Big data analytics using semi-supervised learning methods

Frumosu, Flavia Dalia; Kulahci, Murat

Published in:
Quality and Reliability Engineering International

Link to article, DOI:
10.1002/qre.2338

Publication date:
2018

Document Version
Peer reviewed version

Citation (APA):
Frumosu, F. D., & Kulahci, M. (2018). Big data analytics using semi-supervised learning methods. *Quality and Reliability Engineering International*, *34*(7), 1413-1423. https://doi.org/10.1002/qre.2338

Received: 20 December 2017 | Revised: 25 April 2018 | Accepted: 7 May 2018

DOI: 10.1002/qre.2338

SPECIAL ISSUE ARTICLE

WILEY

Big data analytics using semi-supervised learning methods

Flavia D. Frumosu[1] | Murat Kulahci[1,2]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

[2] Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

Correspondence
Flavia D. Frumosu, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark.
Email: fdal@dtu.dk

Abstract

The expanding availability of complex data structures requires development of new analysis methods for process understanding and monitoring. In manufacturing, this is primarily due to high-frequency and high-dimensional data available through automated data collection schemes and sensors. However, particularly for fast production rate situations, data on the quality characteristics of the process output tend to be scarcer than the available process data. There has been a considerable effort in incorporating latent structure–based methods in the context of complex data. The research question addressed in this paper is to make use of latent structure–based methods in the pursuit of better predictions using all available data including the process data for which there are no corresponding output measurements, ie, unlabeled data. Inspiration for the research question comes from an industrial setting where there is a need for prediction with extremely low tolerances. A semi-supervised principal component regression method is compared against benchmark latent structure–based methods, principal components regression, and partial least squares, on simulated and experimental data. In the analysis, we show the circumstances in which it becomes more advantageous to use the semi-supervised principal component regression over these competing methods.

KEYWORDS

dimension reduction, latent structure methods, multivariate data, production statistics

1 | INTRODUCTION

Since the start of industrialization, technological advances have led to paradigm shifts, which today are recognized as "industrial revolutions." The so-called first industrial revolution started in the field of mechanization through steam power, the second one continued with the intensive use of electricity, while the third one incorporated information technology and robots into the industries. Industry 4.0, the term for the fourth industrial revolution, was first coined in 2010 by the Federal Ministry of Education and Research in Germany and has become a high-tech strategy for Horizon 2020.[1] The vision of the fourth industrial revolution is that based on the advanced digitalization within industries, the combination of internet of things, "smart objects" all powered by sensors and internet networks, will result in another paradigm shift (Figure 1).

This will in fact result into modular and efficient manufacturing systems with the ultimate goal that products control their own manufacturing process.[1] Actually, this will require that manufacturing machines will know their own history and predict the future events making the manufacturing process smarter.[2] All these changes result in the abundance of data, usually referred as big data, which can potentially be used in the optimization of the manufacturing processes.

In data analysis, these advances have led to an environment with an abundant amount of production data, which is often characterized as multidimensional and

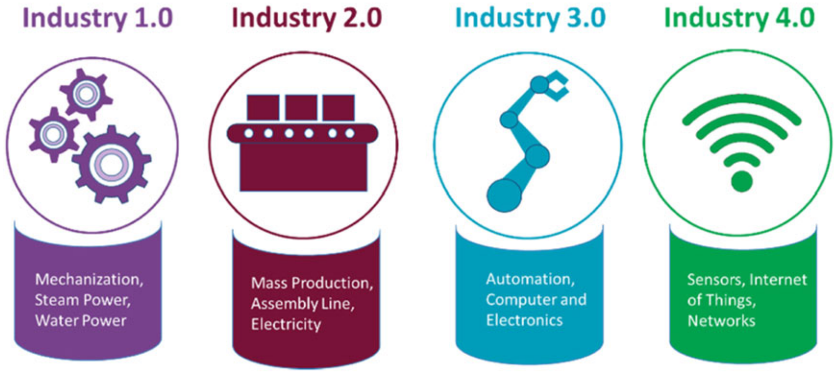

Industry 1.0
Mechanization, Steam Power, Water Power
Industry 2.0
Mass Production, Assembly Line, Electricity
Industry 3.0
Automation, Computer and Electronics
Industry 4.0
Sensors, Internet of Things, Networks

FIGURE 1 Industry revolutions through time [Colour figure can be viewed at wileyonlinelibrary.com]

highly correlated. This has been possible through automated data collection schemes and sensorics, which collected process data such as temperature and pressure at a great frequency. On the response side, eg, product quality characteristics, however, the progress has been slower. Particularly for processes that operate at high production rates, such output data are still scarce. This results in 2 types of process data: labeled (a.k.a. supervised) for which there is a corresponding output data and unlabeled (a.k.a. unsupervised) for which there is no output data. The research question in this work is the effective combination of these 2 types of data to improve prediction of the output. That is, the goal is to investigate ways in which the unlabeled data offer added value in predicting the output in the case of scarce labeled data.

As a tool for analysis, multiple regressions with high-dimensional input data have proven to be inefficient when the number of observations is very low and also when multicollinearity is present. Multicollinearities are frequently indicated by large correlations within subsets of the variables and can lead to large variances of the estimated regression coefficients, which make the regression estimates unstable and potentially misleading.[3] Therefore, multiple regression cannot be used in the above mentioned scenario. To overcome these issues, methods based on latent structures such as principal component regression (PCR) and partial least squares (PLS) regression have been traditionally used. These methods have been studied intensively in the field of chemometrics.[4-6] As indicated by Frank and Friedman,[6] PLS is often recommended over PCR even though both methods are documented in various articles with industrial applications, which state that both methods perform similar in terms of prediction.[7] The difference lays in the number of components as usually PLS needs fewer latent variables than PCR to obtain the same amount of prediction error.[7]

Dimensionality reduction does not necessarily ensure a better prediction when there are fewer observations for the response variable(s). This type of problem, namely, having more observations for the inputs than for the response variable(s), can be found in the literature under semi-supervised (SS) learning. This term was first introduced for classification problems where there were fewer labeled data (predictors/label pairs) than unlabeled data (predictors without responses). By incorporating unlabeled data into the supervised model or directly training the model from both labeled and unlabeled datasets, SS models are reported to perform better than the classical methods.[8-12] Traditional SS learning methods include self-training–based methods, co-training methods, probabilistic generative model–based methods, and graph-based methods.[11] In the literature, more focus has been put on classification problems rather than regression problems.

This paper is motivated by the pursuit of obtaining better predictions for production data by incorporating the unlabeled data into model training. The naïve approach is to discard the unlabeled data and make predictions solely based on the labeled data. With the help of SS methods, we show that the predictions can be improved by simultaneously dealing with dimensionality reduction and scarce observations for the response(s). The argument is that by introducing unlabeled data, we get a better understanding of the overall variation in the inputs (predictors), which ultimately leads to a better prediction.

2 | METHODS

In this section, a short introduction to PCR and PLS is given along with the SS-PCR method.

2.1 | Principal component regression

In multiple linear regression, the model is

$$\mathbf{y} = X\beta + \varepsilon, \tag{1}$$

where $\mathbf{y}$ is the response variable with n observations, the predictors matrix $\mathbf{X}$ is of size $n \times p$, while β is a column vector of p regression coefficients. The error terms are represented by ε, and they are usually assumed for inference purposes to be independently and identically distributed with a normal distribution with mean 0 and constant variance, σ^2.

In the case of high correlation among the predictors, the principal components of the predictors can be obtained from

$$\mathbf{Z} = \mathbf{XA}, \tag{2}$$

where $\mathbf{Z}$ is the matrix representing the principal components, a.k.a. scores, and $\mathbf{A}$ is the loading matrix and has a dimension of (p x p) whose kth column is the kth eigenvector of $\mathbf{X}'\mathbf{X}$ corresponding to its kth largest eigenvalue. The principal components can be seen as the linear combinations of the original variables obtained in such a way that the first linear combination explains the most variation in the data, the second linear combination explains the second largest variation in the data and uncorrelated with the first linear combination, and so on.

Because of the orthonormality of $\mathbf{A}$, ie, $\mathbf{AA}' = \mathbf{I}$ as eigenvectors are orthogonal and of unit length, $\mathrm{X}\beta$ can be rewritten as $\mathbf{XAA}'\beta = \mathrm{Z}\gamma$, where $\gamma = \mathbf{A}'\beta$. The PCR equation can thus be written as

$$\mathbf{y} = \mathrm{Z}\gamma + \varepsilon, \tag{3}$$

which has the predictor variables replaced by their PC's in the regression model. Furthermore, it is conventional in the PCR as well as other chemometric methods to standardize the predictor variables such that $\mathbf{X}'\mathbf{X}$ is proportional to the correlation matrix of the predictor variables. This convention is also followed in this paper.

From an optimization perspective, PCR seeks directions that have high variance, and mathematically, the kth principal component direction a_k[13] is obtained from

$$\begin{gathered}\max_\alpha \mathbf{Var}(\mathrm{X}\alpha)\\ \textbf{subject to } \|\alpha\| = \mathbf{1}, \alpha^{\mathbf{T}}\mathbf{S}\mathbf{a_j} = 0, \mathbf{j}{=}\mathbf{1},\ldots,\mathbf{k}\text{-}\mathbf{1},\end{gathered} \tag{4}$$

where $\mathbf{S}$ is defined as the sample covariance matrix of $\mathbf{X}$. To ensure that $\mathbf{z_k} = \mathrm{X}\alpha$ is uncorrelated with the previous linear combinations $\mathbf{z_j} = \mathbf{X}\boldsymbol{a_j}$, the condition $\alpha^{\mathbf{T}}\mathbf{Sa_j} = \mathbf{0}$ needs to be satisfied. The number of components chosen for $\mathbf{Z}$ can be determined by using different methods. This will be considered in section 2.3.2. For further details we refer to Chapter 6 of Jolliffe.[3]

2.2 | Partial least squares

As in the case of PCR, PLS also builds a set of linear combinations of the inputs for regression.[13] However, unlike PCR, it uses the correlation between the response(s) and the predictors in finding the linear of combinations of the inputs that not only explain the most variation in the input variables but also have the most predictive power to explain the response(s). This can be formulated as an optimization problem for univariate y, where the kth PLS direction $\widehat{\varphi}_{\boldsymbol{k}}$ solves[13]:

$$\begin{gathered}\max_\alpha \mathbf{Corr^2}(\mathbf{y}, \mathrm{X}\alpha)\mathbf{Var}(\mathrm{X}\alpha)\\ \textbf{subject to } \|\alpha\| = \mathbf{1}, \alpha^{\mathbf{T}}\mathbf{S}\widehat{\varphi}_{\mathbf{j}} = 0, \mathbf{j}{=}\mathbf{1},\ldots,\mathbf{k}\text{-}\mathbf{1}.\end{gathered} \tag{5}$$

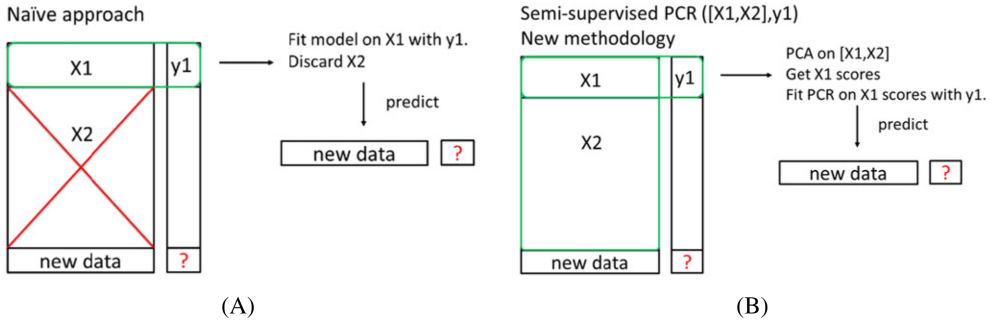
The traditional PLS regression algorithm is based on the nonlinear iterative partial least squares.[4,14,15] The idea in nonlinear iterative partial least squares is to compute the components in a partial fashion, ie, one at a time, until all the variance in the data is explained. The algorithm is iterative, and in each step, residuals are computed from the information explained by the last component subtracted from $\mathbf{X}$ and $\mathbf{y}$.[15] Other implementations for PLS are available in the literature such as SIMPLS[16] or the kernel algorithm for PLS.[15] For further details on PLS, we refer to Wold et al[4] and Höskuldsson.[17]

2.3 | Semi-supervised principal component regression

2.3.1 | Methodology

In the context of manufacturing industry, process data from sensors tend to be abundant while the responses, ie, product quality measures, are very scarce due to sampling and inspection costs. The underlying idea is to make use of unlabeled data, which may otherwise be ignored for prediction purposes. By introducing unlabeled data, we will get a better understanding of the overall predictors' variation that will ultimately lead to a better prediction. One straightforward approach is to modify PCR such that all available data are used for computing the loadings rather than using just the labeled data. This idea has been proposed before when this approach is used to calibrate the model with all available data and compared with standard PCR.[29,30] We follow a slightly modified approach where the goal is to predict the outcome of the new data that was not used in model building. Furthermore, we also provide a comparison of this approach with the commonly used supervised learning method, PLS, and show the circumstances in which PLS is outperformed by the modified PCR approach. For a quick overview of the naïve approach versus the proposed methodology used for prediction, see Figure 2.

One common particularity of industrial systems is that because of the high-frequency nature of the data, the predictors' data collected in time are serially dependent (autocorrelated). This phenomenon can lead to various problems in modelling.[18-20] In the case of labeled data, autocorrelation should not pose a problem since



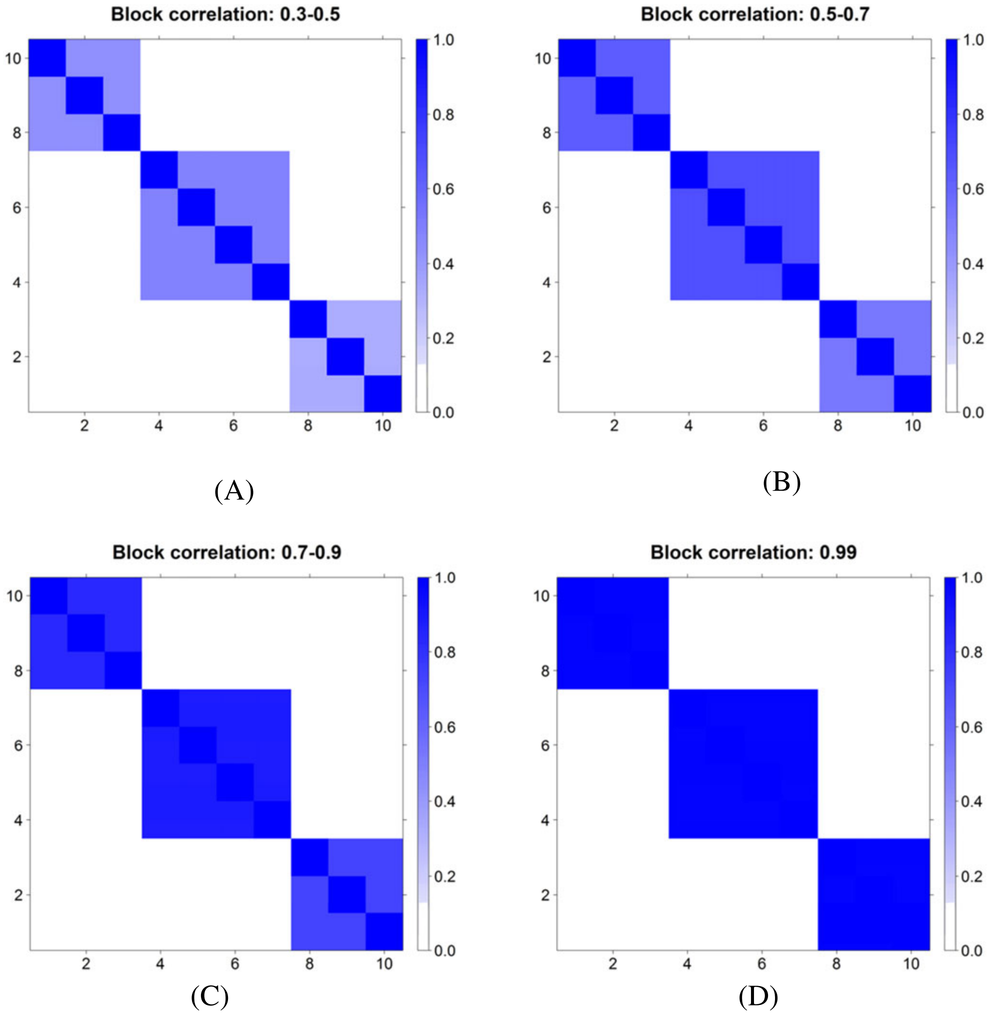
(A)

(B)

FIGURE 2 Comparison between the naïve approach and the proposed methodology semi-supervised principal component regression (SS-PCR) [Colour figure can be viewed at wileyonlinelibrary.com]

the data are collected at distant time intervals, usually because of a high cost of sampling and inspection. Autocorrelation can be a problem in the case of unlabeled data as the data are sampled at a high frequency. Because of the abundance of unlabeled data, a skipping strategy or any other strategy[21,22] can be used to avoid this issue. In our simulations, we did not impose any serial dependence, and hence, our results reflect that particular case.

The mathematical formulation for both SS-PCR methods coincides and is derived below. Let the data for



(A)

(B)

(C)

(D)

FIGURE 3 First simulation study (10 predictors with 3 latent components): Correlation structures among input variables. Each block contains the same random value generated from the block correlation interval. In the case of 0.99, all the blocks have the same value namely 0.99 [Colour figure can be viewed at wileyonlinelibrary.com]

predictors $\mathbf{X}$ ($n \times p$) be split into $\mathbf{X_1}$ ($n_1 \times p$) and $\mathbf{X_2}$ ($n_2 \times p$), hence $n_1 + n_2 = n$. The response $\mathbf{y}$ is of dimension $n_1 \times 1$ and together with $\mathbf{X_1}$ defines the labeled data set, while $\mathbf{X_2}$ is further defined as the unlabeled data set.

Given the relationships from (1) and (2) and because of the orthonormality of $\mathbf{A}$ (the loading matrix $\mathbf{A}$ is computed from the entire predictors data, $\mathbf{X}$), $\mathbf{X_1}\beta$ can be rewritten as $\mathbf{X_1 A A}'\beta = \mathbf{Z_1}\gamma$, where $\gamma = \mathbf{A}'\beta$. The SS-PCR equation can thus be written as

$$\mathbf{y} = \mathbf{Z_1}\gamma + \varepsilon, \tag{6}$$

which has the predictor variables $\mathbf{X_1}$ replaced by the $\mathbf{Z_1}$; the $\mathbf{X_1}$ scores using the loadings from the entire $\mathbf{X}$ in the regression model. It is assumed that both labeled and unlabeled data come from the same distribution. It has been proven through simulations that a drift in the unlabeled data can result in an increased prediction error compared with a model based solely on labeled data.[23] Since there is only a methodology change between the existing and proposed SS-PCR, both are referred under the name SS-PCR. Semi-supervised principal component regression was treated from both a linear and nonlinear perspective in other published work.[11,12] The methods proposed in the papers written by Ge Z et al[11,12] consider the SS-PCR from a probabilistic perspective using a generative model structure. In this approach, the SS-PCR model is built from a data projection perspective. The advantage of this method is the simplicity in which it can be easily applied in industrial cases. Also, no additional coding is needed beyond the standard PCR coding.

2.3.2 | The number of SS-PCR components

An important research question is to determine how the number of latent components is selected for the SS-PCR method. For PCR and PLS, there are several suggestions in the literature on how to deal with estimating the number of latent components, eg, Chapter 6 of Jolliffe[3] and Wold et al.[4] One very common approach as stated in Chun and Keleş[24] is to use cross validation. Coster et al[25] introduced cross validation to find the number of dimensions that minimized the prediction error. All of these techniques can be combined under the measure, root-mean-square error of prediction (RMSEP) performed with cross validation. As indicated in Mevik and Cederkvist,[26] leave-one-out cross validation is preferable when one can afford the computational burden, although adjusted 5-fold or 10-fold cross validation are also viable candidates. For illustration purposes, RMSEP is presented with a K-fold cross validation where the data are randomly divided into $\mathbf{K}$ segments of approximately equal size, $\mathbf{L_k}$,[27]

$$\mathbf{RMSEP_{cv.K}} = \sqrt{\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \sum_{\mathbf{i \in L_k}} \left(\mathbf{y_i} - \widehat{\mathbf{y}}_\mathbf{i}\right)^\mathbf{2}} \tag{7}$$

where $\widehat{\mathbf{y}}_\mathbf{i}$ is the vector of predicted values for the sample $\mathbf{i}$ and $\mathbf{n}$ is the number of total samples. This statistic is also

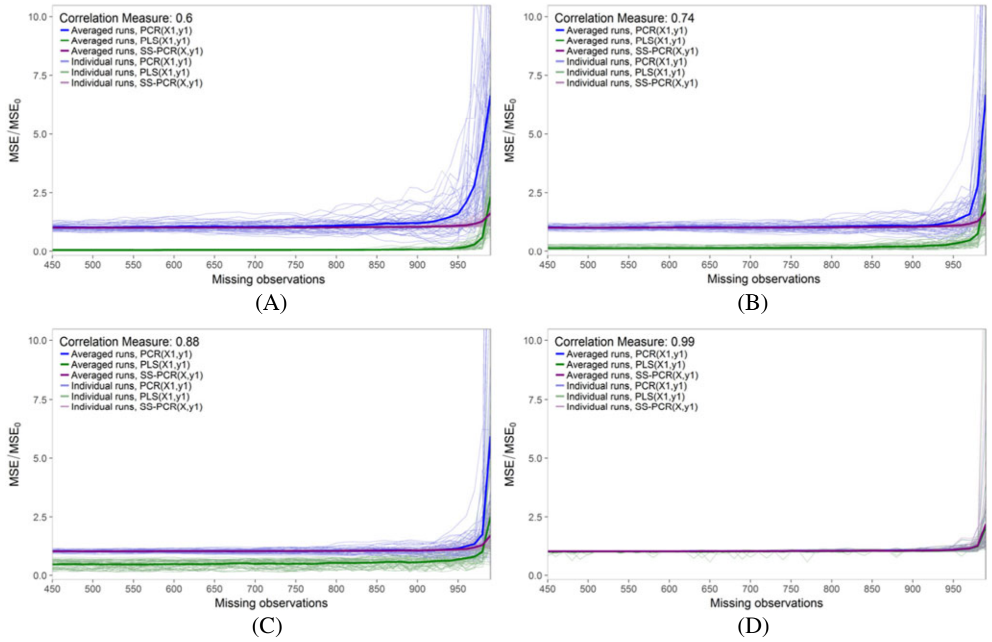


FIGURE 4 First simulation study (10 predictors with 3 latent components): MSE/MSE$_0$ curves averaged over 50 runs (solid green, blue, and purple lines) for supervised partial least squares (PLS), principal component regression (PCR), and semi-supervised PCR. Missing observations are created in steps of 10 from the original 1000 observations [Colour figure can be viewed at wileyonlinelibrary.com]

denoted as root-mean-squared error of cross validation.[26] The number of latent components is chosen for which RMSEP is minimized. In this way, the optimal number of latent components is based on the predictive power. In our case, we compute the RMSEP on the $\mathbf{X_1}$ scores computed by using the loadings from the entire input data, **X**.

2.3.3 | Correlation measure

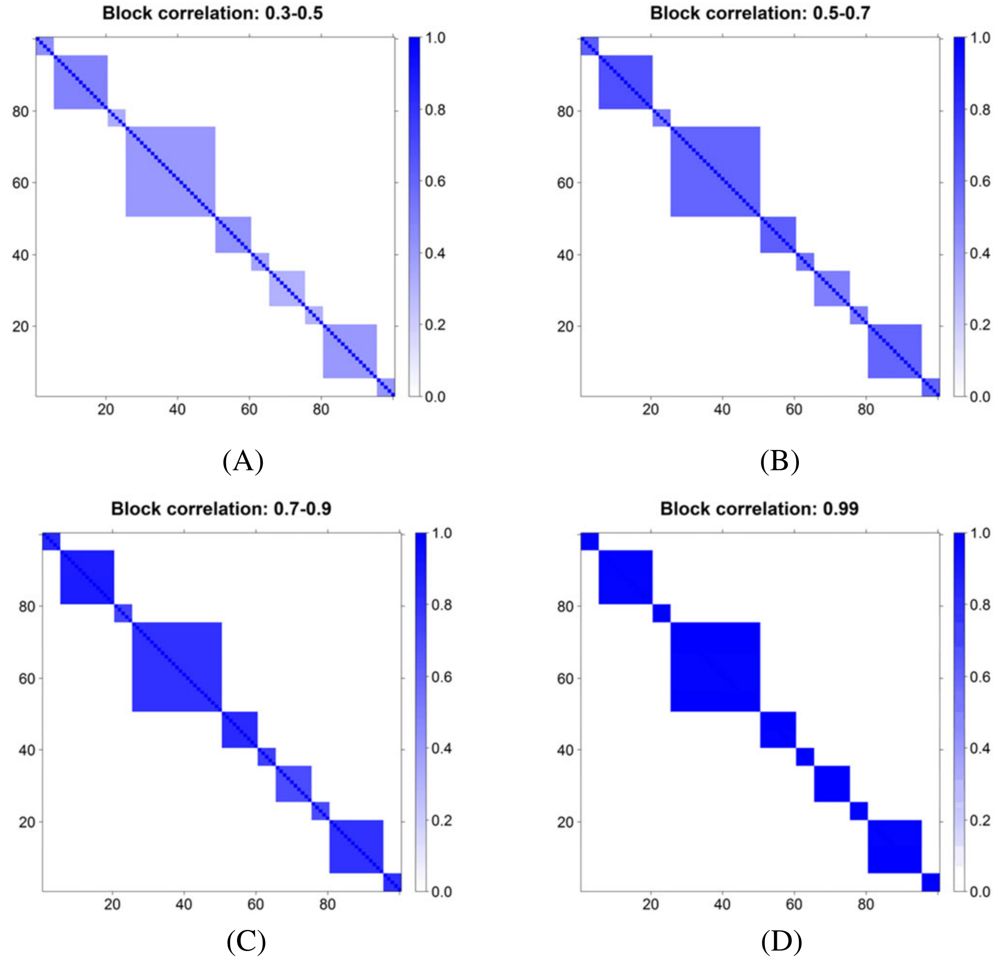
It is conjectured that the contribution of unsupervised data will be related to the amount of supervised data as well as the underlying correlation structure of the input data. Hence, it is of high relevance in this context to define a measure for the correlation among input variables. Given that the number of latent components **h** is retained using the RMSEP procedure as discussed in the previous section, we define the correlation measure (CM) as

$$\mathbf{CM} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{h}} \lambda_{\mathbf{i}}}{\sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_{\mathbf{i}}} \tag{8}$$

where **p** is the initial dimension and λs are the eigenvalues of the correlation matrix of **X**. The CM is between **h/p** (no correlation among variables) and **1**.

3 | SIMULATION STUDY

Three simulation studies have been conducted to compare the SS-PCR method with the benchmark methods PCR and PLS. All studies are based on 1000 observations, and each simulation study contains different number of predictors and latent components. For a fair comparison, the latent components have been artificially imposed by building blocks in the structure of the correlation matrix of the predictors, **X** which is normally distributed with zero mean. Each block contains the same random number generated from an interval as it can be observed in Figures 3, 5, and 7. Though not visible in these figures, 5, 25, and 50 values between 0 and 0.05 for each study are also spread randomly for the "off-diagonal" entries with the only exception of the case of high correlation



(A)

(B)

(C)

(D)

FIGURE 5 Second simulation study (100 predictors with 10 latent components): Correlation structures among input variables. Each block contains the same random value generated from the block correlation interval. In the case of 0.99, all the blocks have the same value namely 0.99 [Colour figure can be viewed at wileyonlinelibrary.com]

(0.99). Since the "true" latent structure is assumed to be fixed and known, the methodology proposed in section 2.3.2 is not followed in the simulation studies. However, in an industrial setting, this is highly recommended if no prior information regarding latent structures is available.

The first 2 simulation studies are based on 10 predictors and 3 latent components, and 100 predictors and 10 latent components, respectively. For the third simulation, the data structure is inspired from an industrial injection molding process for which process data using molding machine sensors are collected at a rapid rate, while the quality measures are very scarce. To simulate this scenario, data are generated with 300 process variables and 20 latent components.

For all 3 simulation studies, the **y** are generated using

$$\mathbf{y} = X\beta + \varepsilon,$$

where β and ε are distributed uniformly and normally, ie, $\beta \sim \mathcal{U}(\mathbf{0}, \mathbf{1})$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{0}.\mathbf{1}^2)$, respectively. At first, **y** contains 1000 observations. For creating the SS scenario, **y** is then systematically reduced by removing observations in increments of 10, which is further denoted as $\mathbf{y_1}$. The removed observations are later used for computing the prediction mean squared error (MSE), which is then scaled with MSE_0 computed from the PCR model on the entire **X** and **y**. All the data have been zero centered but not scaled since the data have the same order of magnitude.

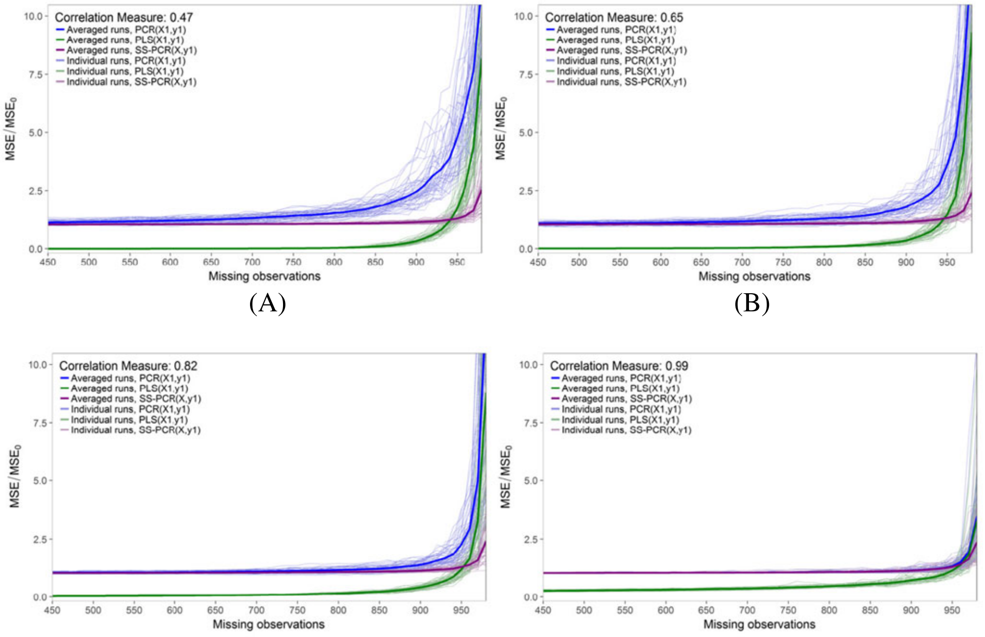
The MSE/MSE_0 curves from applying supervised PLS and PCR on $\mathbf{X_1}$ and $\mathbf{y_1}$, SS-PCR on **X** and $\mathbf{y_1}$ are presented in Figures 4, 5, and 6 with predictions averaged over 50 runs.

3.1 | First simulation study

As specified above, the correlation structure for the first simulation study with 10 predictors and 3 latent components is illustrated in Figure 3.

The comparison between SS-PCR(**X**, $\mathbf{y_1}$), PCR($\mathbf{X_1}$, $\mathbf{y_1}$), and PLS($\mathbf{X_1}$, $\mathbf{y_1}$) is illustrated in Figure 4.

As it can be observed in Figure 4, PLS outperforms SS-PCR and PCR in 3 of the 4 cases by having the lowest MSE/MSE_0 ratio. When correlation is very high among predictors as in Figure 4D, SS-PCR, PCR, and PLS perform similarly. As the results indicate, SS-PCR starts to take advantage of the added unlabeled data in the case when the response is scarce and more importantly when PLS coefficient estimates start to become unstable. For this case, as the CM starts to increase, the PLS coefficient estimates start to become unstable, and this is reflected in the prediction.



(A) (B)

(C) (D)

FIGURE 6 Second simulation study (100 predictors with 10 latent components): MSE/MSE_0curves averaged over 50 runs (solid green, blue, and purple lines) for supervised partial least squares (PLS), principal component regression (PCR), and semi-supervised PCR. Missing observations are created in steps of 10 from the original 1000 observations [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Second simulation study

The correlation structure for this data set with 100 predictors and 10 latent components is illustrated in Figure 5.

The comparison between SS-PCR($\mathbf{X}$, $\mathbf{y_1}$), PCR($\mathbf{X_1}$, $\mathbf{y_1}$), and PLS($\mathbf{X_1}$, $\mathbf{y_1}$) is illustrated in Figure 6.

As opposed to the results obtained for the first study, SS-PCR starts making use of the available unlabeled data earlier for prediction when the response is scarce. This result is valid for all 4 cases; however, the difference lies in the CM, which dictates that a better performance of SS-PCR is obtained when the CM is lower as opposed to the first simulation. What is interesting to note is that even though the blocks of the correlation matrix contain the same intervals for randomly generating the correlations, the CM is very different. This result is consistent with Jolliffe's statement,[3] which indicates that as the number of dimensions increases, a fixed number of retained latent components will explain less variation in the data.

3.3 | Third simulation study

The correlation structure for this data set with 300 predictors and 20 latent components is illustrated in Figure 7.

The comparison between SS-PCR($\mathbf{X}$, $\mathbf{y_1}$), PCR($\mathbf{X_1}$, $\mathbf{y_1}$), and PLS($\mathbf{X_1}$, $\mathbf{y_1}$) is illustrated in Figure 8.

As it can be depicted from Figure 8, in all 4 cases, PLS performs remarkably well compared with PCR. However, when the size of unsupervised data gets large, SS-PCR starts to outperform PLS. As stated before, what is interesting to observe is that SS-PCR outperforms PLS depending on the CM present in the original $\mathbf{X}$. As the CM gets larger, it is more difficult for SS-PCR to outperform PLS as observed in the second study. This is because of the fact that with high correlation, the estimation of the covariance matrix of $\mathbf{X}$ does not require as many observations as in the case of low correlation among the inputs. In that sense, the contribution of the unsupervised data gets attenuated with high correlation among inputs.

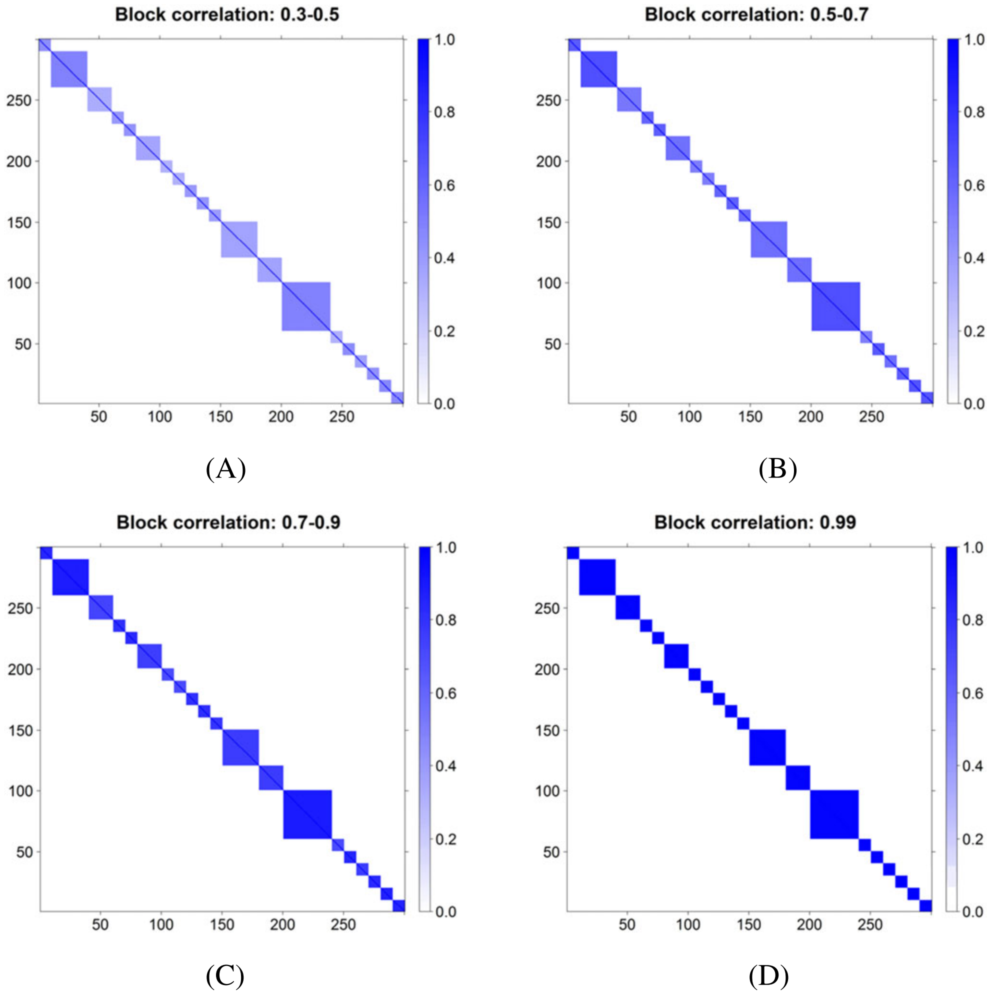


FIGURE 7 Third simulation study (300 predictors with 20 latent components): Correlation structures among input variables. Each block contains the same random value generated from the block correlation interval. In the case of 0.99, all the blocks have the same value namely 0.99 [Colour figure can be viewed at wileyonlinelibrary.com]

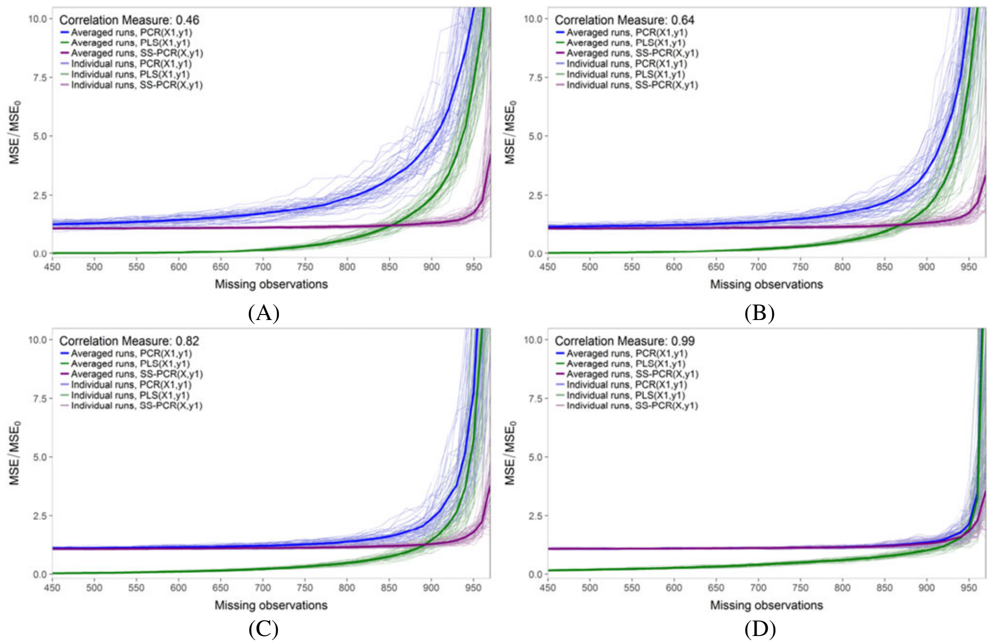


FIGURE 8 Third simulation study (300 predictors with 20 latent components): MSE/MSE$_0$curves averaged over 50 runs (solid green, blue and purple lines) for supervised partial least squares (PLS), principal component regression (PCR), and semi-supervised PCR. Missing observations are created in steps of 10 from the original 1000 observations [Colour figure can be viewed at wileyonlinelibrary.com]

4 | EXPERIMENTAL DATA

Most of the comparisons between PCR and PLS involve chemometric applications, and therefore, we use a near-infrared (NIR) spectroscopy benchmark data set for applying the SS-PCR methodology.

The data set consists of NIR spectra as predictors and octane numbers as responses of 60 gasoline samples. The NIR spectra were measured by using diffuse reflectance as log(1/R) from 900 to 1700 nm in 2 nm intervals, giving 401 wavelengths (Figures 9 and 10).[28]

To mimic the SS scenario as before, y is systematically reduced by removing observations in increments of 1. The number of labeled observations spans between 6 and 57. The data have been mean centered when applying the PCR, PLS, and SS-PCR methods.

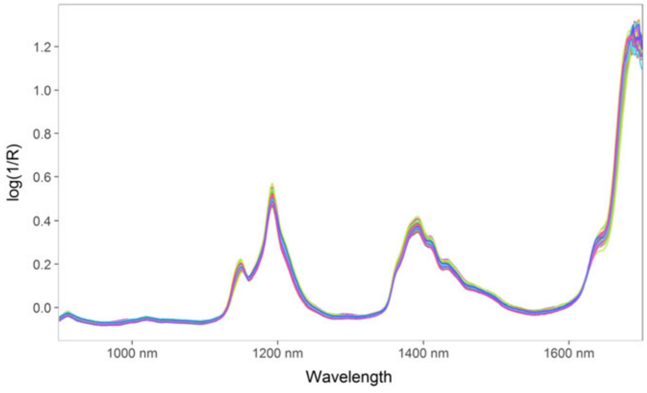


FIGURE 9 Gasoline near-infrared spectra [Colour figure can be viewed at wileyonlinelibrary.com]

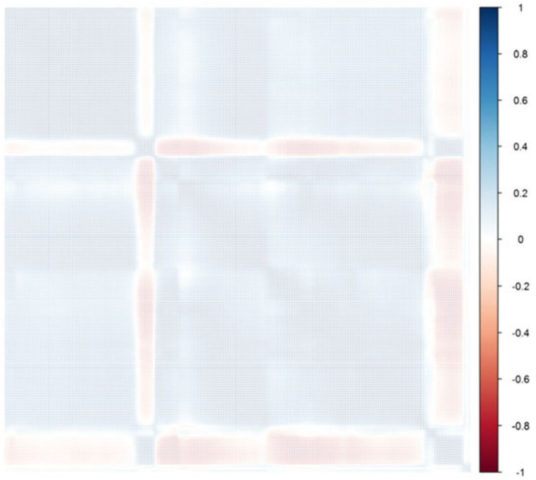
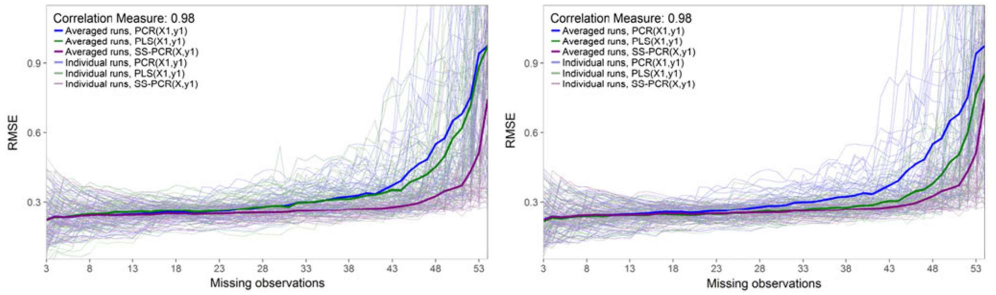


FIGURE 10 Correlation matrix of the gasoline near-infrared spectra [Colour figure can be viewed at wileyonlinelibrary.com]

For choosing the number of components, since the data set is small, RMSEP along with the leave-one-out cross validation have been used. For PLS, 3 or 4 components seem to be optimal, while for PCR and SS-PCR, 4 components are used. The CM for this data set is 0.98.

The comparison between SS-PCR, PCR, and PLS are illustrated in Figure 11. For a fair comparison, the original data have been sampled without replacement 50 times. In other words, each individual run consists of



(A)
PLS – 3 components
PCR, SS-PCR – 4 components

(B)
PLS – 4 components
PCR, SS-PCR – 4 components

FIGURE 11 Root-mean-square error curves averaged over 50 randomly sampled labeled and unlabeled data sets (solid green, blue, and purple lines) for supervised partial least squares (PLS), principal component regression (PCR), and semi-supervised PCR. Missing observations are created in steps of 1 [Colour figure can be viewed at wileyonlinelibrary.com]

randomly sampled labeled and unlabeled data sets from the original data set.

Semi-supervised principal component regression method performs better than PCR and PLS in the case of scarce outputs. When the number of components is 3 for PLS, SS-PCR starts to outperform PLS already after 20 missing observations, whereas for 4 components, SS-PCR outperforms PLS after 30 missing observations. This result confirms once more that unlabeled data can provide information to achieve a better prediction for scarce outputs scenarios.

5 | FINAL DISCUSSION

With the new data collection schemes in automated manufacturing, abundant production data are widely available, and consequently, benchmark data analysis methods need to be updated to take care of the newly created challenges. This paper investigates a SS-PCR method,[29,30] which uses both labeled and unlabeled data for prediction. Four test cases where PLS, PCR, and SS-PCR are compared have been performed on simulated data of varying dimensionality. A comparison between these 3 methods is also performed on a NIR spectroscopy benchmark data set.

As indicated already in the literature, PCR and PLS performs well in supervised learning applications. However, as we show in this paper, PCR and PLS are negatively affected when the response data are scarce or the CM is low for high-dimensional predictor data. In many current industrial settings where input data are abundant but the response data are extremely scarce, SS methods are recommended. By using the entire input data including both labeled and unlabeled data, a better representation of the covariance of the input variables is obtained. In addition, when the correlation among the input variables is relatively low yet still requiring dimension deduction methods in regression, the addition of unlabeled data improves the estimation of the overall correlation of the input variables. On the other hand for low-dimensional data, SS-PCR seems to outperform the benchmark methods when the response data are really scarce. However, since the implementation of the proposed methodology is relatively simple and does not require any additional software or coding compared with standard PCR, the use of SS-PCR for prediction purposes is highly recommended when abundant unlabeled process data is available.

ACKNOWLEDGEMENTS

The authors would like to thank MADE SPIR—Strategic Platform for Innovation and Research, Denmark, for making this work possible.

REFERENCES

1. Lasi H, Kemper H-G, Fettke P, Feld T, Hoffmann M. Industry 4.0, *Business and Information Systems Engineering*. 2014;6(4): 239-242. https://doi.org/10.1007/s12599-014-0334-4

2. Gilchrist A. Industry 4.0, Apress 2016; 195–215. https://doi.org/10.1007/978-1-4842-2047-4

3. Jolliffe IT. Principal component analysis, Springer Series in Statistics, 2002; 113, 167–198. ISBN-13: 978–0387954424

4. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intel Lab Syst*. 2001;58(2):109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

5. Naes T, Martens H. Comparison of prediction methods for multicollinear data. *Comm Stat Simulat Comput*. 1985;14(3): 545-576. https://doi.org/10.1080/03610918508812458

6. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35:109. https://doi.org/10.2307/1269656

7. Wentzell PD, Montoto LV. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemom Intel Lab Syst*. 2003;65(2):257-279. https://doi.org/10.1016/S0169-7439(02)00138-7

8. Zhu W. Semi-supervised learning literature survey, Preprint. https://doi.org/10.1.1.103.1693

9. Pise NN, Kulkarni P. A survey of semi-supervised learning methods, *Proceedings - 2008 International Conference on Computational Intelligence and Security*. 2008;2:30-34. https://doi.org/10.1109/CIS.2008.20477

10. Chapelle O, Schölkopf B, Zien A. *Semi-supervised Learning*. MIT Press; 2006. ISBN:9780262033589.

11. Ge Z, Song Z. Semi-supervised bayesian method for soft sensor modeling with unlabeled data samples. *Aiche J*. 2011;57(8): 2109-2119. https://doi.org/10.1002/aic.12422

12. Ge Z, Huang B, Song Z. Nonlinear semi-supervised principal component regression for soft sensor modeling and its mixture form. *J Chemometr*. 2014;28(11):793-804. https://doi.org/10.1002/cem.2638

13. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning* , *Springer Series in Statistics* 2008; ISBN: 9780387848570

14. Siong Ng K. A simple explanation of partial least squares, Preprint. https://doi.org/10.1.1.352.4447

15. Lindgren F, Geladi P, Wold S. The kernel algorithm for PLS. *J Chemometr*. 1993;7(1):45-59. https://doi.org/10.1002/cem.1180070104

16. De Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemom Intel Lab Syst*. 1993;18(3):251-263. https://doi.org/10.1016/0169-7439(93)85002-X

17. Höskuldsson A. PLS regression methods. *J Chemometr*. 1988;2(3):211-228. https://doi.org/10.1002/cem.1180020306

18. Rato T, Reis M, Schmitt E, Hubert M, De Ketelaere B. A systematic comparison of PCA-based statistical process monitoring methods for high-dimensional, time-dependent processes. *Aiche J*. 2016;62(5):1478-1493. https://doi.org/10.1002/aic.15062

19. Vanhatalo E, Kulahci M. The effect of autocorrelation on the Hotelling T-2 control chart. *Qual Reliab Eng Int*. 2015;31(8): 1779-1796. https://doi.org/10.1002/qre.1717

20. Vanhatalo E, Kulahci M, Bergquist B. On the structure of dynamic principal component analysis used in statistical process monitoring. *Chemom Intel Lab Syst*. 2017;167:1-11. https://doi.org/10.1016/j.chemolab.2017.05.016

21. Runger GC, Willemain TR. Batch-means control charts for autocorrelated data. *Iie Trans*. 1996;28(6):483-487. https://doi.org/10.1080/07408179608966295

22. Vanmann K, Kulahci M. A model-free approach to eliminate autocorrelation when testing for process capability. *Qual Reliab Eng Int*. 2008;24(2):213-228. https://doi.org/10.1002/qre.887

23. Gujral P, Amrhein M, Ergon R, Wise BM, Bonvin D. On multivariate calibration with unlabeled data. *J Chemometr*. 2011;25:456-465. https://doi.org/10.1002/cem.1389

24. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J Roy Stat Soc B Stat Meth*. 2010;72:3-25. https://doi.org/10.1111/j.1467-9868.2009.00723.x

25. Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet Sel Evol*. 2010;42(1):9. https://doi.org/10.1186/1297-9686-42-9

26. Mevik BH, Cederkvist HR. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J Chemometr*. 2004;18(9):422-429. https://doi.org/10.1002/cem.887

27. Davison AC, Hinkley DV. Bootstrap methods and their application, Cambridge University Press 1997; ISBN: 0521573912

28. Kalivas JH. Two data sets of near infrared spectra. *Chemom Intel Lab Syst*. 1997;37(2):255-225. https://doi.org/10.1016/S0169-7439(97)00038-5

29. Thomas EV. Incorporating auxiliary predictor variation in principal component regression models. *J Chemometr*. 1995;9(6):471-481. https://doi.org/10.1002/cem.1180090605

30. Isaksson T, Naes T. Selection of samples for calibration in near-infrared spectroscopy. II. Selection based on spectral measurements. *Appl Spectrosc*. 1990;44(7):1152-1158. https://doi.org/10.1366/0003702904086533

Flavia Dalia Frumosu holds an MSc in dynamic modeling from the Technical University of Denmark where she is currently a PhD student. Her current research lies in the field of big data analysis with applications to manufacturing industry. She also has expertise in the domain of risk and safety where she worked as a consultant for 3 years in the oil and gas industry.

Murat Kulahci is a Professor in the Department of Business Administration, Technology and Social Sciences at Luleå University of Technology in Sweden and an Associate Professor in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark. His research focuses on design of physical and computer experiments, statistical process monitoring, time series analysis and forecasting, and financial engineering. He has presented his work in international conferences and published over 80 articles in archival journals. He is the coauthor of 2 books on time series analysis and forecasting.

How to cite this article: Frumosu FD, Kulahci M. Big data analytics using semi-supervised learning methods. *Qual Reliab Engng Int*. 2018;1–11. https://doi.org/10.1002/qre.2338