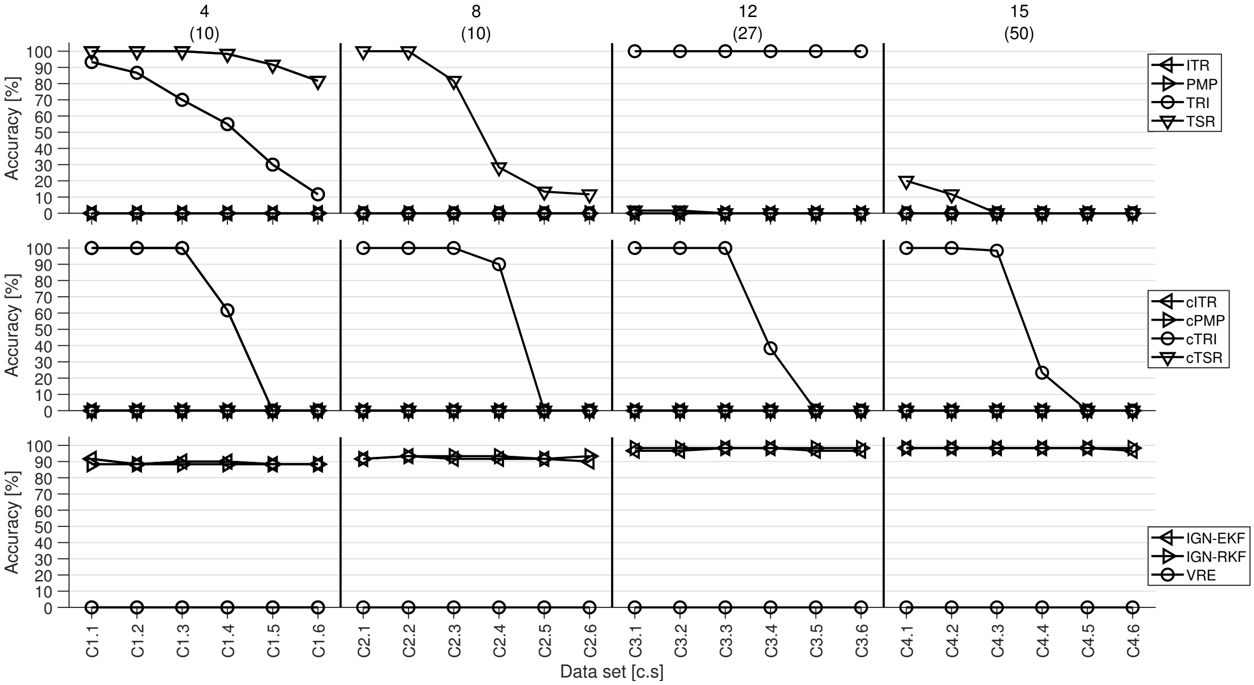


Figure 9: **Simulation data classes C1 to C4 – Accuracy.** The fraction of correctly identified number of PCs is shown for every data type in the C classes (C1.1 until C4.5) and model selection criterion. Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE. The IGN-EKF and IGN-RKF are the only criteria that achieve an accuracy above 85% for all data sets. IGN-EKF and IGN-RKF are the most accurate criteria for noise levels of 40% or higher (C1.4-C1.5, C2.4-C2.5, C3.4-C3.5) and for the whole C4 class (C4.1-C4.5).

Figure 10 shows the relative fraction of non-noisy variance to the total variance for each data type in the C class in the top panel and the same fraction divided by the known number of PCs in the bottom panel. The top panel mainly displays the effect of adding noise, which decreases the amount of non-noisy variance. The fraction of the non-noisy variation divided by the number of PCs increases both with the added noise fraction and the data type index. As with the B1 and B2 classes, the accuracy of the IGN-EKF and IGN-RKF criterion is especially insensitive to low values for this proportion.

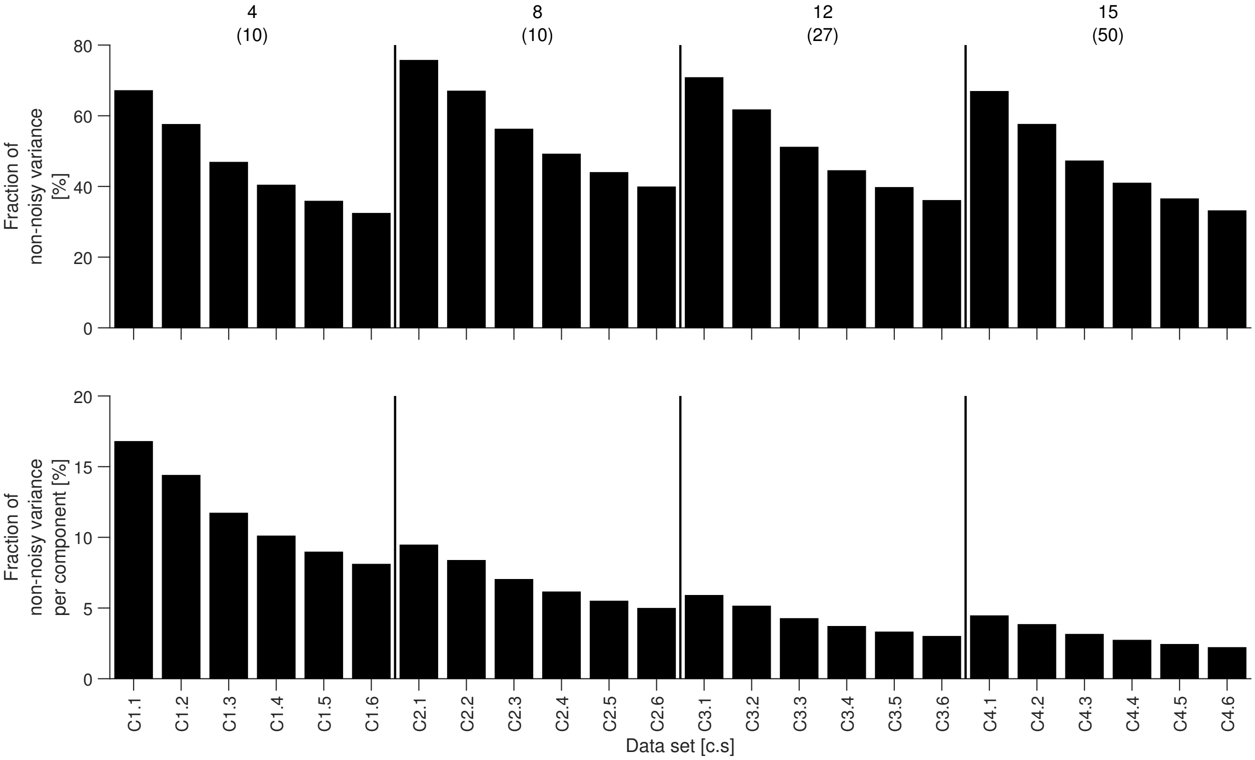


Figure 10: **Simulation data classes C1 to C4 – Non-noisy variance.** Top: Fraction of non-noisy variance to total variance. Bottom: Fraction of non-noisy variance to total variance per principal component.

Figure 11 shows the distribution of identified numbers of PCs for all data set types in the C4 class. The equivalent figures for the C1, C2, and C3 classes are provided in the *Supporting Information* (Section F). Across the C classes, the ITR, PMP, TSR, and VRE criteria underestimate the dimensionality while the cITR, cPMP, and cTSR lead to overestimation. The same happens when using cTRI at high noise levels. The TRI criterion underestimates the number of PCs with at most 3 PCs. The IGN-EKF and IGN-RKF criteria either overestimate or underestimate the number of PCs although with at most 1 PC for all data types.

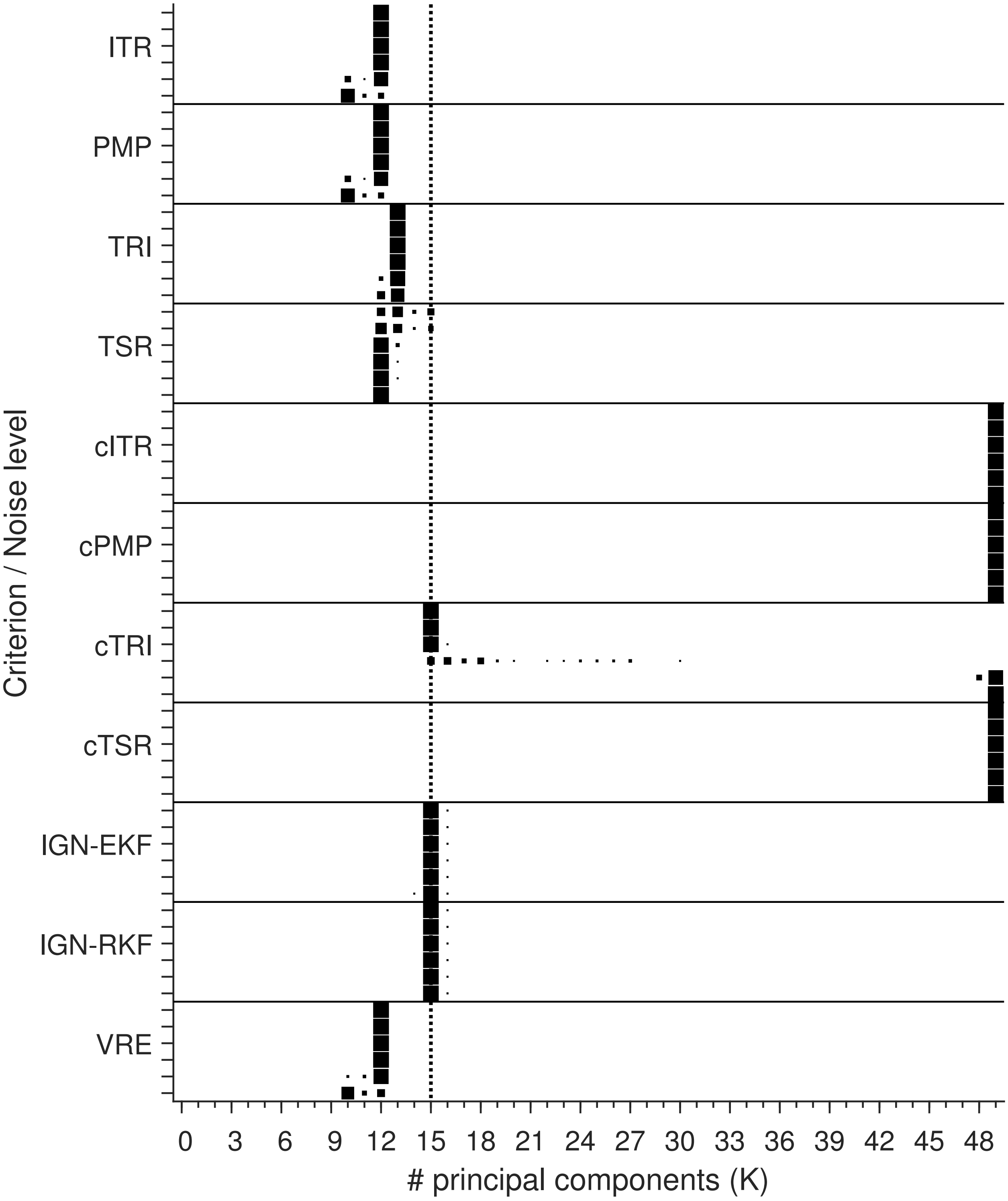


Figure 11: **Simulation data class C4 – Distribution of the identified number of principal components.** A black box is shown with a surface proportional to the number of data instances for which K component are selected. This is executed for each number of PCs (K, left to right), for every criterion (top to bottom), and for every noise level (increasing from top to bottom for each criterion). The vertical line indicates the simulated value of K (15).

Figure 12 displays the average computational effort, measured in seconds, required to complete a single run of each criterion. These results lead to similar conclusions as for the B1 class. However, the data types in the C class exhibit changing numbers of variables and noise, unlike the B1 and B2 classes. The computational requirements for the imputation-based criteria and the IGN-EKF increase with about a factor of 100 when increasing the number of variables from 10 (C1, C2) to 50 (C4). Each of these criteria uses the EKF pattern. In contrast, the computational requirements increase less for the IGN-RKF and VRE criterion, with a factor around 10.

Note that the average run times remain fairly low for all criteria due to the use of a 16-fold cross-validation pattern in the row/sample direction. Exploratory experiments (not shown) suggest that the computational time depends linearly on the number of folds. A ball-park estimate for the run time when applying a leave-one-out pattern in the row direction can be obtained by multiplying the computed run times by 64.

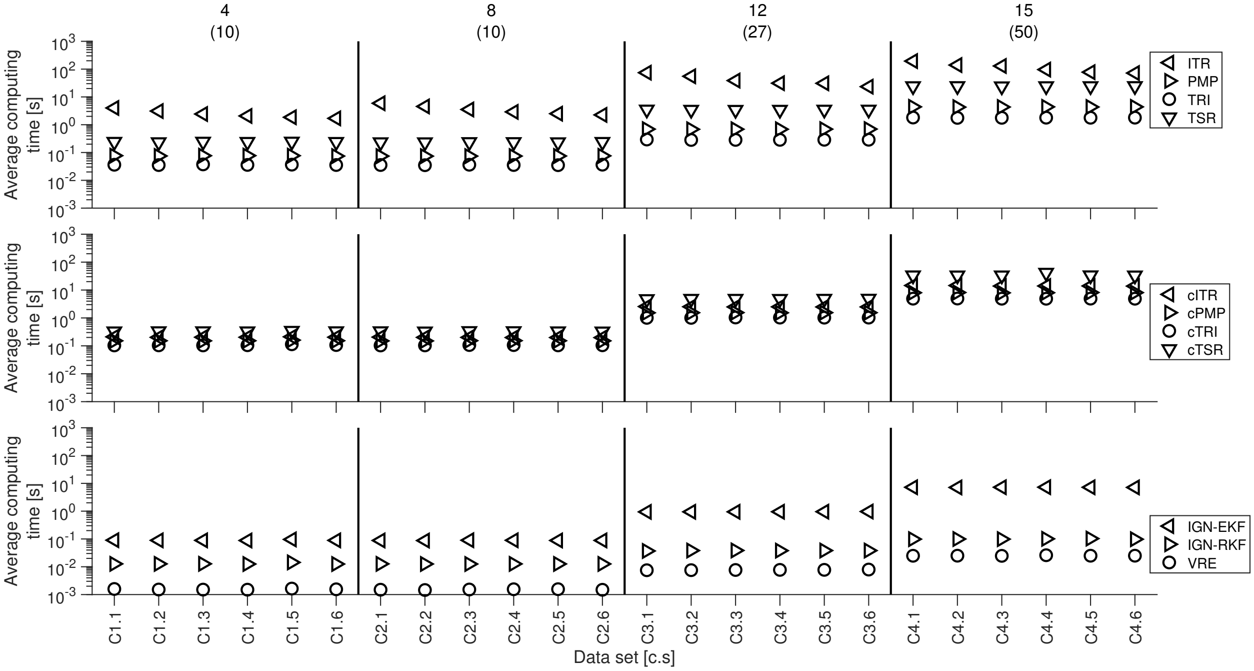


Figure 12: **Simulation data classes C1 to C4 – Average computation time.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE. The numbers at the top of the figure are *(a)* the known number of PCs and *(b)* the number of variables (in parentheses).

Experimental data sets

Nitrogen species data sets

Nitrogen species data set 0 – Wavelengths 200-735 nm Figure 13 displays the complete set of absorbance measurements associated with the first validation block. One can see that the spectra are sensitive to variations of nitrite and nitrate concentrations in the range from 200 nm until about 420 nm. One can also clearly observe the known secondary absorbance peak of nitrate around 300 nm and of nitrite around 355 nm.[39] On the left hand side of the spectra, i.e. below 250 nm, one can observe high but rather insensitive absorbance measurements. This is a region where the Beer-Lambert law for absorbance measurements does not apply as the device is subject to saturation phenomena, i.e. virtually all light in this wavelength range is absorbed leading to meaningless readings by the device, as was demonstrated before for this kind of device.[40] As a result, analyzing these data by PCA, let alone using these data to test the proposed model selection criteria, makes little sense. For this reason, we continue below with an analysis based on a subset of the absorbance measurements, which are expected to depend linearly on the nitrite and nitrate concentrations in the experimental solutions.

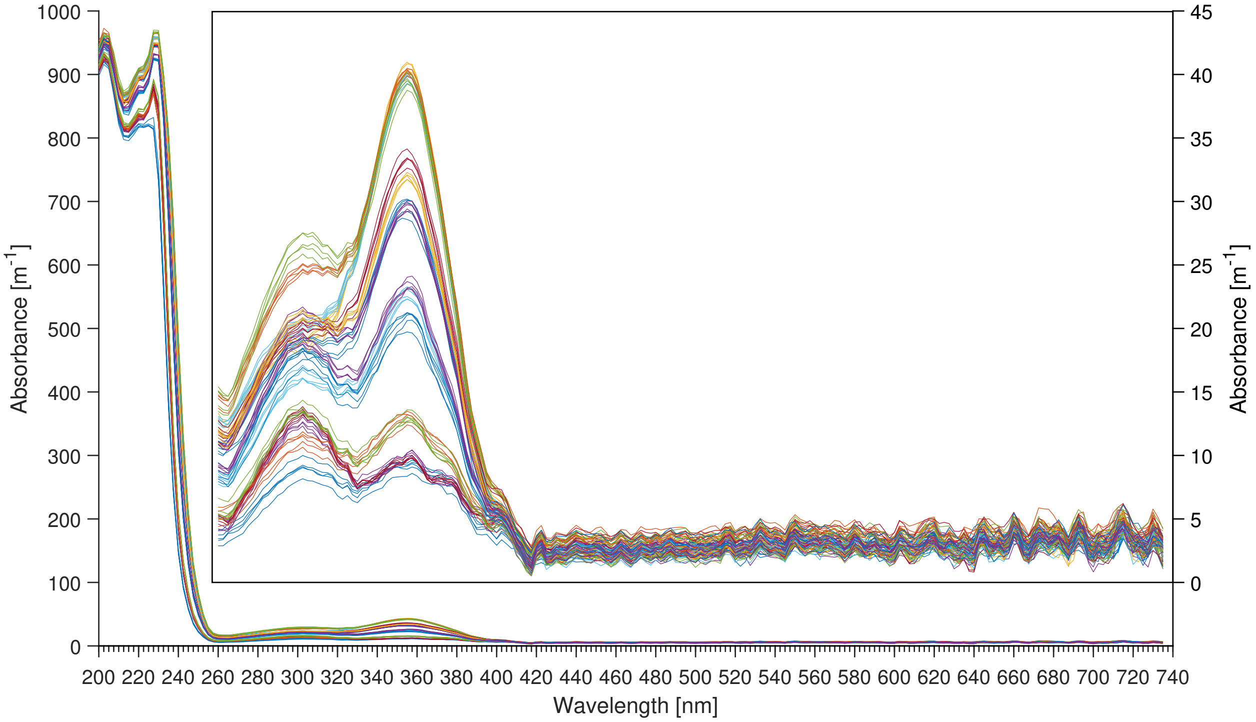


Figure 13: **Nitrogen species data – Validation block 1.** Absorbance measurements are shown as a function of the wavelength. The inset on the top-right shows a detailed view on the measurements taken at 260 nm or higher wavelengths.

Nitrogen species data set 1 – Wavelengths 285-735 nm All model selection criteria are tested after the absorbance measurements corresponding to wavelengths below 285 nm are removed (nitrogen species data set 1). Figure 14 shows the profiles of all model selection criteria. These profiles all share the property that the model selection criterion decreases fast over the first two to three PCs, as one might expect given the anticipated number of PCs (2). However, the minima of these criteria are observed from five PCs (TRI) to 180 PCs (e.g., TSR, cTSR). TRI and cTRI are the only criteria delivering reasonable estimates of five PCs (TRI) and 20 PCs (cTRI). All other criteria deliver dimensionality estimates that are significantly higher (above 50 PCs). This is not so surprising for these imputation-based criteria as these were shown to select too many PCs with the simulated data sets also. This is more surprising for IGN-EKF and IGN-RKF, which select the maximum number of PCs here, a stark contrast with the accurate estimates obtained with the simulated data sets.

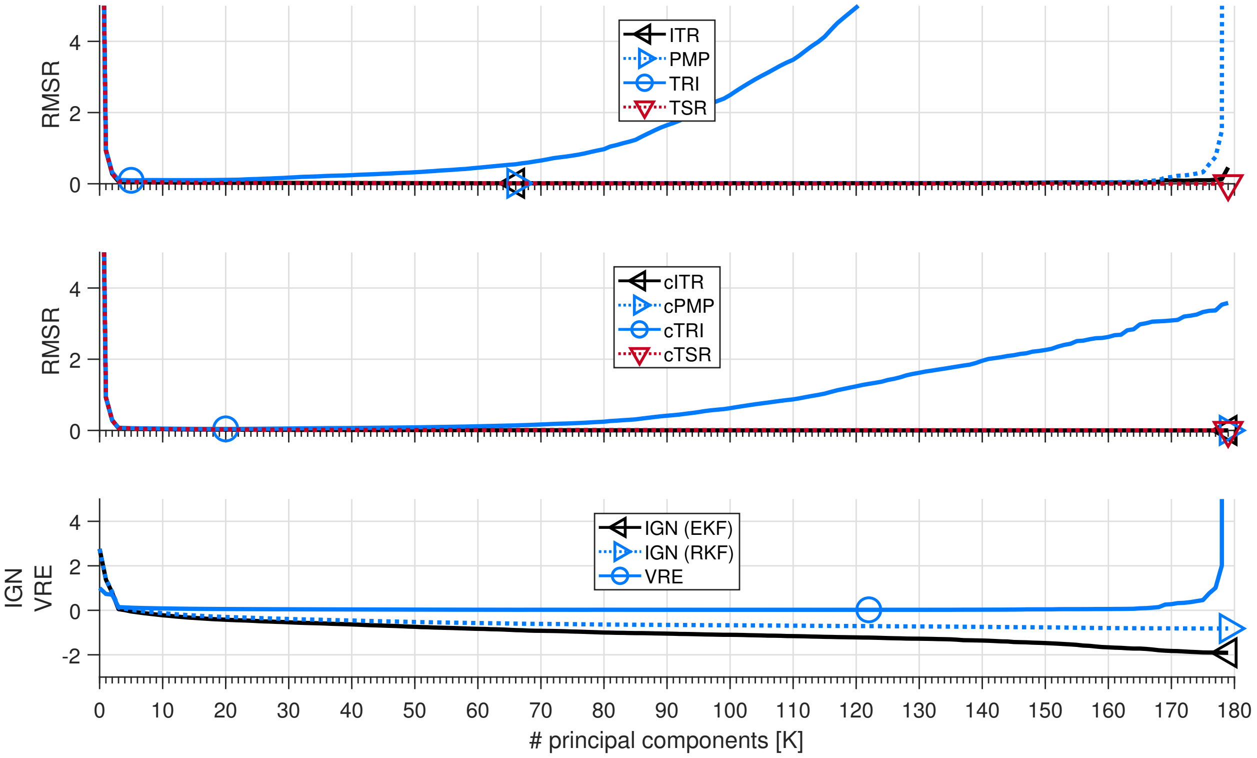


Figure 14: **Nitrogen species data set 1 – Model selection criteria profiles with original data.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE.

To investigate this further, Figure 15 shows the eigenvectors corresponding to the 3rd, 4th, and 5th PC. Assuming that the Beer-Lambert law is valid - and thus also the PCA model structure - these loading vectors should mainly reflect uncorrelated noise properties of the data. In contrast, one can see that the 3rd PC appears to explain a uniform effect across the wavelengths in the visible range (400-735 nm). This cannot be explained by an effect of the absorbing nitrogen species and is therefore considered an artifact in the data, suggestive of strongly correlated type of measurement errors. The scores for this PC were inspected visually to check for a temporal effect in the experimental data but none could be identified. The 4th and 5th PC appear to represent strongly correlated features with relatively large magnitudes in the red range of the spectrum (600-700 nm). This cannot be explained by an effect of nitrite or nitrate either. Since the oscillations follow a regular pattern with a peak-to-peak distance of about 30 nm and there is no physical relationship to known light absorbing properties of nitrite and nitrate, we speculate that this is a result of strongly correlation measurement errors at neighboring absorbance wavelengths. Similar patterns are observed for higher-order PCs as well ($k \geq 6$, not shown).

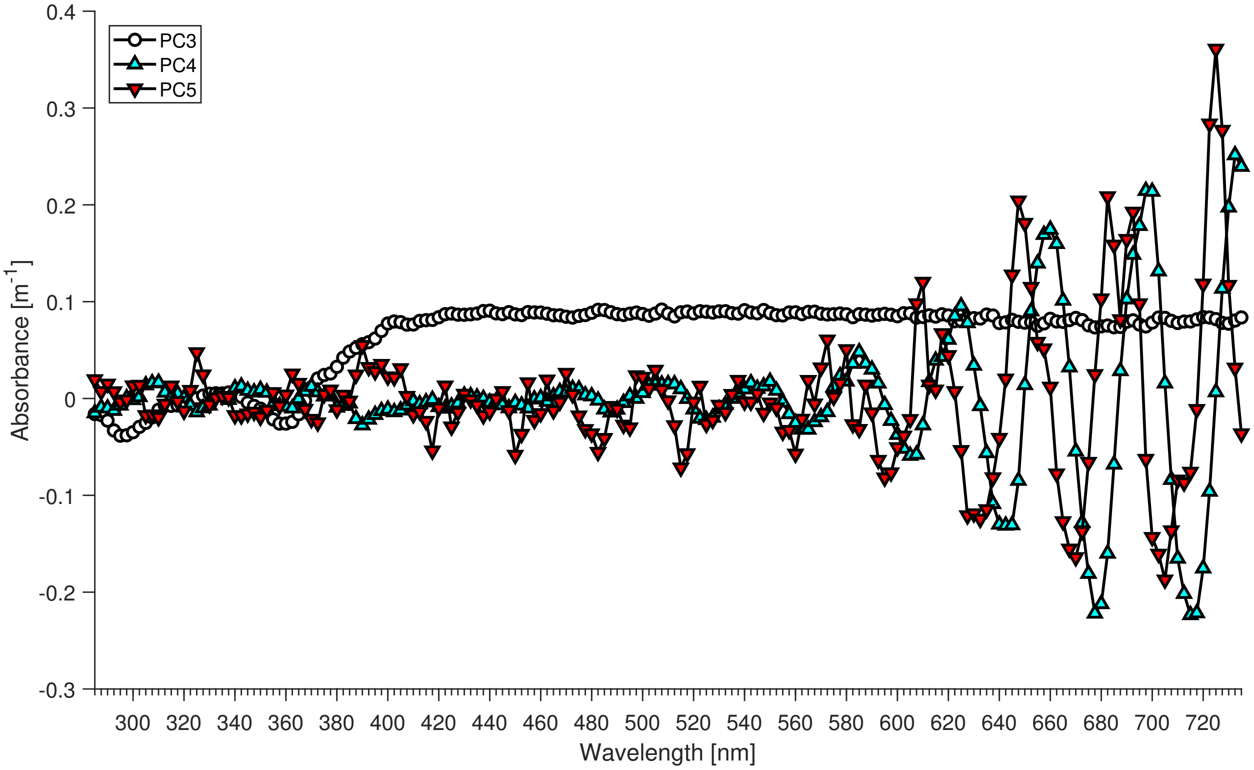


Figure 15: **Nitrogen species data set 1 – Loading vectors for PC 3 to 5.**

Nitrogen species data set 2 – Wavelengths 285-385 nm Considering that the higher-order PCs appear to describe variation in the visible range of the absorbance spectra (400-750 nm) primarily, we now apply the studied model selection criteria to the nitrogen species data set 2, which only contains absorbance measurements for wavelengths between 285 and 385 nm. Figure 16 shows the obtained results. In this case, TRI and cTRI deliver the anticipated number of PCs (2). This suggests that the distribution of the noise in the non-absorbing region of the spectra affects the accuracy of these two criteria. In contrast, all other criteria select 30 PCs or more. A possible explanation is that the presence of correlated data features identified before could not be removed entirely and that TRI and cTRI are relatively robust to such features.

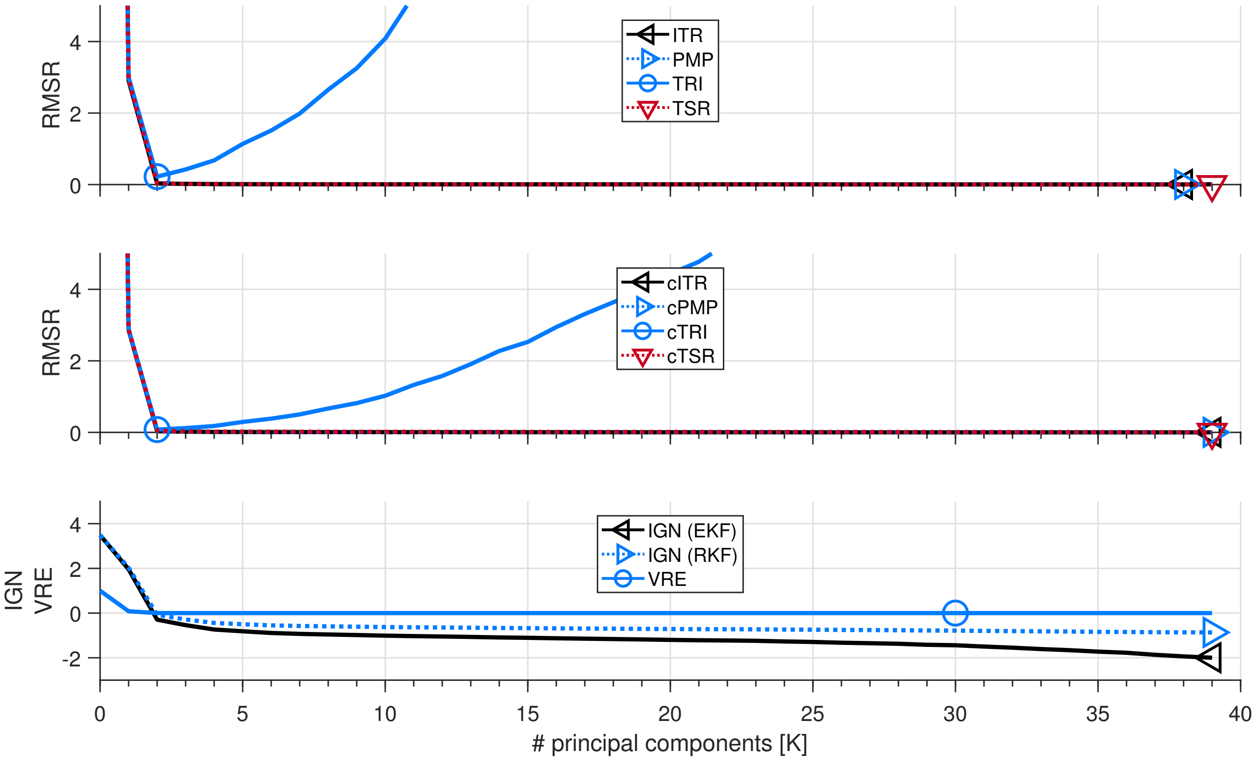


Figure 16: **Nitrogen species data set 2 – Model selection criteria profiles with original data.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE.

Figure 17 shows the eigenvectors associated with PC 3, 4, and 5. Each of these exhibits an oscillating profile, suggesting that correlated noise remains present in the absorbance measurements in the ultraviolet range, i.e. where both nitrite and nitrate absorb light.

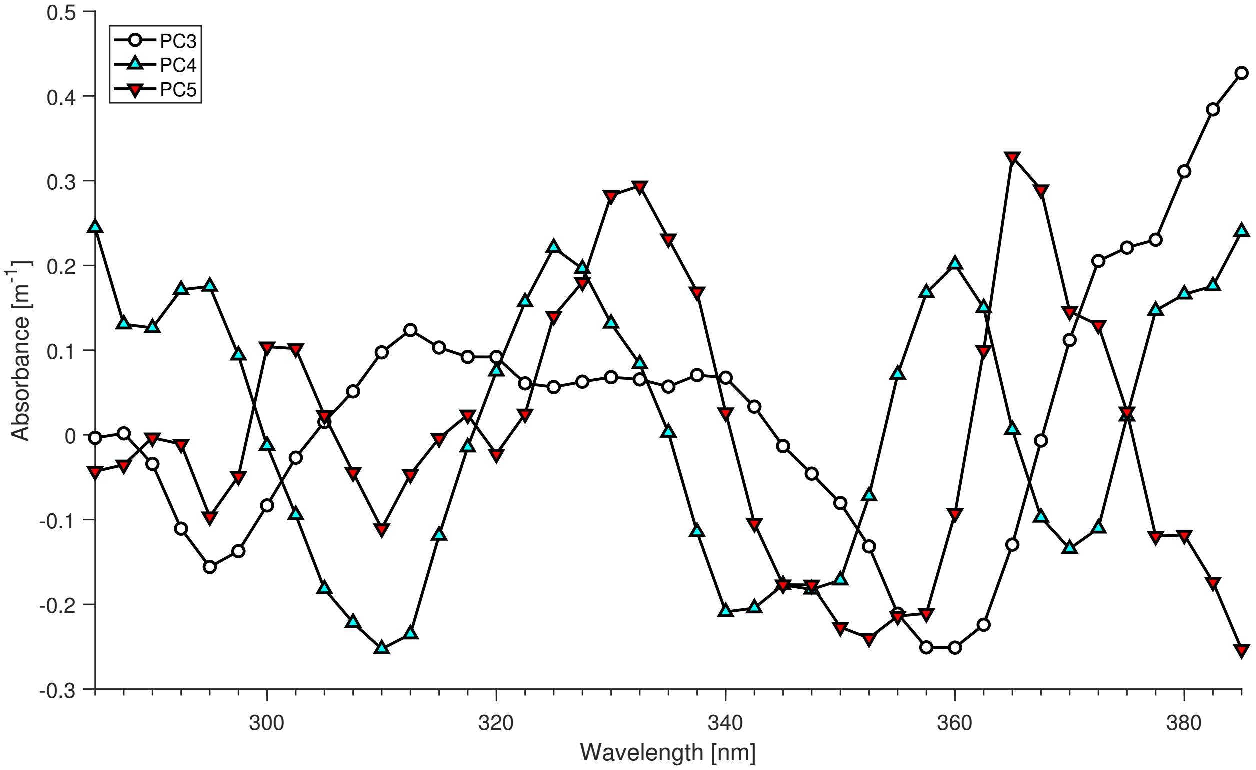


Figure 17: **Nitrogen species data set 2 – Loading vectors for PC 3 to 5.**

Simulated data set – Wavelengths 285-385 nm

To evaluate the idea that non-spherical noise is a key factor in the performance of the model selection criteria, we simulate data with a PPCA model with 2 PCs identified with the nitrogen species data set 1. That is, we identify the mean and $\Sigma^{(K)}$ with the complete data set and assuming $K = 2$. We simulate a new data set of the same dimensions according to the PPCA model. We use the computed scores obtained with (17) to do this. This means that the distribution of the simulated components is similar to the distribution of the two most important principal components in the experimental data. The resulting data does not adhere to a normal distribution, due to the experimental design, while the distribution of the measurement noise adheres to the spherical noise assumption. We then repeat each of the proposed model selection criteria on this artificial data set. The result of this is shown in Figure 18. In this case, 7 out of the 11 criteria select two PCs (ITR, PMP, TSR, cTRI, IGN-EKF, IGN-RKF, VRE). TRI selects one PC in this case whereas the remaining criteria continue to overestimate the number of PCs (cITR, cPMP, cTSR). By forcing the noise properties of the data to adhere to the assumed PCA model structure, we obtained the expected result with most model selection criteria.

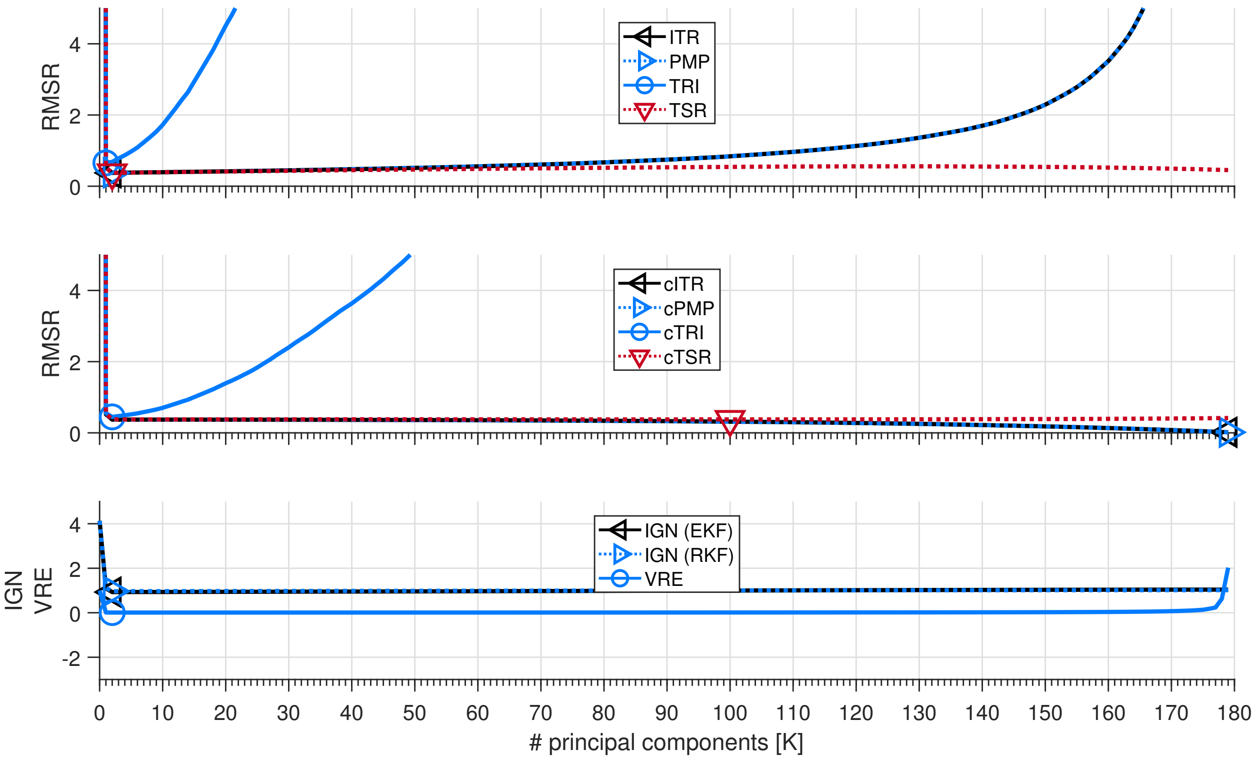


Figure 18: **Nitrogen species data set 1 – Model selection criteria profiles with simulated data.** Top: ITR, PMP, TRI, TSR; Center: cITR, cPMP, cTRI, cTSR; Bottom: IGN-EKF, IGN-RKF, VRE.

Metal ion data set

The results obtained with the metal ion data set are shown in the *Supporting Information* (Section G). The nature of these results is very similar to the results obtained with the nitrogen species data sets. However, the higher-order loading vectors (PC 4 and higher) do not exhibit an oscillatory pattern with a regular peak-to-peak distance in this case. This leads us to speculate that nonlinear effects of the absorbing metal ions on the absorbance spectra could contribute to the difficulty in identifying the known number of experiment factors. Still, this data set is considered to exhibit correlated noise as well[33].

Discussion

Cross-validated ignorance score as a tool for dimensionality selection

With this study, we propose a new criterion for dimensionality selection in PCA. It is based on the application of the ignorance score to the PPCA model. Simulation results show that using the ignorance score delivers excellent accuracy in identifying the correct number of PCs when the assumed PCA model structure is correct. Most interesting is that this model selection criterion clearly outperforms all imputation-based criteria in overall accuracy of the selected dimensionality. Moreover, this approach appears most fruitful when the fraction of non-noisy variance (information) is small and spread thinly over many components. In addition, the ignorance score can be applied successfully with a row-wise K-fold cross-validation (RKF) pattern. The row-wise K-fold cross-validated ignorance score is therefore the first known PCA model selection criterion with the following properties: *(i)* it produces an accurate and meaningful minimum in the cross-validated performance criterion, *(ii)* it is tuned well to the purpose of data compression, and *(iii)* its efficiency scales well with the dimensionality of the data set. Both IGN-EKF and IGN-RKF are reliable estimators as long as the linear model structure and the least-squares objective matches the analyzed data. In practice, this level of match between model and reality may be hard to attain, so that a human-in-the-loop approach remains advised for PCA model selection as of yet.

It is worth noting that the proposed IGN-RKF criterion is the only model selection criterion which can be interpreted as a distance in the J-dimensional data space. In addition, it is the only known RKF-based criterion producing a clear minimum when plotted against the number of principal components, while using validation data only during model testing and not during model calibration. Thus, the model selection criterion effectively avoids data leakage,[41] which consists of using the same data twice for calibration and prediction. We speculate that this is one of the reasons for the rather poor performance of some of the tested criteria. For instance, the cITR, cPMP, cTRI, and cTSR criteria all use the imputed data during the data augmentation step before treating these as missing data, thus leading to imputation errors which depend on the imputed data. A definite proof that the prevention of data leakage is the key factor in the robust performance of IGN-EKF and IGN-RKF is considered beyond the scope of this work.

Cross-validated ignorance score as a tool for detection of model structure deficits

Tests with experimental data sets indicate that all considered criteria select a number of PCs that is higher than expected. Results generally improve when variables that contain mostly noise are removed or when using simulated approximations of the experimental data without correlated noise. This suggest that the experimental data contain artefacts that cannot be explained by the Beer-Lambert law. Instead, it is likely that the data exhibit nonlinear effects and measurement errors with unequal variances or strong correlation. Indeed, the PCA and PPCA models are only optimal in the maximum likelihood sense when the measurement errors are drawn from the same univariate normal distribution. The presence of non-spherical measurement errors and nonlinear effects may in part explain why relatively complex models, i.e. with a large number of PCs, are necessary to obtain good predictive performance in practice.[40] In the opinion of the authors, this means that the studied model selection criteria can be a viable tool to detect deviations from the assumed model structure, possibly including the presence of heteroskedastic and/or correlated noise or nonlinear effects in the data.

Utility of the imputation-based model selection criteria

This study concentrates on finding a good model selection criterion for the purpose of data compression with PCA. We conclude that many imputation-based methods are ill-suited for this purpose as they lead to inaccurate dimensionality selection. It is important to note however that these criteria select models based on their ability to impute missing data. It follows that the best model selection criteria for data compression are not the same as the best model selection criteria for optimal imputation of missing data. This is one of the main messages in earlier work[27] and is also supported by this study. A cursory look into the RMSR profiles (not shown) suggests that the TSR method is optimal for least-squares data imputation, in line with existing recommendations[27]. The utility of describing missing data with both a mean and a variance, as used for computation of IGN-EKF, remains to be studied in detail however.

VRE as a proxy for ITR and PMP

The profiles of the ITR, PMP, and VRE criteria are very similar for most simulated data sets. The similarity between ITR and PMP is not a surprise as they deliver the same result except in numerically challenging cases[27]. ITR and PMP imputation both minimize the reconstruction error of a single variable in the least-squares sense so that ITR- and PMP-based cross-validation leads to a minimization of the variance of this reconstruction error. The VRE computed for faults in single sensors, as in this study, is an estimate of this variance. It is therefore also not surprising that the VRE criterion could be an excellent and fast approximation to the ITR and PMP criterion. Note however that VRE assumes that the PCA model structure is correct during model selection whereas the ITR and PMP criterion do not. Note that the ITR and PMP criteria and the VRE criterion are not as similar in the case of the experimental data sets. This corroborates the idea that the PCA model structure may be inadequate for these experimental data sets.

Links to work in statistical process control

The ignorance score as used for IGN-RKF was shown to be equivalent to the well-known Mahalanobis distance, however computed with the covariance matrix estimated through the PPCA model. This makes it similar – yet not equal – to *(a)* the Mahalanobis distance based on the empirical covariance matrix,[2,42,43] *(b)* the Mahalanobis distance based on exploratory factor analysis,[44] or *(c)* the combined index composed of the Hotelling's T^2 statistic and the squared prediction error statistic.[45] Logically, this means that ignorance score could be a useful statistic for anomaly and fault detection based on principal component analysis. Importantly, the ignorance score corresponds to a distance in the original J-dimensional space and neither in the principal component or the residual space alone. This is similar to the combined index.[45] For this reason, we speculate that interpreting this combined index as a log-likelihood also enables its use for model selection based on RKF, similar to the IGN-RKF criterion.

Open avenues for research

In view of clarity, this work is focused on demonstrating the use of the ignorance score for dimensionality selection in the most trivial case for principal component analysis, i.e. assuming linear effects and homoskedastic and uncorrelated noise. Considering that our results suggest that the experimental data studied in this work do not share this property, we explore below whether the ignorance score could also be applied to alternative model structures, which may be better adjusted to these data sets.

Variational auto-encoders[46,47] and Gaussian process latent variable models[48] are interpreted as nonlinear versions of PPCA and permit a generative, probabilistic interpretation. Exploratory factor analysis[49] and target factor analysis[50,51] may be used to find an optimal K-dimensional hyper-plane describing a data set similar to PCA, yet allowing for unequal noise variance estimates in the diagonal error covariance matrix. Other models, such as combined PCA-ICA models[52] and the heteroscedastic latent variable model[53] are explicitly developed to account for non-Gaussian distributions of the non-noisy variations in the data. These models deal explicitly with nonlinear effects but not with deviations from the uncorrelated noise assumption.

In recent years, several modified PCA models have been proposed to allow for non-diagonal forms for the error covariance matrix. This matrix can be assumed known,[34] estimated independently,[33,36] or estimated simultaneously.[9] In a recent article[54], this type of models have been evaluated for the purpose of missing data imputation, but not yet for the purpose of model selection. Note that each of these models are similar to PCA in the sense that the eigenvalues associated with the residual space are not modified. Applying the ignorance score for dimensionality selection in such models thus requires modifications of these models, e.g. by applying the variance deflation step as in PPCA.

Other approaches to deal with correlated noise may consist of feature engineering prior to PCA analysis. For example, multi-scale principal component analysis[55–57] and functional principal component analysis[6,58] are both based on the computation of new features, which typically are linear combinations of the original data prior to model calibration. This transformation may very well produce features with a noise covariance matrix approximating a diagonal matrix more closely. When so, this may improve the fit of the PPCA model and reduce the selected dimensionality. Conversely, specialized PCA models and feature generation should be explored as a way to enhance the robustness of the cross-validated ignorance score, specifically by accounting for nonlinear effects and for unknown or poorly understood noise properties.

Conclusions

This work addresses the important yet challenging selection of the optimal number of latent variables in principal component analysis (PCA). Two variations of a newly proposed cross-validated ignorance score for PCA model selection are compared to established model selection criteria. Our most important findings are:

- Benchmarking results reveal that the proposed ignorance score delivers performances of 80% or higher for every simulated data set. This is unlike any other model selection criterion included in our study, all of which deliver an accuracy of 0% for at least one of the data types.

- Simulation results show that our proposed ignorance score is the most accurate model selection criterion for data sets with relatively low proportions of non-noisy variation spread over relatively large numbers of principal components.
- Experimental results revealed that devices for in-situ measurement of spectrophotometric absorbance spectra in aquatic systems are prone to produce data that violate the PCA model structure. This is true for two experimental data sets collected for the development and evaluation of latent variable models. One probable cause for this is the presence of measurement errors with a non-spherical distribution, i.e. with heteroskedastic and correlated measurement errors.
- The ignorance score is a valuable addition to the tool set for both PCA model selection. It is likely useful to detect PCA model structure deficits and may be a promising statistic for fault detection as well.

Acknowledgement

The authors would like to thank Karin Rottermann, Sylvia Richter, and Kai Udert for their contributions to the work presented in this paper. The study has been made possible by the Swiss National Foundation (project: 157097) and Eawag Discretionary Funds (grant number: 5221.00492.012.02, project: DF2018/ADASen).

Supporting Information Available

The Supporting Information consists of a single package including:

- Detailed procedures for data simulation
- Detailed procedures for experimental data collection
- Detailed procedures for PCA-based imputation

- Detailed results not discussed in the main body of the text
- Self-contained software, which produces all results, presented in this work
- All experimental data collected for the purpose of this study

Table 2: List of symbols

Symbol	Description	Dimensions
ξ ($\tilde{\xi}$)	Fault direction	$J \times 1$
$\boldsymbol{\epsilon}$	Vector of measurement errors	$J \times 1$
$\boldsymbol{\theta}$	Distribution parameters	
$\boldsymbol{\lambda}$	Vector of eigenvalues	$K \times 1$
$\tilde{\Sigma}$ ($\boldsymbol{\Sigma}^{(K)}$)	Covariance matrix (estimate)	$J \times J$
σ_{ϵ}	Estimated noise variance	1×1
ϕ	Expected variance of imputed value	1×1
$\boldsymbol{\Psi}$	Diagonal matrix with latent variable variances on the diagonal	$K \times K$
$\boldsymbol{\psi}$	Vector of latent variable variances	$K \times 1$
$\mathbf{A}$	Matrix of regression coefficiens	$K \times K$
B	Number of data blocks	1×1
b	Data block index	1×1
c	Simulated data class	1×1
$\boldsymbol{d}$	Vector of intercepts	$K \times 1$
e	Simulated data noise level	1×1
$\mathbf{I}_J$	Identity matrix	$J \times J$
$\mathcal{I}(y)$	Scalar-valued ignorance function	1×1
I ($I^{(c)}$, $I^{(v)}$)	Number of samples / rows (for calibration, validation)	1×1
i ($\mathbf{i}^{(c)}$, $\mathbf{i}^{(v)}$)	Sample / row index (rows for calibration, rows for validation)	1×1
J	Number of variables / columns	1×1
j	Variable / column index	1×1

Continued on next page

Table 2 – *Continued from previous page*

Symbol	**Description**	**Dimensions**
$-j$	Indices of all variables except j	$(J-1)\times 1$
K ($\overline{K}$)	Number of principal components (maximum)	1×1
k	Index of principal component	1×1
$L(\bullet)$	Scalar-valued likelihood function	1×1
l	Variable / column index	1×1
$\mathbf{Q}$	Matrix of model performance measures	$I\times J$
$\boldsymbol{q}$ ($\boldsymbol{q}^{(v)}$, $\boldsymbol{q}^{(v,j)}$)	Vector of model performance measures (for block v, variable j)	$I\times 1$ ($I^{(v)}\times 1$)
r	Simulated data repetition	1×1
$\mathbf{S}$	Diagonal matrix of singular values	$\overline{K}\times\overline{K}$
s	Simulated data type	1×1
$\mathbf{T}$ ($\mathbf{T}^{(c)}$, $\mathbf{T}^{(v)}$)	Principal scores (calibration, validation)	$I\times\overline{K}$ $\left(I^{(c)}\times\overline{K}\right)$
$\mathbf{U}$ ($\mathbf{U}^{(c)}$)	Standardized principal scores (calibration)	$I\times\overline{K}$ $\left(I^{(c)}\times\overline{K}\right)$
$\mathbf{V}$ ($\mathbf{V}^{(*)}$	Matrix of loading vectors (for augmented data)	$J\times\overline{K}$ $(J+(\overline{K})\times\overline{K})$
v	Validation data block index	1×1
$\mathbf{W}$	Matrix of data-generating loading vectors	$J\times K$
$\mathbf{X}$	Matrix of latent variables	$I\times K$
$\boldsymbol{x}$	Vector of latent variables	$K\times 1$

Continued on next page

Table 2 – *Continued from previous page*

Symbol	Description	Dimensions
$\mathbf{Y}$ ($\mathbf{Y}^{(c)}$, $\mathbf{Y}^{(v)}$)	Noisy measurement matrix (calibration, validation)	$I \times J$ ($I^{(c)} \times J$, $I^{(v)} \times J$)
$\mathbf{Y}^{(c)}$ ($\mathbf{Y}^{(c,*)}$, $\mathbf{Y}^{(v,*)}$)	Augmented calibration data (calibration, validation)	$I \times (J+K)$ ($I^{(c)} \times (J+K)$, $I^{(v)} \times (J+K)$)
$\hat{\mathbf{Y}}$ ($\hat{\mathbf{Y}}^{(c)}$, $\hat{\mathbf{Y}}^{(v)}$)	Estimated measurement matrix (calibration, validation)	$I \times J$
$\boldsymbol{y}$	Vector of measured variables	$J \times 1$
y	Scalar measurement	1×1

References

(1) Wise, B. M.; Ricker, N. L.; Veltkamp, D. F.; Kowalski, B. R. A theoretical basis for the use of principal component models for monitoring multivariate processes. *Process Control and Quality* **1990**, *1*, 41–51.

(2) Tong, H.; Crowe, C. M. Detection of gross erros in data reconciliation by principal component analysis. *AIChE Journal* **1995**, *41*, 1712–1722.

(3) Joliffe, I. *Principal component analysis*, 2nd ed.; Springer, New York, USA, 2002; p 487.

(4) Ramaker, H.-J.; van Sprang, E.; Westerhuis, J.; Smilde, A. The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chem. Intell. Lab. Syst.* **2004**, *73*, 181–187.

(5) Villez, K.; Ruiz, M.; Sin, G.; Colomer, J.; Rosén, C.; Vanrolleghem, P. A. Combining multiway principal component analysis and clustering for efficient data mining of historical data sets of SBR processes. *Water Science and Technology* **2008**, *57*, 1659–1666.

(6) Maere, T.; Villez, K.; Marsili-Libelli,; S.,; Naessens, W.; Nopens, I. Membrane bioreactor fouling behaviour assessment through principal component analysis and fuzzy clustering. *Water Research* **2012**, *46*, 6132–6142.

(7) Perelman, L.; Arad, J.; Housh, M.; Ostfeld, A. Event detection in water distribution systems from multivariate water quality time series. *Environmental Science & Technology* **2012**, *46*, 8212–8219.

(8) Yin, S.; Ding, S. X.; Xie, X.; Luo, H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics* **2014**, *61*, 6418–6428.

(9) Narasimhan, S.; Bhatt, N. Deconstructing principal component analysis using a data reconciliation perspective. *Computers & Chemical Engineering* **2015**, *77*, 74–84.

(10) Villez, K.; Habermacher, J. Shape Anomaly Detection for Process Monitoring of a Sequencing Batch Reactor. *Computers & Chemical Engineering* **2016**, *91*, 365–379.

(11) Zhang, J.; Hou, D.; Wang, K.; Huang, P.; Zhang, G.; LoÃąiciga, H. Real-time detection of organic contamination events in water distribution systems by principal components analysis of ultraviolet spectral data. *Environmental Science and Pollution Research* **2017**, *24*, 12882–12898.

(12) Smyth, P. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **2000**, *10*, 63–72.

(13) Xiao, Y.; Wang, H.; Xu, W. Hyperparameter selection for Gaussian process one-class classification. *IEEE transactions on neural networks and learning systems* **2015**, *26*, 2182–2187.

(14) Wang, S.; Liu, Q.; Zhu, E.; Porikli, F.; Yin, J. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. *Pattern Recognition* **2018**, *74*, 198–211.

(15) Ferré, L. Selection of components in principal component analysis: a comparison of methods. *Computational Statistics & Data Analysis* **1995**, *19*, 669–682.

(16) Valle, S.; Li, W.; Qin, S. J. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research 38*, 4389–4401.

(17) Minka, T. P. Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems 13 (NIPS 2000)* **2001**, 598–604.

(18) Bouveyron, C.; Celeux, G.; Girard, S. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters* **2011**, *32*, 1706–1713.

(19) Josse, J.; Husson, F. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis* **2012**, *56*, 1869–1879.

(20) Bro, R.; Kjeldahl, K.; Smilde, A. K.; Kiers, H. A. L. Cross-validation of component models: a critical look at current methods. *Analytical and Bioanalytical Chemistry* **2008**, *390*, 1241–1251.

(21) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning. Data Mining, Inference, and Prediction*; Springer, NY, USA, 2001; p 533.

(22) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence. 1995; pp 1137–1145.

(23) Saccenti, E.; Camacho, J. On the use of the observation-wise k-fold operation in PCA cross-validation. *Journal of Chemometrics* **2015**, *29*, 467–478.

(24) Arteaga, F.; Ferrer, A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics* **2002**, *16*, 408–418.

(25) Arteaga, F.; Ferrer, A. Framework for regression-based missing data imputation methods in on-line MSPC. *Journal of Chemometrics* **2005**, *19*, 439–447.

(26) Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **1978**, *20*, 397–405.

(27) Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects. *Chemometrics and Intelligent Laboratory Systems* **2014**, *131*, 37–50.

(28) Camacho, J.; Ferrer, A. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects. *Journal of Chemometrics* **2012**, *26*, 361–373.

(29) Gneiting, T.; Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **2007**, *102*, 359–378.

(30) Tipping, E.; Bishop, C. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1999**, *61*, 611–622.

(31) Saccenti, E.; Camacho, J. Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems* **2015**, *149*, 99–116.

(32) Peres-Neto, P. R.; Jackson, D. A.; Somers, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis* **2005**, *49*, 974–997.

(33) Wentzell, P.; Lohnes, M. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *J. Chemometr.* **1999**, *45*, 65–85.

(34) Hoefsloot, H. C.; Verouden, M. P.; Westerhuis, J. A.; Smilde, A. K. Maximum likelihood scaling (MALS). *Journal of Chemometrics* **2006**, *20*, 120–127.

(35) Narasimhan, S.; Shah, S. Model identification and error covariance matrix estimation from noisy data using PCA. *Control Engineering Practice* **2008**, *16*, 146–155.

(36) Wentzell, P.; Andrews, D.; Hamilton, D.; Faber, K.; Kowalski, B. Maximum likelihood principal component analysis. *J. Chemometr.* **1997**, *11*, 339–366.

(37) Schuermans, M.; Markovsky, I.; Wentzell, P.; Van Huffel, S. On the equivalence between total least squares and maximum likelihood PCA. *Anal. Chim. Acta* **2005**, *544*, 254–267.

(38) Boucher, M. A.; Perreault, L.; Anctil, F. Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. *Journal of Hydroinformatics* **2009**, *11*, 297–307.

(39) Spinelli, S.; Gonzalez, C.; Thomas, O. In *UV spectra library. UV-Visible Spectrophotometry for Water and Wastewater*; Thomas, O., Burgess, C., Eds.; Elsevier, Amsterdam, 2007; pp 267–357.

(40) Ma*(v)* sić, A.; Santos, A. T. L.; Etter, B.; Udert, K. M.; Villez, K. Estimation of nitrite in source-separated nitrified urine with UV spectrophotometry. *Water Research* **2015**, *85*, 244–254.

(41) Smialowski, P.; Frishman, D.; Kramer, S. *Pitfalls of supervised feature selection* **2009**, *26*, 440–443.

(42) De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. **2000**, *50*, 1–18.

(43) Brereton, R. G. The Mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics* **2015**, *29*, 143–145.

(44) Wu, N.; Zhang, J. Factor-analysis based anomaly detection and clustering. *Decision Support Systems* **2006**, *42*, 375–389.

(45) Yue, H. H.; Qin, S. J. Reconstruction-based fault identification using a combined index. *Industrial & Engineering Chemistry Research* **2001**, *40*, 4403–4414.

(46) Dilokthanakul, N.; Mediano, P. A.; Garnelo, M.; Lee, M. C.; Salimbeni, H.; Arulkumaran, K.; Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648* **2016**,

(47) Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* **2016**,

(48) Lawrence, N. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research* **2005**, *6*, 1783–1816.

(49) Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. *Psychometrika* **1967**, *32*, 443–482.

(50) Bonvin, D.; Rippin, D. W. T. Target factor analysis for the identification of stoichiometric models. *Chemical Engineering Science* **1990**, *45*, 3417–3426.

(51) Harmon, J. L.; Duboc, P.; Bonvin, D. Factor analytical modeling of biochemical data. *Computers & Chemical Engineering* **1995**, *19*, 1287–1300.

(52) Ge, Z.; Song, Z. Process monitoring based on independent component analysis - principal component analysis (ICA-PCA) and similarity factors. *Industrial & Engineering Chemistry Research* **2007**, *46*, 2054–2063.

(53) Reis, M. S.; Saraiva, P. M. Heteroscedastic latent variable modelling with applications to multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems* **2006**, 57–66.

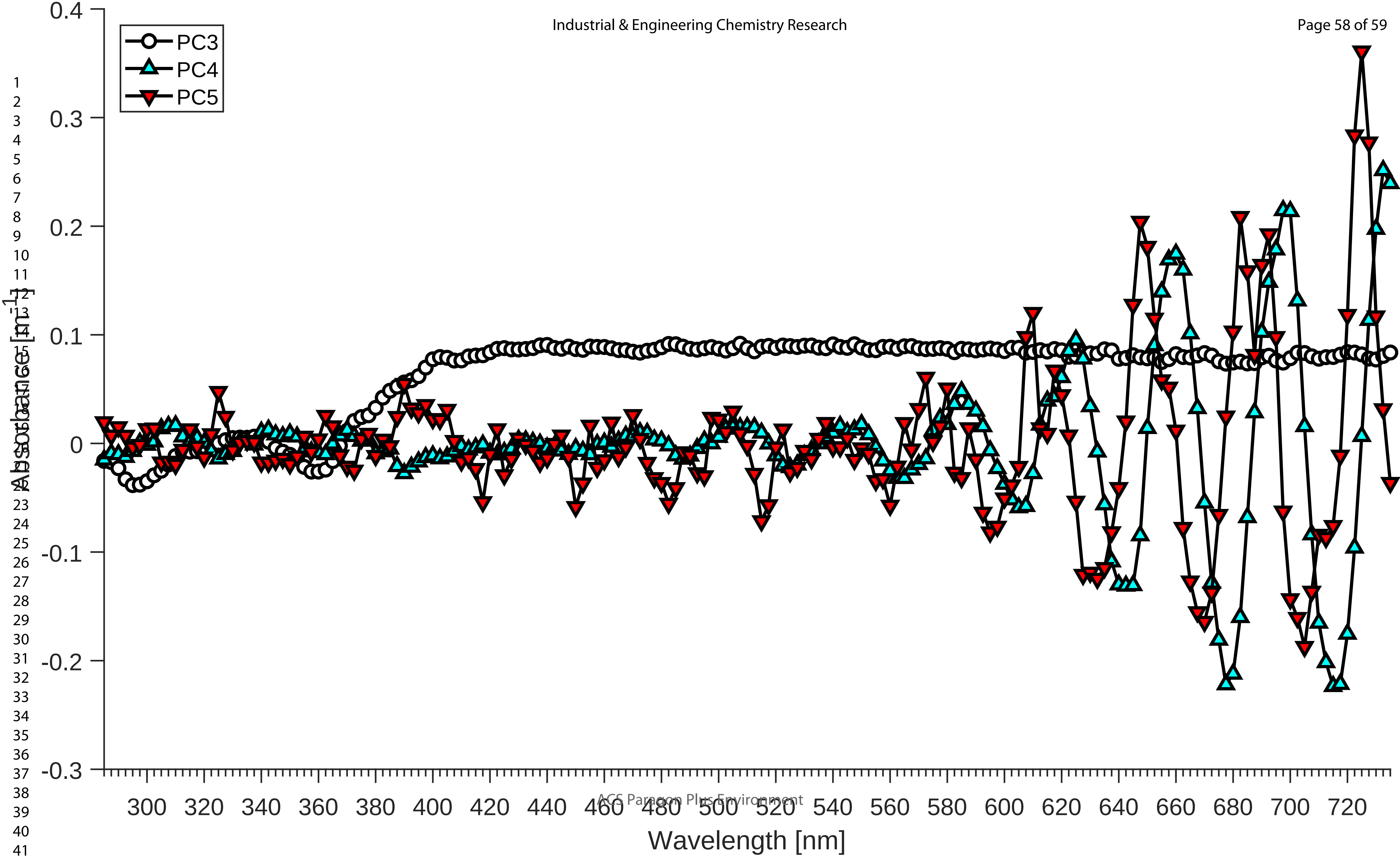
(54) Folch-Fortuny, A.; Arteaga, F.; Ferrer, A. Assessment of maximum likelihood PCA missing data imputation. *Journal of Chemometrics,* **2016**, *30*, 386–393.

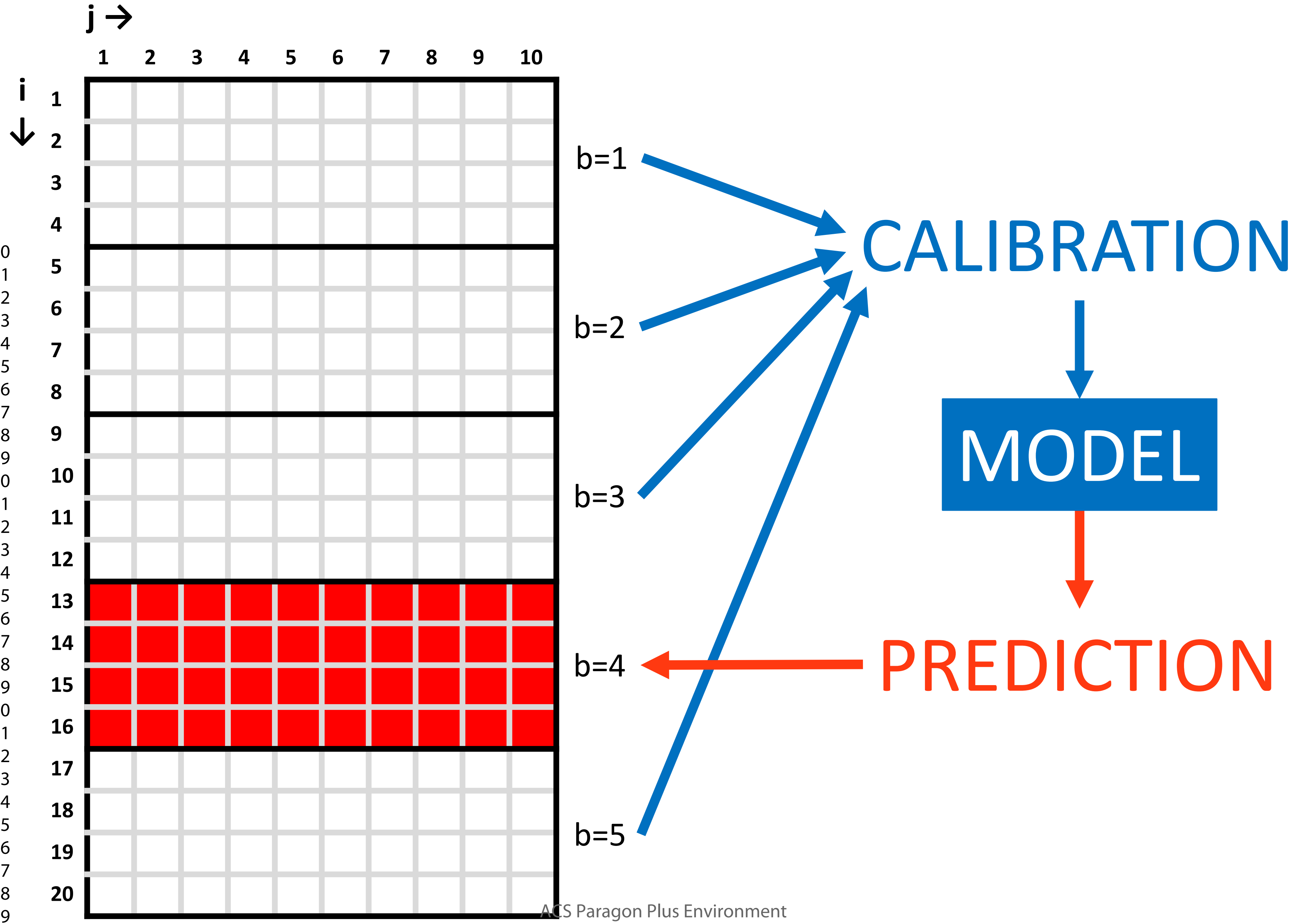
(55) Bakshi, B. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal* **1998**, *44*, 1596–1610.

(56) Rosén, C.; Lennox, J. A. Multivariate and multiscale monitoring of wastewater treatment operation. *Water Research* **Rosen2001**, *35*, 3402–3410.

(57) Lee, D. S.; Park, J. M.; Vanrolleghem, P. A. Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. *J. Biotechnol.* **2005**, *116*, 195–210.

(58) Baert, A.; Villez, K.; Steppe, K. Functional unfold principal component analysis for automatic plant-based stress detection in grapevine. *Functional Plant Biology* **2012**, *39*, 519–530.


PC3
PC4
PC5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Absorbance [m^{-1}]
300
320
340
360
380
400
420
440
460
480
500
520
540
560
580
600
620
640
660
680
700
720
Wavelength [nm]


j →
1
2
3
4
5
6
7
8
9
10
i ↓
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
b=1
b=2
b=3
b=4
b=5
CALIBRATION
MODEL
PREDICTION