Important Variable Selection in Partial Least Squares for Industrial Process Understanding and Control

An Academic Essay in Chemometrics

Nelson Lee Afanador
Important Variable Selection in Partial Least Squares for Industrial Process Understanding and Control

Proefschrift

ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 22 september 2014
om 12.30 uur precies

door

Nelson Lee Afanador

geboren op 10 januari 1972
te Philadelphia, Pennsylvania, USA
Promotor: Prof. dr. Lutgarde Buydens
Copromotor: Dr. Thanh Tran (Merck Sharp & Dohme, Oss)
Manuscriptcommissie:
  o prof. dr. Tom Heskes
  o prof. dr. Jakob de Vlieg (Bayer, Mijdrecht)
  o prof. dr. Alberto Ferrer (Universitat Politècnica de València, Spanje)
Important Variable Selection in Partial Least Squares for Industrial Process Understanding and Control

Doctoral thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. S.C.J.J. Kortmann,
according to the decision of the Council of Deans
to be defended in public on Monday, September 22, 2014
at 12.30 hours
by
Nelson Lee Afanador
Born on January 10, 1972
in Philadelphia (Pennsylvania, USA)
Acknowledgements

I owe my deepest gratitude to Dr. Thanh Tran, Prof. dr. Lutgarde Buydens, and Lori Pfahler who made this research and thesis possible, and supported me in every possible way. I also would like to thank all of my colleagues at the Center for Mathematical Sciences, Merck Sharp & Dohme for constructive discussions. I am also very grateful to those who have collaborated with me, in particular Dr. Lionel Blanchet.

For her constant encouragement and support in my education since childhood, my mother deserves special thanks. Last, but certainly not least, I thank my wife for her patience and support throughout.
## Contents

Chapter 1: Introduction .......................................................................................................................... 8
References ............................................................................................................................................... 11

Chapter 2: Use of the Bootstrap and Permutation Methods for a more Robust Variable Importance in the Projection Metric for PLS .......................................................................................... 13
Abstract .............................................................................................................................................. 13
Introduction ........................................................................................................................................ 14
Methods ............................................................................................................................................. 14
  Partial Least Squares ................................................................................................................... 14
  Variable Importance in the Projection ..................................................................................... 15
  Jackknife Procedure .................................................................................................................. 16
  Bootstrapping and Permutation for VIP Overview .............................................................. 16
  Bootstrapping Procedure ........................................................................................................ 17
  Permutation Procedure .......................................................................................................... 17
Experimental ....................................................................................................................................... 18
  Datasets ........................................................................................................................................ 19
Results .............................................................................................................................................. 20
Discussion ......................................................................................................................................... 22
Conclusion ......................................................................................................................................... 26
References ......................................................................................................................................... 27

Chapter 3: An Assessment of the Jackknife and Bootstrap Procedures on Uncertainty Estimation in the Variable Importance in the Projection Metric .................................................................................. 28
Abstract ........................................................................................................................................... 28
Introduction ....................................................................................................................................... 29
Methods ......................................................................................................................................... 29
  Partial Least Squares ................................................................................................................ 29
  PLS-VIP ...................................................................................................................................... 30
  Jackknife procedure .................................................................................................................. 32
  Bootstrapping procedure ........................................................................................................ 32
  Bootstrap Confidence Interval Estimation ......................................................................... 33
Experimental ................................................................................................................................. 36
  Experimental Structure ........................................................................................................... 36
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>84</td>
</tr>
<tr>
<td>Methods</td>
<td>84</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>84</td>
</tr>
<tr>
<td>Autocorrelation Correction Factor</td>
<td>86</td>
</tr>
<tr>
<td>Significant multivariate correlation (SMC)</td>
<td>87</td>
</tr>
<tr>
<td>Experimental</td>
<td>88</td>
</tr>
<tr>
<td>Results</td>
<td>91</td>
</tr>
<tr>
<td>Discussion and Conclusion</td>
<td>93</td>
</tr>
<tr>
<td>Appendix A</td>
<td>94</td>
</tr>
<tr>
<td>References</td>
<td>95</td>
</tr>
<tr>
<td>Chapter 7: Conclusion</td>
<td>96</td>
</tr>
<tr>
<td>Summary</td>
<td>98</td>
</tr>
<tr>
<td>Samenvatting</td>
<td>99</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>101</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

The objective of this research is the statistical evaluation of variable selection methodologies in Partial Least Squares regression, with the goal of identifying a method that can be applied across varying data analysis problems. Generally speaking, these problems can often be classified into two areas: real-time and off-line applications. Real-time applications involve multivariate process monitoring and control activities, whereas off-line applications can encompass process optimization, root cause determination for process changes, and product development activities. This research centers on off-line applications and is driven by the need to better understand, troubleshoot and optimize industrial manufacturing processes, using methods that reliably identify important variables driving the current manufacturing performance. Although off-line applications are the focus of this research, many of the conclusions presented here can be considered for real-time applications.

An integral part of interpreting process performance for subsequent optimization is a multivariate understanding of the manufacturing process where process parameters are linked to product critical quality attributes. In many instances process performance information can be gained via the analysis of historical data. The application of robust variable importance metrics to these historical datasets helps ensure continued process knowledge throughout the life of the product. This knowledge can then be more confidently applied to either correct or optimize the manufacturing process. Failure to continuously understand process performance can lead to situations where products are no longer capable of consistently meeting product specifications. This unfavorable situation can have a large impact to a company’s ability to meet market demand. In the context of pharmaceutical manufacturing, this inability can translate into public health concerns where patients are unable to receive needed medications.

A common hindrance to understanding process performance in historical data is the varying degrees of autocorrelation observed in batch manufacturing wherein the current batch is correlated to some degree with previous batches. This autocorrelation may result in a higher false positive rate in identified non-important variables. As such, the overarching goal of this research is to determine which variable selection methods perform best in the presence of autocorrelation, and determine if an autocorrelation correction factor provides any positive adjustment as a function of increasing autocorrelation.

Traditional linear regression methods, such as ordinary least squares, have been popular among data analysts for many years. With the explosion of data being generated and captured as part of routine batch manufacturing, issues associated with missing data, multicollinearity, noisy predictor variables, and the number of monitored process variables being greater than the number of batches, has made the use of traditional linear regression methods prohibitive. Consequently, Partial Least Squares (PLS) has gained popularity within industry for its ability to relate a set of manufacturing processing
predictor variables (i.e. input parameters, intermediate performance and controllable parameters) to a product's quality attribute(s). This is accomplished via a multivariate linear model capable of modeling data in the presence of the aforementioned data concerns and challenges, while at the same time being able to accommodate multiple y-variables, as needed. Beyond being able to provide a predictive model, PLS offers the ability to not only identify important process variables with respect to a particular quality attribute, but also provides a better understanding of the internal X-data structure in both object and variable space.

Within the area of variable importance selection methods, such as the VIP, there has been limited research on the impact to said metrics in the presence of auto-correlation, varying correlation between predictors, and varying signal to noise ratio. Often times, the aforementioned predictor space conditions coalesce to produce a model in which many parameters are deemed important. Adding further complexity to the problem, given the duality of PLS where \( X \) is both characterized and related to \( Y \) \([1]\), it is possible that any weak relationship between \( X \) and \( Y \) could be masked. In the case of the VIP, this has made a practical interpretation of assigned important variables very difficult, thus diminishing the benefits of employing Partial Least Squares analysis for certain problems in manufacturing. This has required the use of additional data mining methods, such as Random Forest, in order to build consensus around identified important variables \([2]\). However, given that the main focus of many data mining methods is on minimizing prediction error, they alone may not provide a good understanding of the internal relevant \( X \)-structure that improves model interpretation. This problem if often brought to bear in industrial manufacturing processes wherein the signal-to-noise ratio is very small and process engineers are attempting to understand the current manufacturing process performance in the presence of special cause variation associated with either favorable or unfavorable process performance. Given that predicting future process performance in the presence of special cause variation is analogous to setting process control limits on an out-of-control process, and as such not recommended, the focus of this research is on methods that help identify the most important variables correlated with the observed process performance.

Although work has been done to determine improved measures of variable importance for predictors \([3, 4, 5, 6]\), these have predominantly focused on regression coefficients. This again points to a central focus on minimizing prediction error; an approach that as stated above may not be aligned with the goal of understanding current process performance. As alluded to above, the Variable Importance in the Projection (VIP) statistic is one metric that has proven to be very useful in understanding what are the reduced \( X \) space variables that may explain changes in process performance \([2, 7]\). Given its incorporation of information from both the predictor matrix, \( X \), and the response(s), \( Y \), the VIP became a good initial candidate metric for assessment in our goal of determining an optimal important variable selection method.

The initial approach in this research was the statistical evaluation of strategies based on resampling methods such as the Jackknife and Bootstrap for calculating the uncertainty around the estimate of a variables VIP. Permutation tests were also used in order to make
a straightforward comparison in the resampled VIP estimate as compared to its randomly permuted estimate. Our initial research indicated that a parameter uncertainty approach using the lower-bound of the 95% jackknife confidence interval being greater than the recommended PLS-VIP cut-off value of 1 provided a reasonable estimate of the most important variables in a model. Given that the jackknife is a linear approximation to the bootstrap [8], that can in some instances under-estimate the variability around an estimate [9], further research was performed as to whether the jackknife estimate of uncertainty is comparable to several bootstrap confidence interval approaches.

During this latter investigation it was discovered that with the addition of new latent variables the performance of the VIP around an important variable was degraded. This problem was found to result from the decomposition of orthogonal variances common to latent modeling methods such as PLS regression and PCA [10]; an orthogonalization mechanism that has been recently used as a basis for model interpretation in Orthogonal-PLS, Target Projection [11] and variable selection in Selectivity Ratio [11]. A focus to understand this orthogonalization mechanism, and its impact to PLS modeling and related methods, was subsequently ensued.

It is ultimately shown in this work that the aforementioned orthogonalization can result in not only regression bias in PLS, but also biased estimation of orthogonal variances in filtering based methods such as Target Projection, SR, and the VIP, due to the use of the cross-product score matrix. We find that this bias is dependent on the degree of rotation away from the regression vector to what we term a basic vector, in which the resultant rotation can result in misinterpretation of a predictor variables importance.

As a result of this finding, the method Significant Multivariate Correlation (SMC) was proposed for the purpose of statistically assessing a predictor variables importance for PLS regression, and classification, taking into account a better understanding of the basic rotation effect presented in this work. In addition, an adjustment factor formulation for the error term of the SMC, used in an associated F-test for determining a variables importance, was presented and resulted in a substantial improvement in controlling the increased Type I error rate that can result from moderately to highly autocorrelated data.

An outline of this thesis proceeds is as follows. In Chapter 2 a study is presented of how the use of bootstrapping, in conjunction with permutation tests, can provide avenues for improving the selection of variables responsible for manufacturing process changes via the Variable Importance in the Projection statistic. Chapter 3 extends the work of Chapter 2 by providing an assessment of the performance of seven resampling based methods of uncertainty estimation with the goal of assessing which method performs best in reducing the false positive rate, while at the same time not impacting the true positive rate. The results and conclusion presented in Chapter 3 prompted the work presented in Chapter 4 in order to better understand the relationship between the regression coefficients and orthogonally decomposed variances in PLS in both prediction and model interpretation, and resulted in the introduction of the Significant Multivariate Correlation (SMC) as a novel method of determining variable importance. Chapter 5 extends the work of Chapter 4 by performing an assessment of the SMC method on both
simulated and real data sets in order to illustrate its performance over several commonly used important variable selection methods. Finally, Chapter 6 presents an autocorrelation correction formulation that is shown to provide a favorable correction as a function of increasing autocorrelation.

As alluded to above, this research was driven by the need to better understand, troubleshoot and optimize industrial manufacturing processes. Our goal was to further advance the application of robust variable importance metrics, accounting for both multicollinearity and autocorrelation, in order to improve the process knowledge needed to more confidently correct or optimize the current manufacturing process. Given these objectives, the stated goals of this research were accomplished through a better understanding of resampling based methods for determining the uncertainty in the estimation of PLS parameters and the VIP, the introduction of a new variable importance method, Significant Multivariate Correlation (SMC), and an autocorrelation adjustment formulation for the SMC to better control the increased false positive rate associated with autocorrelated data.

**References**


Chapter 2: Use of the Bootstrap and Permutation Methods for a more Robust Variable Importance in the Projection Metric for PLS

Abstract

Bio-pharmaceutical manufacturing is a multifaceted and complex process where in the manufacture of a single batch hundreds of processing variables and raw materials are monitored. In these processes, identifying the candidate variables responsible for any changes in process performance can prove to be extremely challenging. Within this context, Partial Least Squares (PLS) has proven to be an important tool in helping determine the root cause for changes in biological performance, such as cellular growth or viral propagation. In spite of the positive impact PLS has had in helping understand bio-pharmaceutical process data, the high variability in measured response ($Y$) and predictor variables ($X$), and weak relationship between $X$ and $Y$, has at times made root cause determination for process changes difficult. Our goal is to demonstrate how the use of bootstrapping, in conjunction with permutation tests, can provide avenues for improving the selection of variables responsible for manufacturing process changes via the Variable Importance in the Projection (PLS-VIP) statistic. Although applied uniquely to the PLS-VIP in this article, the generality of the aforementioned methods can be used to improve other variable selection methods, in addition to increasing confidence around other estimates obtained from a PLS model.

Introduction

Vaccine manufacturing, within the bio-pharmaceutical industry, is a complex process where information on hundreds of processing variables and raw materials is monitored and collected [1]. In order to better troubleshoot or optimize a vaccine manufacturing process it is important to both the scientist and engineer to have functional parsimonious models in the subset of important predictor variables that provide good measures of model fit as well as being predictive of future performance. PLS has gained popularity within the pharmaceutical industry for its ability to relate $X$ and $Y$ via a multivariate linear model that is able to model data in the presence of multicollinearity [2], and instances in which the number of cases is less than the number of predictor variables. A standard Partial Least Squares analysis provides model fit statistics, in addition to model parameter estimates and Variable Importance in the Projection (PLS-VIP) statistics. It is this measure, the PLS-VIP, which has been found to be useful in understanding the reduced $X$ space predictor variables that best explain process performance.

Although the PLS-VIP is provided as part of a PLS analysis there are few guidelines about its performance or cut-off value under varying predictor space conditions, namely the proportion of relevant predictors, structure of regression coefficients, magnitude of correlation between predictors, and magnitude of signal-to-noise [3]. Often times, the aforementioned predictor space conditions coalesce to produce a model in which many parameters are deemed important, as per PLS-VIP cut-off guidelines, thus making a practical interpretation of the PLS-VIP challenging.

Our goal is to demonstrate how the use of tools common to machine learning and classical statistics, namely bootstrapping and permutation methods as applied to the PLS-VIP, may provide avenues for improvement of the important variable selection process.

Methods

Partial Least Squares

PLS has gained popularity for its ability to relate $X$ and $Y$ via a multivariate linear model that is able to model data in the presence of multicollinearity, and instances in which the number of cases is less than the number of predictor variables. In the case where we have a single response variable, $y$, the PLS regression model with $h$ latent variables can be expressed as per Eq.'s (1, 2) [3]:

(1) \[ X = T'P + E \]  
(2) \[ y = Tc + f \]
Where $X(n \times p)$ is the matrix of predictors, $T(n \times h)$ is the $X$-score matrix of latent variables, $P(p \times h)$ is the matrix of $X$-loadings, $y(n \times 1)$ is the univariate response variable, $c(n \times 1)$ are the $y$-loadings, and where $E(n \times p)$ and $f(n \times 1)$ are the random errors of $X$ and $y$, respectively.

The goal of PLS is to maximize the covariance between $T$ and $y$ [4]. This maximization is achieved as per Eq’s (3 – 8), where $t_k$, $p_k$, $w_k$, stand for the $k$-th column of $T$, $P$, and $W$, respectively ($k = 1, 2, \ldots, h$). It is assumed that both $X$ and $y$ have been either mean-centered [4], wherein the mean of each column is subtracted from each of the observations within that column, or standardized to unit variance, wherein the mean of each column is subtracted from each of the observations within that column, and then divided by the corresponding column standard deviation. The practice of standardizing to unit-variance is common in the analysis of manufacturing process data given the varying scales of the predictor variables. As demonstrated in Eq.’s (3 – 8), when there is only one $y$-variable, the PLS algorithm converges after one iteration, and the need to loop between steps 1 – 4 is removed [4].

(3) \[ w_k = X'(k)y(k)/\|X'(k)y(k)\| \]

(4) \[ t_k = X(k)w_k \]

(5) \[ \hat{c}_k = t'_k y(k)/t'_k t_k \]

(6) \[ p_k = X'_k t(k)/t'_k t_k \]

(7) \[ X_{(k+1)} = X_{(k)} - t_k p'_k \]

(8) \[ y_{(k+1)} = y_{(k)} - t_k \hat{c}_k \]

The algorithm is then repeated beginning with step 1 using $X_{(k+1)}$ and $y_{(k+1)}$ until the required number of latent variables, $T$, are obtained. This step is determined by the data analyst, depending on the application, and is often supported by the use of cross-validation.

Variable Importance in the Projection

The variable importance in the projection method (PLS-VIP) [3], scores the importance of the $j$\textsuperscript{th} variable per Eq. (9) where $p$ in this instance is equal to the number of predictor variables.

(9) \[ VIP_j = \sqrt{p \sum_{k=1}^{h} (\hat{e}_k^2 t'_k t_k)(w_{(jk)})^2 / \sum_{k=1}^{h} \hat{e}_k^2 t'_k t_k} \]
The PLS-VIP measures the contribution of each predictor variable to the model by taking into account the covariance between $\mathbf{X}_{(k)}$ and $\mathbf{y}_{(k)}$, as expressed by $(\mathbf{w}_j)^2$, weighted by the proportion of $\mathbf{y}_{(k)}$ that is explained by the $k$th dimension $(\hat{c}_k^2\hat{t}_k^t\mathbf{t}_k)$. The average of the squared PLS-VIP scores is equal to one; hence the "VIP scores greater than one" rule is generally used as the criterion for important variable selection. For this study, the 95\% jackknife confidence interval for the PLS-VIP was calculated and is subsequently used for important variable selection.

Jackknife Procedure

The jackknife procedure works by repeatedly re-computing the statistic of interest, $\hat{\theta}_{(i)}$, by leaving out the $i$th observation from the dataset. It then calculates the overall parameter estimate, $\hat{\theta}_{(*)}$, by taking the average of the aforementioned replicate estimates. An estimate for the standard deviation of said statistic, $\hat{\sigma}_j$, can then be calculated using both $\hat{\theta}_{(*)}$ and the replicates from the repeated re-computations, $\hat{\theta}_{(i)}$. The overall estimate for the parameter of interest, $\hat{\theta}_{(*)}$, and corresponding standard deviation, $\hat{\sigma}_j$, was calculated per Eq.'s (10, 11) [5]:

\begin{align*}
(10) & \quad \hat{\theta}_{(*)} = (1/n)\sum_{i=1}^{n} \hat{\theta}_{(i)} \\
(11) & \quad \hat{\sigma}_j = \left[\frac{(n-1)}{n} \sum_{i=1}^{n} \left( \hat{\theta}_{(i)} - \hat{\theta}_{(*)} \right)^2 \right]^{1/2}
\end{align*}

where $\hat{\theta}_{(*)} = VIP_j$, $\hat{\sigma}_j = \hat{vip}_j$, and $j = j$th predictor variable. Using the above estimate for $\hat{\sigma}_j$, 95\% confidence intervals were calculated using the appropriate quantile from the t-distribution, $t_{1-\alpha/2,n-1}$.

Bootstrapping and Permutation for VIP Overview

Bootstrapping can aid statistical modeling by addressing the multiplicity-of-models phenomenon [6] and estimation of a parameter's sampling distribution [5]. It has been stated for decision trees that a very different tree (i.e. model), with almost the same cross-validation error, could be obtained by just slightly perturbing the original dataset by randomly removing 2-3\% of the data [6]. This could be attributed to many different models being clustered together with approximately the same cross-validated error [6], but very different model structures, resulting in a different subset of important variables. Hence, performing the analysis over a large set of pseudo-independent datasets, like those generated by bootstrapping, with competing models, could reduce the non-uniqueness problem [6]; fitting individual models to each of these datasets allows for a distribution of the parameter of interest, in our case the PLS-VIP, which takes into account all the
potential realizations of the dataset with respect to the parameter. A variable that is truly important can be expected to remain important across the multitude of datasets being analyzed. In order to add more confidence with respect to the variables true importance, after fitting the model to the bootstrap dataset, the effect of randomly permuting each variable in turn, capturing the permuted variables PLS-VIP, and comparing this to the original PLS-VIP, via measuring its difference, generates the analogue of a matched pairs analysis; one can then examine the magnitude of the differences relative to zero difference.

Bootstrapping Procedure

Bootstrapping involves sampling \( X(\hat{b}) \) datasets, each consisting of \( n \) observations randomly selected with replacement from the original dataset. These \( X(\hat{b}) \) datasets are termed "bootstrap datasets". Given that bootstrapping follows a Poisson distribution, approximately 63\% (1 - \( e^{-1} \)) of the original cases will be in each bootstrap dataset. Parameter estimation follows a straightforward approach wherein one sequentially fits independent models to each of the bootstrap datasets. The overall estimate for the parameter of interest, \( \hat{\theta} \), and corresponding standard deviation, \( \hat{\sigma}_b \), is calculated per Eq.'s (12, 13) \[5\], where \( B \) is the number of bootstrap datasets:

\[
\hat{\theta} = \frac{\sum \hat{\theta}^{*b}}{B}
\]

\[
\hat{\sigma}_b = \left[ \sum_{b=1}^{B} \left( \hat{\theta}^{*b} - \hat{\theta} \right)^2 / (B-1) \right]^{1/2}
\]

where \( \hat{\theta}^{*b} = VIP_j, \hat{\sigma}_b = \hat{\sigma}_{vip_j} \), and \( j = j^{th} \) predictor variable; 95\% confidence intervals are calculated by multiplying \( \hat{\sigma}_b \) by the appropriate quantile from the t-distribution, \( t_{1-\alpha/2,n-1} \).

Permutation Procedure

The rationale behind randomly permuting each predictor variable in turn is that its original relationship with the response variable is disrupted. If the permuted variable is important it is expected that the PLS-VIP value will decreases substantially. Thus, a reasonable measure for variable importance is the difference between the actual PLS-VIP and the randomly permuted PLS-VIP. As defined in Eq. 14, the permutation method corresponds to a null hypothesis of independence between the randomly permuted variable, \( x_j \), and both the response, \( y \), and the remaining non-permuted predictor variables, \( Z \), \[7\]:

\[
H_0: x_j \perp y, Z
\]
The bootstrap and permutation method is applied to the PLS-VIP as follows:

1) Using the original dataset, \textit{a priori} determine the number of PLS components (latent variables) required.

2) Obtain, via bootstrapping, \( X_{(l)} \) bootstrap datasets.

3) PLS models are fit to each of the bootstrap datasets using the number of PLS components determined in Step 1, and the PLS-VIP for each predictor variable is obtained.

4) Immediately following Step 3, randomly permute each variable in turn and re-fit PLS models to the modified dataset, where each variable is randomly permuted in sequence within each bootstrap dataset.

5) Calculate the PLS-VIP score difference between the non-permuted bootstrap variable (Step 3) and the randomly permuted bootstrap dataset for that variable (Step 4).

6) Normalize the difference vector for each variable obtained from Step 5 to its corresponding standard deviation and construct 95\% confidence intervals around the differences. This normalization has the effect of penalizing those sets of differences with higher variability by scaling down their individual difference scores, and rewarding those with lower variability. As stated, the 95\% lower-bound confidence interval around the individual differences is used to determine the importance of a variable per the guidelines listed below.

\begin{itemize}
  \item 95\% confidence interval lower-bound > \( t_{1-\alpha/2,n-1} \) = Important
  \item 95\% confidence interval lower-bound > 0 = Marginally Important
  \item 95\% confidence interval lower-bound < 0 = Not Important
\end{itemize}

\section*{Experimental}

Important variable selection is performed in one of two ways [8]: dimension-wise, where one determines important variables one PLS component at a time; and model-wise, where the PLS model is first obtained and then a variable selection procedure is applied. For this study the model-wise approach was used with the PLS-VIP and BP-VIP methods being applied to four datasets used in [8] and [9], in addition to a dataset obtained from a vaccine manufacturing process. All dataset variables were standardized to unit-variance prior to performing the analyses and applying the variable selection methods.

Important variable selection using the PLS-VIP method was effected using the criterion that the lower-bound on the 95\% jackknife confidence interval did not encompass one. The BP-VIP method was implemented using a total of 300 bootstrap datasets, using the important variable selection criterion that the 95\% confidence interval lower-bound for
the standardized differences was $> t_{1-a/2,n-1}$. Once the most important predictor variables were selected using both PLS-VIP and BP-VIP, and to avoid the introduction of any bias in the selection of the relevant PLS components, SIMCA-P+ v12.0.1.0 [10] was used to determine the final PLS model in the subset of important predictor variables. These models were then used to obtain the comparison metrics listed in Table 1, with the exception of OLS-CV-R2, which was performed in R [11].

For this study, the PLS analysis, PLS-VIP calculation, and graphical display were done using [12, 13]. The jackknife procedure for the PLS-VIP, the BP-VIP bootstrap and permutation procedures, and BP-VIP variable selection were performed using a script written by the first author in the R programming language [14].

Note: These types of studies across multiple datasets require a large number of graphs. Due to a lack of space the graphs presented will be solely based on the VACCINE dataset. Additional graphs will be provided upon request addressed to the first author.

Datasets

Each dataset listed below was split into a calibration dataset (training dataset) and test dataset for the purpose of determining how well the model in the subset of selected important variables predicted an entirely new dataset (test dataset).

ADPN

The ADPN dataset is taken from the industrial manufacturing process of adiponitrile (ADPN), an intermediary of the chemical synthesis of Nylon 6-6. This process is described by the explanatory variables of flow, pressure, temperature and of compositions of reactional mixtures evolving in time, in addition to the response "nickel loss" [9].

- ADPN data set with $n = 57$, $y = 1$, $p = 100$ (training dataset)
- ADPN data set with $n = 14$, $y = 1$, $p = 100$ (test dataset)

LATEX

The LATEX dataset is taken from emulsion polymerisation batch operations of the industrial manufacturing process for latex. The explanatory variables are temperature, level time, monomer input rate, catalyst level and reactive concentration. The response variable, "insoluble products" is the amount of secondary products [9].

- LATEX data set with $n = 210$, $y = 1$, $p = 117$ (training dataset)
- LATEX data set with $n = 52$, $y = 1$, $p = 117$ (test dataset)
Chapter 2

OXY

The OXY dataset is taken from the industrial manufacturing process of titanium oxide, a chemical product with very high tonnage that enters into the composition of many finished goods, particularly, paint. One of the responses, the whiteness property, can be explained as a function of the variables of the hydrolysis, maturity and calcination phases [9].

- OXY data set with \( n = 20, y = 1, p = 95 \) (training dataset)
- OXY data set with \( n = 5, y = 1, p = 95 \) (test dataset)

SPIRA

The SPIRA dataset is taken from the fermentation process used to manufacture Spiramycine, an antibiotic. Observations were made of fermentation operations in several fermentation reactors described by process variables such as stirring power, temperature level, oxygen consumption peaks and the times at which these peaks occur; information recorded over a period of 240 h at regular intervals [9].

- SPIRA data set with \( n = 115, y = 1, p = 96 \) (training dataset)
- SPIRA data set with \( n = 30, y = 1, p = 96 \) (test dataset)

VACCINE

The VACCINE dataset is obtained from a vaccine manufacturing process wherein process variables pertaining to cellular growth and viral propagation are monitored. With respect to the VACCINE dataset, the two most important predictor variables, in their order of importance, are C_CONC and S_E_.

- VACCINE dataset with \( n = 50, y = 1, p = 67 \) (training dataset)
- VACCINE dataset with \( n = 16, y = 1, p = 67 \) (test dataset)

Results

The goal of the following experiment is to determine whether the BP-VIP method or PLS-VIP using the lower-bound on the 95% jackknife confidence interval, is more successful at identifying the most important variables that are driving the manufacturing process that is being modeled. Once the most important variables have been selected by each method (Ref. Table 1 – Variables Selected), subsequent higher value metric results, as can be found in the Test Set CV R2 column in Table 1, imply that the variables chosen by a particular method were indeed the ones that were the most important variables in predicting \( y \). Table 1 summarizes the comparative results between the PLS-VIP and BP-VIP variable selection approaches.
The following criteria were used in order to determine the performance of the variables selection methods:

- **Variables Selected** – Number of important variables selected by either the PLS-VIP or BP-VIP method. These were determined using the training dataset where all predictor variables were considered candidate variables for importance selection.

- **OLS-CV-R2** – Ordinary least squares cross-validated R2 explains how well the variables chosen by the respective method performs using ordinary least squares regression (as measured by the cross-validated R2).

- **LV** – Number of relevant components, \( N \), shows the complexity of the PLS model in the subset of the most important variables selected.

- **R2X** – Cumulative \( X \)-variation modeled after \( N \) components illustrates how well the PLS model in the subset of important variables explains variation in \( X \).

- **R2Y** – Cumulative \( Y \)-variation modeled after \( N \) components shows how much variation is explained in the response variable via the PLS model in the subset of important variables selected.

- **Q2(cum)** – Cumulative overall cross-validated R2Y is the cross-validated R2 for the PLS model in the subset of important variables selected.

- **Test Set CV R2** – The prediction R2 demonstrates how well the PLS model, in the subset of important variables selected, generalizes to an entirely new dataset.

\(^1\)Calculated using the training dataset and only the most important variables selected within each variable selection method.

\(^2\)Calculated using the test dataset and only the most important variables selected within each variable selection method.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Variable Selection Method</th>
<th>Variables Selected</th>
<th>OLS-CV-R2</th>
<th>LV</th>
<th>R2X</th>
<th>R2Y</th>
<th>Q2</th>
<th>Test Set CV R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>VACCINE</td>
<td>PLS-VIP</td>
<td>11</td>
<td>.67</td>
<td>2</td>
<td>.70</td>
<td>.71</td>
<td>.63</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>BP-VIP</td>
<td>2</td>
<td>.60</td>
<td>1</td>
<td>.72</td>
<td>.63</td>
<td>.62</td>
<td>.53</td>
</tr>
<tr>
<td>SPIRA</td>
<td>PLS-VIP</td>
<td>10</td>
<td>.50</td>
<td>2</td>
<td>.65</td>
<td>.52</td>
<td>.48</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>BP-VIP</td>
<td>7</td>
<td>.47</td>
<td>2</td>
<td>.66</td>
<td>.52</td>
<td>.49</td>
<td>.50</td>
</tr>
<tr>
<td>LATEX</td>
<td>PLS-VIP</td>
<td>20</td>
<td>.76</td>
<td>3</td>
<td>.73</td>
<td>.70</td>
<td>.68</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>BP-VIP</td>
<td>26</td>
<td>.76</td>
<td>4</td>
<td>.75</td>
<td>.72</td>
<td>.67</td>
<td>.57</td>
</tr>
<tr>
<td>ADPN</td>
<td>PLS-VIP</td>
<td>15</td>
<td>.69</td>
<td>2</td>
<td>.73</td>
<td>.70</td>
<td>.64</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>BP-VIP</td>
<td>3</td>
<td>.52</td>
<td>1</td>
<td>.76</td>
<td>.55</td>
<td>.55</td>
<td>.43</td>
</tr>
<tr>
<td>OXY</td>
<td>PLS-VIP</td>
<td>2</td>
<td>.44</td>
<td>1</td>
<td>.69</td>
<td>.63</td>
<td>.53</td>
<td>-.56</td>
</tr>
<tr>
<td></td>
<td>BP-VIP</td>
<td>2</td>
<td>.44</td>
<td>1</td>
<td>.69</td>
<td>.63</td>
<td>.53</td>
<td>-.56</td>
</tr>
</tbody>
</table>

Table 1: A discussion of the results listed in Table 1, and relevant Figures, will only be done for the VACCINE dataset, with the exception of the Test Set CV R2 results, where the results across all the datasets used in the analysis are briefly discussed.
Figure 1: VACCINE Dataset PLS-VIP Important Variables with 95% jackknife confidence intervals for each predictor variable in the VACCINE dataset. Variables whose 95% jackknife confidence interval lower-bound did not encompass one were designated important variables. The horizontal line corresponds to the average of the squared PLS-VIP scores for the non-permuted predictor variables.

With respect to the BP-VIP method, Figure 2 displays the 95% confidence intervals between the non-permuted and permuted PLS-VIP scores, whereas Figure 3 displays the actual differences between the bootstrapped PLS-VIP scores and its corresponding randomly permuted PLS-VIP scores, sorted by the magnitude of the mean of this difference vector. The approach detailed in the Materials and Methods section in which the differences displayed in Figure 3 are normalized to their corresponding standard deviation are displayed in Figure 4. The horizontal line in this graph corresponds to the cut-off limit specified in the importance guidelines in the Materials and Methods section, wherein a variable is scored as important if the 95% confidence interval lower-bound $> t_{1-\alpha/2,n-1}$. 

Figure 2: VACCINE Dataset Bootstrapped and Randomly Permuted PLS-VIPs. Bootstrapped PLS-VIP scores (red) and their corresponding randomly permuted scores (blue), with 95% bootstrap confidence intervals around the individual scores, sorted on the magnitude of the PLS-VIP mean. The horizontal line corresponds to the average of the squared PLS-VIP scores for the non-permuted predictor variables.

Figure 3: VACCINE Dataset Non-normalized Bootstrapped and Permuted PLS-VIP. Difference Scores Difference vectors between the bootstrapped PLS-VIP scores and their
corresponding randomly permuted scores, with 95% confidence intervals. Variables are sorted on the magnitude of the mean of the difference vector. The color groupings correspond to the Wilcoxon-Sign Rank statistic, which is solely being used to group variables, by color, that lie at approximately the same distance from zero (horizontal line).

Figure 4: VACCINE Dataset BP-VIP Important Variables. Normalized difference vector between the bootstrapped PLS-VIP scores and their corresponding randomly permuted scores with 95% confidence intervals. Given the normalization, the y-axis can be interpreted as the distance from zero, in standard deviations. The horizontal line in this graph corresponds to the cut-off limit specified in the importance guidelines in the Materials and Methods section for an important predictor variable.

*For the OXY dataset, neither the BP-VIP nor PLS-VIP methods identified important variables using their respective criteria. The values displayed in Table 1 for this dataset reflect the results obtained using the same two variables ranked as 1st and 2nd in importance by both methods based solely on the magnitude of their PLS-VIP score. Although these two variables appear to explain a significant proportion in the overall variability in terms of OLS-CV-R2, LV, R2X, R2Y, and Q2, they do not predict well when using a new dataset (Test Set CV R2). Hence, applying the criteria explained above using either the BP-VIP or PLS-VIP methods may help in identifying only those variables that are predictive of $y$.

**VARIABLES SELECTED (VACCINE DATASET)**

From Figure 4 we can see that two variables, namely C_CONC and S_E_, meet the BP-VIP criterion that the lower-bound on the 95% confidence interval for the standardized...
Chapter 2

differences not encompass 2.01 ($t_{1-a/2,n-1}$). With respect to the PLS-VIP method, from Figure 1 we find that 11 variables meet the criterion that the 95% lower-bound on the jackknife confidence interval not encompass one. Hence, a lesser number of variables were chosen as important by the BP-VIP method. For the VACCINE dataset we are privy to the actual important variables, namely C_CONC and S_E_, and can see from Figure 4 that the BP-VIP method identified said variables in their corresponding importance rank of 1 and 2 (Figure 1).

OLS-CV-R2 – ORDINARY LEAST SQUARES CROSS-VALIDATED R2 (VACCINE DATASET)

When calculating the OLS-CV-R2 using only the most important variables selected within each variable selection method we find a difference between the BP-VIP (60%) and PLS-VIP method (67%), wherein the PLS-VIP scores higher. The additional gains in this metric by the PLS-VIP method are arrived at with 9 supplementary predictor variables.

LATENT VARIABLES SELECTED (VACCINE DATASET)

When fitting models in the subset of identified important variables, the BP-VIP method chose a less complex model (one latent variable) when compared to the PLS-VIP method (two latent variables).

R2X (VACCINE DATASET)

In terms of the amount of variability explained in the predictor matrix, the BP-VIP method (72%), with a less complex model, did just as well as the PLS-VIP method (70%) for the VACCINE dataset.

R2Y (VACCINE DATASET)

With respect to the amount of variability explained in the response variable, the BP-VIP method with a less complex model explains 63% of the variability in the response variable, whereas the PLS-VIP method explained 71%.

Q2 (VACCINE DATASET)

In terms of the cross-validated R2 (Q2), the BP-VIP method (62%) with a lesser number of identified important variables and less complex model did as well as the PLS-VIP method (63%).

TEST SET CV R2 (ALL DATASETS)

In the prediction of a test dataset in the subset of identified important predictor variables the BP-VIP method, with a fewer subset of important variables, either outperformed, or in practical terms did the same, as the PLS-VIP for the VACCINE (53% vs. 47%), ADPN (43% vs. 26%), and SPIRA (50% vs. 51%) datasets.
With respect to the LATEX dataset, although a fewer subset of variables was selected as important by the PLS-VIP method (47%), the Test Set CV R2 was in practical terms lower when compared to the BP-VIP method (57%), demonstrating that the latter method may have been successful in identifying additional variables that were important for predicting $y$ that were omitted by the PLS-VIP method. For the OXY dataset there was no difference in methods and indeed both methods did equally poorly in terms of the Test Set CV R2 (-.56). The Oxy dataset results may point to the ability of both methods to be similarly capable of not detecting important variables for predicting $y$ when there are indeed no relevant predictors.

**Conclusion**

In this article we explored the use of the 95% lower-bound on the jackknife confidence interval for the PLS-VIP versus the combination of bootstrap and permutation methods (BP-VIP) for determining variable importance across five datasets. Given the results listed in Table 1 it can be concluded that the use of the jackknife confidence interval approach is sufficient for ensuring that the correct important variables are selected. In terms of the most important variables the PLS-VIP may be subject to resulting in a less optimal model in the subset of important variables when compared to the BP-VIP method. Additionally, it can be observed from Figure 1 that the jackknife confidence intervals approximate the bootstrap confidence intervals of the non-permuted VIP scores in Figure 2. With respect to this apparent relationship between the jackknife and bootstrap, the jackknife is a linear approximation to the bootstrap [5], and can be considered a more conservative estimate of the variance [15]. Consequently, adopting the bootstrap procedure for estimation of the PLS-VIP sampling distribution, and applying the aforementioned permutation method, combines two powerful methods for important variable selection. The advantage of using the lower-bound on the jackknife confidence interval for the PLS-VIP currently lies in its computational efficiency, whereas the difficulty in implementing both bootstrap and permutation methods, as in the BP-VIP, is that they are computational intensive. In the end, the need to confidently identify the most important variables underlying a manufacturing process may provide ample justification for the implementation of a more robust method of variable selection that is computationally intensive in lieu of a less optimal selection method.
References

Chapter 3: An Assessment of the Jackknife and Bootstrap Procedures on Uncertainty Estimation in the Variable Importance in the Projection Metric

Abstract

Industrial manufacturing processes can be very complex systems where in the manufacture of a single batch hundreds of processing variables and raw materials is monitored. In these processes, where there is a high degree of multicollinearity between predictor variables, identifying the candidate variables responsible for any changes in product quality can prove to be extremely challenging. Within this context partial least squares (PLS), in conjunction with the variable importance in the projection (PLS-VIP) metric, is currently an important tool in determining the most correlated variables and helping to determine the root cause for changes in a product's quality attributes. Using the standard 'greater than one' important variable cut-off rule for the PLS-VIP, our approach is to measure the performance of seven methods of uncertainty estimation with the goal of assessing which method performs best in reducing the false positive rate while at the same time not impacting the true positive rate. Our findings demonstrate that the implementation of either the normal or basic bootstrap confidence intervals for the PLS-VIP will result in a more consistent determination of the important variables. If computation speed is a concern, the use of the bias-corrected jackknife confidence interval is recommended in place of the un-corrected jackknife.

Introduction

Partial least squares (PLS) has gained popularity within the manufacturing industry for its ability to relate a large number of correlated explanatory variables to a response via a multivariate linear model, thus proving helpful in driving at the variables most correlated to product quality changes. A standard PLS analysis provides model fit statistics, parameter estimates, and in many cases the variable importance in the projection (PLS-VIP) statistics. It is this latter metric, the PLS-VIP, which has been found useful in identifying variables associated with the current manufacturing process performance [1, 2]. However, often times complex predictor space conditions coalesce to produce a model in which many explanatory variables are deemed important, as per the PLS-VIP N1 cut-off guideline [2], thus making a practical interpretation of the PLS-VIP, as related to changes in product quality, very challenging. In spite of this limitation [5] determined that a parameter uncertainty approach using the lower-bound of the 95% jackknife confidence interval being greater than the PLS-VIP cut-off value of 1 does indeed provide a reasonable estimate of the most important variables in a model. Given this conclusion the question arises as to whether the jackknife estimate of uncertainty is comparable to a general bootstrap confidence interval approach given that the jackknife is a linear approximation to the bootstrap [3], that can in some instances under-estimate the variability around an estimate [4]. As such, the goal of this study is to compare the coverage properties of the jackknife confidence interval, and its bias-corrected analogue, to five different methods of estimating confidence intervals via the bootstrap, and how this relates to important variable selection via the PLS-VIP.

The motivation for assessing five bootstrap procedures is due to their varying approaches for determining confidence intervals. Hence, their inclusion in this study allows for a level of competition between the bootstrap confidence interval methods presented, and allows us to examine if one general approach is applicable for the PLS-VIP when compared to the jackknife. The general conclusion may be applicable for other model parameters too. The results from our study demonstrate that implementation of either the normal or basic bootstrap confidence intervals for the PLS-VIP will result in a more consistent determination of the important variables currently driving a manufacturing process. If computation speed is a concern, the use of the bias-corrected jackknife confidence interval is recommended in place of the un-corrected jackknife.

Methods

Partial Least Squares

In this paper we only consider the case of a single response variable, \( y \). As such, the PLS regression model with \( h \) latent variables can be expressed as per Eqs. (1) and (2) [2].

\[
X = T'P + E
\]
\( y = Tc + f \)

Where \( X(n \times p) \) is the matrix of predictors, \( T(n \times h) \) is the \( X \)-score matrix of latent variables, \( P(p \times h) \) is the matrix of \( X \)-loadings, \( y(n \times 1) \) is the univariate response variable, \( c(n \times 1) \) are the PLS regression coefficients, and where \( E(n \times p) \) and \( f(n \times 1) \) are the random errors of \( X \) and \( y \), respectively.

The goal of PLS is to maximize the covariance between \( T \) and \( y \) [6]. This maximization is achieved as per Eqs. (3) – (8), as per the NIPALS algorithm where \( t_k, p_k, w_k \), stand for the \( k \)-th column of \( T \), \( P \), and \( W \), respectively \( (k = 1, 2, \ldots, h) \).

\[
(3) \quad w_k = X'_k y(k) / \| X'_k y(k) \|
\]

\[
(4) \quad t_k = X(k) w_k
\]

\[
(5) \quad \hat{c}_k = t'_k y(k) / t'_k t_k
\]

\[
(6) \quad p_k = X'_k t(k) / t'_k t_k
\]

\[
(7) \quad X_{(k+1)} = X_{(k)} - t_k p'_k
\]

\[
(8) \quad y_{(k+1)} = y_{(k)} - t_k \hat{c}_k
\]

The algorithm is then repeated beginning with step 1 using \( X_{(k+1)} \) and \( y_{(k+1)} \) until the required number of latent variables, \( h \), are obtained. This step is determined by the data analyst and is often supported by the use of cross-validation.

PLS-VIP

The variable importance in the projection (PLS-VIP) [2], scores the importance of the \( j \)th predictor variable per Eq. (9) where \( p \) in this instance is equal to the number of predictor variables.

\[
(9) \quad VIP_j = \sqrt{p \sum_{k=1}^{h} \left( \frac{\hat{c}_{k}^2 t'_k t_k}{(w_{(jk)})^2 / \sum_{k=1}^{h} \hat{c}_{k}^2 t'_k t_k} \right)}
\]

The PLS-VIP measures the contribution of each predictor variable to the model by taking into account the covariance between \( X_{(k)} \) and \( y_{(k)} \), as expressed by \( (w_{(jk)})^2 \), weighted by the proportion of \( y_{(k)} \) that is explained by the \( k \)th dimension \( (\hat{c}_{k}^2 t'_k t_k) \). The average of the squared PLS-VIP scores is equal to one; hence the "PLS-VIP score >1" rule is generally used as the criterion for important variable selection, wherein simply the magnitude of the PLS-VIP score for a variable needs to exceed this value. Throughout
this paper “PLS-VIP > 1” will be used to designate this important variable selection criterion.

We would like to note that in the classical use of the VIP a decision as to the ranking of the variables could be made solely on the point estimate of the VIP. As an example, one could choose the top 3 important variables with VIP scores greater than the cut-off criterion of 1 simply based on their descending VIP magnitude. However, this assumes that the VIP score is perfectly estimated given the data. When taking in account the degree of uncertainty in the estimation of the VIP it might be shown that its score is not significantly different than the cut-off criterion of 1. In this instance, the variable should not be counted as an important variable because its score, from a statistical standpoint, could be <1 when taking into account their uncertainty. Hence the most elevated VIP could possibly be discarded because of its wide confidence intervals. Furthermore, their ranking can also be changed accordingly. Our contention is that this estimate of uncertainty can from a theoretical stand-point be correctly estimated via the bootstrap, and its application can help in reducing the Type I error rate (false positives). The rationale of using two-sided intervals as opposed to a one sided lower-bound is to allow the data analyst the ability to compare the degree uncertainty estimation between variables. This comparison can inform the data analyst as to which parameters are best estimated given the data. As such, rather than using the point estimate for the VIP of each variable, VIP_j, as described above, we will now explore methods to estimate the confidence intervals around this estimate for the purpose of objectively determining a variables importance and ranking. The scientific notation used for the confidence interval methods is as follows:

**Notation for Confidence Interval Methods:**

\[ \theta = \text{population parameter} \]
\[ \hat{\theta} = \text{sample estimate of } \theta \]
\[ \hat{\sigma} = \text{sample estimate of the population parameter } \sigma \]
\[ \hat{\theta}_{(i)} = \text{jackknife replicate estimate of } \hat{\theta} \text{ with the } i^{\text{th}} \text{ observation removed} \]
\[ \hat{\theta}_* = \text{jackknife estimate of } \hat{\theta} \text{ across all jackknife replicates} \]
\[ \hat{\theta}_j = \text{jackknife estimate of } \hat{\sigma} \]
\[ \hat{B}_j = \text{jackknife estimate of bias} \]
\[ \hat{\theta}^{*b} = \text{bootstrap replicate sample estimate of } \hat{\theta} \]
\[ \hat{\theta}^* = \text{bootstrap estimate of } \hat{\theta} \text{ across all bootstrap replicates} \]
\[ \hat{\sigma}_b = \text{bootstrap estimate of } \hat{\sigma} \]
\[ \hat{B}^* = \text{bootstrap estimate of bias} \]
\[ \alpha = \text{statistical significance level} \]
Jackknife procedure

The jackknife procedure, popular in chemometric applications, works by repeatedly re-computing the statistic of interest, $\hat{\theta}_{(i)}$, by leaving out the $i^{th}$ observation from the dataset. It then calculates the overall jackknife estimate of the parameter, $\hat{\theta}_{(\ast)}$, by taking the average of the aforementioned replicate estimates (10). An estimate for the standard deviation of said statistic, $\hat{\sigma}_J$, can then be calculated using both $\hat{\theta}_{(i)}$ and the replicates from the re-computations, $\hat{\theta}_{(\ast)}$ (11) [7].

\begin{equation}
\hat{\theta}_{(\ast)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}
\end{equation}

\begin{equation}
\hat{\sigma}_J = \left[ \frac{(n-1)}{n} \sum_{i=1}^{n} \left( \hat{\theta}_{(i)} - \hat{\theta}_{(\ast)} \right)^2 \right]^{1/2}
\end{equation}

Using the above estimate for $\hat{\sigma}_J$, 95% confidence intervals can be calculated using the appropriate quantiles from the $t$-distribution (12).

\begin{equation}
(\theta_{\alpha}, \theta_{1-\alpha}) = \hat{\theta} \pm t_{(1-\alpha/2, n-1)} \hat{\sigma}_J
\end{equation}

An additional property of the jackknife procedure is that it allows for the estimation of bias between the current estimate and the target parameter (13) [8]. This bias-correction, $\hat{B}_j$, can then be applied to the jackknife estimate to obtain a bias corrected estimate (14).

\begin{equation}
(\theta_{\alpha}, \theta_{1-\alpha}) = \hat{\theta} \pm t_{(1-\alpha/2, n-1)} \hat{\sigma}_J
\end{equation}

\begin{equation}
\hat{B}_j = (n-1)(\hat{\theta}_{(\ast)} - \hat{\theta})
\end{equation}

Bootstrapping procedure

Bootstrapping involves sampling $B$ datasets, each consisting of $n$ observations randomly selected with replacement from the original dataset. These $B$ datasets are termed "bootstrap sample replicates". Parameter estimation follows a straightforward approach wherein one sequentially fits independent models to each of the bootstrap samples. The overall estimate for the parameter of interest, $\hat{\theta}^*$ (Eq. (15)), and corresponding standard
deviation, $\hat{\sigma}_B$, is calculated as detailed in [9], where $B$ is the number of bootstrap samples (Eq. (16)).

$$\hat{\theta}^{**} = \frac{\sum_{b=1}^{B} \hat{\theta}^{*b}}{B}$$  \hspace{1cm} (15)

$$\hat{\sigma}_B = \left[ \sum_{b=1}^{B} \left( \hat{\theta}^{*b} - \hat{\theta}^{**} \right)^2 / (B-1) \right]^{1/2}$$  \hspace{1cm} (16)

Bootstrap Confidence Interval Estimation

Bootstrap confidence intervals are generally classified into two types [10]: pivotal, where symmetry is imposed, and non-pivotal, in which symmetry is not imposed. The determination of which one to use should be preceded by an examination of the bootstrap distribution [12], especially when dealing with small sample sizes. This assessment should take into account the extent to which the bootstrap distribution approximates a normal distribution and the overall estimate of bias, thus helping ensure that the correct bootstrap confidence intervals are chosen so as to maximize the likelihood of coverage for the population parameter of interest.

PIVOTAL

Pivotal methods use the quantity $1 - \alpha = P(L \leq \hat{\theta} - \hat{\theta} \leq U)$ in order to arrive at the desired confidence interval [8]. These methods work well when the assumption of normality is met, and having an unbiased estimate of $\hat{\theta}$, the population parameter. Having a biased estimate of $\hat{\theta}$ can at times be assuaged by applying a bias-correction [11].

NORMAL APPROXIMATION BOOTSTRAP CONFIDENCE INTERVALS

This method relies on obtaining an unbiased estimate of $\hat{\sigma}$ and the sample estimate of $\theta$ being adjusted for any bias [13]. The derivation of the normal bootstrap confidence intervals via the above pivotal quantity results is shown in Eq. (17), where $\hat{B}^* = \hat{\theta}^{**} - \hat{\theta}$ is the bootstrap estimate of bias, and $z$ are the appropriate quantiles from the standard normal distribution.

$$\hat{\theta}^{*b} = \hat{\theta} - \hat{\sigma}_B z_{1-\alpha}$$  \hspace{1cm} (17)

BASIC BOOTSTRAP CONFIDENCE INTERVAL

In calculating the Basic bootstrap confidence intervals [15], the goal is to arrive at a lower ($L$) and upper bound ($U$) as per (18).
(18) \[ 1 - \alpha = P(L \leq \hat{\theta} - \theta \leq U) \]

By replacing the parameter estimate \( \hat{\theta} \) with the bootstrap estimate \( \hat{\theta}^* \), and the population parameter, \( \theta \), with \( \hat{\theta} \) we arrive at Eqs. (19) and (20).

(19) \[ 1 - \alpha = P(L \leq \hat{\theta}^* - \hat{\theta} \leq U) \]

(20) \[ 1 - \alpha = P(L + \hat{\theta} \leq \hat{\theta}^* \leq U + \hat{\theta}) \]

Using the bootstrap distribution to obtain the percentiles, \( q\left(\frac{\alpha}{2}\right) \) and \( q\left(1 - \frac{\alpha}{2}\right) \), and setting these equal to the endpoints of the interval defined in Eq. (18), the lower (\( L \)) and upper (\( U \)) bounds can be solved for algebraically (Eq. (21)):

(21) \[ 2\hat{\theta} - q\left(1 - \frac{\alpha}{2}\right) \leq \theta \leq 2\hat{\theta} - q\left(\frac{\alpha}{2}\right) \]

Although not obvious, there is a bias correction associated with the basic bootstrap confidence interval. In the Normal approximation bootstrap confidence intervals, \( \left(\hat{\theta} - \hat{B}^*\right) - \theta = 2\hat{\theta} - \hat{\theta}^{**} \), is the bias-corrected estimate; the basic bootstrap simply replaces \( \hat{\theta}^{**} \) with the percentiles from the bootstrap distribution, \( q\left(1 - \frac{\alpha}{2}\right) \) and \( q\left(\frac{\alpha}{2}\right) \), as shown in Eq. (21).

**STUDENT’S T CONFIDENCE INTERVAL**

The Student's t confidence interval is similar to the jackknife confidence interval with the difference being that the bootstrap estimate \( \hat{\sigma}_B \) is used as the 'plug-in' standard error estimate for \( \hat{\theta} \) [14]. Using, \( \hat{\sigma}_B \), 95% confidence intervals can be calculated using the appropriate quantiles from the t-distribution (22).

(22) \[ (\theta_{1 - \alpha}, \theta_{1 - \alpha}) = \hat{\theta} \pm t_{(1 - \alpha/2, n-1)} \hat{\sigma}_B \]

**NON-PIVOTAL**

Non-pivotal methods [10] rely on the use of the percentiles from the actual bootstrap distribution to determine confidence intervals for the parameter of interest at some pre-stated \( \alpha \)-level.
PERCENTILE BOOTSTRAP CONFIDENCE INTERVAL

The Percentile bootstrap confidence interval method uses the empirical percentiles of the bootstrap distribution to determine the width of the confidence intervals [16]. The percentile method makes no assumptions about the underlying distribution of the parameter of interest and provides a simple and straightforward non-parametric approach for determining confidence intervals (23), where $B$ is the number of bootstrap sample replicates. The quantity $(B + 1)$ in (23) and (24) helps account for total bootstrap sample replications that are odd in number in order to avoid interpolation in the determination of the required percentiles [17]:

$$
\hat{\theta}^{*b}_{[(B+1)\alpha/2]} < \theta < \hat{\theta}^{*b}_{[(B+1)(1-\alpha/2)]}
$$

BOOTSTRAP BCA CONFIDENCE INTERVAL

The bootstrap bias-corrected and accelerated confidence intervals (BCa) are an adjustment of the Percentile method that accounts for both bias and the rate of change of the standard error of $\hat{\theta}$ with respect to $\theta$ [18]. Bias adjustment proceeds by estimating the proportion of bootstrap samples that are at or below $\hat{\theta}$ (24) [18],

$$
\hat{z} = \Phi^{-1}\left[\frac{\sum_{i=1}^{B} (\hat{\theta}_{i} - \hat{\theta}_{(*)})}{\hat{a} \left(\sum_{i=1}^{B} (\hat{\theta}_{i} - \hat{\theta}_{(*)})^{2}\right)^{3/2}}\right]
$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the standard-normal cumulative distribution function, and $\hat{z}$ represents the calculated quantile from the standard normal. The acceleration parameter, $\hat{a}$, is used to adjust for skewness in the calculation of the BCa confidence intervals [18]. The theoretical background with respect to how $\hat{a}$ provides an estimate of the acceleration constant that adjusts for skewness can be found in [19], but can be shown to be well estimated by one-sixth of the jackknife estimate for skewness (Eq. (25)) [19].

$$
\hat{a} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{i} - \hat{\theta}_{(*)})^{3}}{6 \left(\sum_{i=1}^{n} (\hat{\theta}_{i} - \hat{\theta}_{(*)})^{2}\right)^{3/2}}
$$

After calculating the bias and acceleration correction factors we obtain the appropriate percentiles from the bootstrap distribution (Eqs. (26) and (27)) in order to determine the BCa confidence intervals (Eq. (28)) [20].

$$
\alpha_{1} = \Phi\left[\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{a}(\hat{z} + z_{\alpha/2})}\right]
$$
(27) \[ \hat{z}_2 = \Phi \left( \tilde{z} + \frac{\tilde{z} + z_{1-\alpha/2}}{1 - \hat{a}(z + z_{1-\alpha/2})} \right) \]

(28) \[ \hat{\theta}^{(B+1)u_i} < \theta < \hat{\theta}^{(B+1)u_2} \]

where \( \Phi(\bullet) \) is the standard-normal cumulative distribution function, and the values \( z_{\alpha/2} \) and \( z_{1-\alpha/2} \) represent quantiles from the standard normal at the required \( \alpha \) level.

**Experimental**

**Experimental Structure**

The first part of this study focuses on quantifying the sensitivity and specificity of each method via a simulation experiment. This portion of the study will help gauge how each method performs in detecting variables induced to correlate with the response variable in the presence of varying degrees of multicollinearity.

The second part of this study focuses on four real manufacturing process datasets with process changes that are representative of those observed in processing. With the exception of the VACCINE dataset, the authors are not privy to which variables are most correlated with the current process performance displayed in Figure 2. As such the focus is to solely gauge each method’s ability to identify candidate variables that correlate with the observed process behavior. Given that predicting future process performance in the presence of what appears to be special cause variation is analogous to setting process control limits on an out-of-control process the focus will be on internal cross-validation as the gauge of whether each of the methods investigated identify the most important variables correlated with the observed process performance.

A general agreement in the results of both simulation experiment and actual process datasets should provide initial guidance on which approach is most suitable for assigning important variables in a current manufacturing process via the use of the PLS-VIP.

For the estimation of all the bootstrap confidence intervals 999 bootstrap replications were performed; all confidence intervals were set at 95% confidence.

**SIMULATED MANUFACTURING PROCESS EXPERIMENTAL SECTION**

A factorial experiment was performed using a simulated manufacturing process composed of seven unit operations, 51 process variables, and 60 batches. The seven unit operations were chosen to adequately simulate a manufacturing batch process, wherein multiple processing steps are grouped within specific sequential correlated manufacturing operations (Diagram 1). For each unit operation the within-unit-operation correlation structure is listed (corr=). This correlation structure was chosen to mimic the varying
degrees of multicollinearity that may be observed in a multifaceted batch manufacturing process. In some instances there may multiple levels of multicollinearity present within one unit operation, as shown in Unit Operations 3, 4, and 7. The complexity of the simulation correlation structure has been purposefully designed in the hopes that the results will be more relevant and generalizable to complex manufacturing systems. At each simulation iteration observations intended to simulate the manufacturing process were sampled from a multivariate normal distribution as per the correlation structure defined in Diagram 1. The three processing variables correlated to the response were placed in a sequence intended to simulate multiple influential events whose magnitude was dependent on their proximity to the end of the manufacturing process. An example of the predictor variable correlation structure defined as the baseline correlation is displayed in Fig. 1. This baseline correlation structure was varied in two ways, by first increasing and then decreasing it by 25%. The motivation for this was to see if any one important variable selection method was heavily influenced by the predictor variable correlation structure. One-hundred experimental replications of each factorial combination were performed, and the average of these results was used to assess the performance of each method. The performance assessment of each variable selection method in identifying the important variables was based on the approach outlined in [21] wherein the geometric mean of both sensitivity and specificity, is assessed via the metric G (34). Hence, the greater a method's G value, the better that method did in identifying the variables induced to correlate with the response. For the calculation of Eq. (29) sensitivity is defined as the number of true important variables identified by each method, divided by the total number of actual important variables. Specificity was measured as the number of true unimportant variables identified by each method, divided by the total number of actual unimportant variables.

\[
G = \sqrt{\text{Sensitivity} \times \text{Specificity}}
\]

Diagram 1 – Simulated Manufacturing Process Structure: This simulation is designed to represent a manufacturing process composed of 7 unit operations. Total number of simulated production batches was 60.
Figure 1 – Simulated Manufacturing Process Baseline Correlation Structure
MANUFACTURING PROCESS EXPERIMENTAL STUDY OVERVIEW

MANUFACTURING PROCESS DATASETS

Figure 2 – Run-charts of actual manufacturing process datasets.

ADPN

The ADPN dataset is taken from the manufacture of adiponitrile (ADPN), an intermediary of the chemical synthesis of Nylon 6-6. The main step in the preparation of ADPN lies in a complex catalyzed nickel reaction, described by flow, pressure, temperature and reaction mixture compositions. The response variable for this dataset is specified as “nickel loss”, whose decrease as time passes, results in a decrease of the yield of ADPN. It can be assumed that the identification and adjustment of the levels of certain explanatory variables can limit the loss in nickel while at the same time maintaining a minimal level of productivity of ADPN.[22].

- ADPN data set with $n = 57$, $y = 1$, $p = 100$
Chapter 3

LATEX

The LATEX dataset is taken from an emulsion polymerization manufacturing process for latex. The explanatory variables are temperature, level time, monomer input rate, catalyst level and reactive concentration. The response variable is the amount of undesired secondary products. The objective is to determine a parsimonious model of the most important variables that explain the amount of these secondary products [22].

- LATEX data set with n = 210, y = 1, p = 117

SPIRA

The SPIRA dataset is taken from a fermentation process used to manufacture the antibiotic, Spiramycine (biotechnology). Observations were made of fermentation operations in several fermentation reactors described by process variables such as stirring power, temperature level, oxygen consumption peaks and the times at which these peaks occur. Given that fermentation is a long process the existence of a relevant model may make it possible to stop this reaction early if the content predicted by the model within the first 50 hours is too low in comparison with production standards [22].

- SPIRA data set with n = 115, y = 1, p = 96

VACCINE

The VACCINE dataset results from the manufacture of a vaccine where the response is specified as “yield”, which shows an increase as time passes. The explanatory variables in this dataset represent variables known to potentially impact cellular growth that viral propagation and can result in the observed yield increase. It can be assumed that there is a root cause for the observed increase in yield and that the identification of this root cause can provide increased process knowledge that may be directly linked to better process control.

- VACCINE dataset with n = 50, y = 1, p = 67

Note: More detailed descriptions of the ADPN, LATEX, and SPIRA can be found in [22].

PLS models were built for all four datasets. In order to avoid over-fitting in the selection of the relevant PLS components, cross-validation was used to determine the number of latent variables to model each process dataset. The procedures outlined in the Methods section were then applied to each of the four process datasets and the most important variables designated by each method were identified.

The primary ranking of each variable selection method was based on the magnitude of the CV- $R^2$. In instances where important variable selection methods had a final cross-validated $R^2$ within 3% points of one-another, indicating comparable cross-validation results, the method with the least number of important variables selected was chosen. When the number of important variables chosen was the same, the method with the least model complexity, as measured by the final number of latent variables, was chosen as
best. The above scheme for method ranking was implemented to ensure that variable selection methods with a very close CV- \( R^2 \) were adequately assessed in terms of their overall performance and results, and no one method was favored over another solely based on a single metric.

Note: These types of studies across multiple datasets require a large number of graphs. Due to a lack of space the graphs presented will be based on the ADPN dataset.

Results

PROCESS SIMULATION

The results from the manufacturing process simulation are outlined in Table 1, wherein the results for Sensitivity, Specificity, and \( G \) represent the rounded average across the 100 experimental replicates. It can be observed from this table that the bootstrap Normal and Basic confidence interval approaches are consistently favored, as indicated by their \( G \) scores. Although the bootstrap methods appear to be consistently favored across the simulated process datasets, it can also be noted that the bias-corrected jackknife confidence interval also did well in the three simulated datasets when compared to the uncorrected jackknife confidence interval approach. Although the non-confidence PLS-VIP approach did well (PLS-VIP > 1 rule), a high false positive rate was associated with using a cut-off value of one. The uncorrected jackknife tended to have the lowest \( G \) score among all the methods evaluated.

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>( G )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>Base</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Jackknife CI</td>
<td>Base</td>
<td>0.38</td>
<td>0.98</td>
<td>0.61</td>
</tr>
<tr>
<td>Jackknife BC CI</td>
<td>Base</td>
<td>0.51</td>
<td>0.93</td>
<td>0.69</td>
</tr>
<tr>
<td>Bootstrap Normal CI</td>
<td>Base</td>
<td>0.65</td>
<td>0.95</td>
<td>0.78</td>
</tr>
<tr>
<td>Bootstrap Basic CI</td>
<td>Base</td>
<td>0.65</td>
<td>0.95</td>
<td>0.78</td>
</tr>
<tr>
<td>Bootstrap Student CI</td>
<td>Base</td>
<td>0.59</td>
<td>0.97</td>
<td>0.76</td>
</tr>
<tr>
<td>Bootstrap Percentile CI</td>
<td>Base</td>
<td>0.48</td>
<td>0.98</td>
<td>0.68</td>
</tr>
<tr>
<td>Bootstrap BCa CI</td>
<td>Base</td>
<td>0.48</td>
<td>0.86</td>
<td>0.65</td>
</tr>
<tr>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>Increased by 25%</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Jackknife CI</td>
<td>Increased by 25%</td>
<td>0.36</td>
<td>0.98</td>
<td>0.59</td>
</tr>
<tr>
<td>Jackknife BC CI</td>
<td>Increased by 25%</td>
<td>0.49</td>
<td>0.92</td>
<td>0.67</td>
</tr>
<tr>
<td>Bootstrap Normal CI</td>
<td>Increased by 25%</td>
<td>0.62</td>
<td>0.95</td>
<td>0.77</td>
</tr>
<tr>
<td>Bootstrap Basic CI</td>
<td>Increased by 25%</td>
<td>0.61</td>
<td>0.95</td>
<td>0.76</td>
</tr>
<tr>
<td>Bootstrap Student CI</td>
<td>Increased by 25%</td>
<td>0.56</td>
<td>0.97</td>
<td>0.74</td>
</tr>
<tr>
<td>Bootstrap Percentile CI</td>
<td>Increased by 25%</td>
<td>0.46</td>
<td>0.98</td>
<td>0.67</td>
</tr>
<tr>
<td>Bootstrap BCa CI</td>
<td>Increased by 25%</td>
<td>0.50</td>
<td>0.90</td>
<td>0.67</td>
</tr>
<tr>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>Decreased by 25%</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Jackknife CI</td>
<td>Decreased by 25%</td>
<td>0.46</td>
<td>0.97</td>
<td>0.67</td>
</tr>
<tr>
<td>Jackknife BC CI</td>
<td>Decreased by 25%</td>
<td>0.55</td>
<td>0.94</td>
<td>0.72</td>
</tr>
<tr>
<td>Bootstrap Normal CI</td>
<td>Decreased by 25%</td>
<td>0.65</td>
<td>0.95</td>
<td>0.79</td>
</tr>
<tr>
<td>Bootstrap Basic CI</td>
<td>Decreased by 25%</td>
<td>0.65</td>
<td>0.96</td>
<td>0.79</td>
</tr>
<tr>
<td>Bootstrap Student CI</td>
<td>Decreased by 25%</td>
<td>0.59</td>
<td>0.98</td>
<td>0.76</td>
</tr>
<tr>
<td>Bootstrap Percentile CI</td>
<td>Decreased by 25%</td>
<td>0.52</td>
<td>0.98</td>
<td>0.71</td>
</tr>
<tr>
<td>Bootstrap BCa CI</td>
<td>Decreased by 25%</td>
<td>0.52</td>
<td>0.93</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 1 – Results from the Manufacturing Process Simulation Experiment: Methods with a high \( G \) value (bold font) are considered to have performed best within the simulation experiment as compared to those with a lower \( G \) metric.
Chapter 3

Process Datasets

The results from the analysis of the process datasets are outlined in Table 2. Those variables designated as important within each method were extracted from the overall dataset and sorted in descending order within each method. For the confidence interval based methods the approach of using the magnitude of the lower-bound confidence interval was used for variable sorting. With respect to the PLS-VIP >1 rule approach, the magnitude of a variables PLS-VIP score was used to sort important variables.

After sorting the important variables within each method a forward elimination procedure was applied wherein the important variables were introduced into the model sequentially in their descending order of importance. In this manner the effect of a newly introduced variable was assessed after the introduction of a previous more highly ranked variable(s). This was done to better determine which method performed best in selecting the most important variables in their correct importance rank. During this forward selection procedure the cross-validated R$^2$ was monitored; once this value began to decrease, or reached a plateau, the model building process was deemed complete. The rationale for using a forward selection process arises from the nature of the hypothesis question. As can be seen from the process response variable charts (Figure 2), we are attempting to assess which variable(s) are potentially responsible for the observed process performance. The forward selection approach aids by ensuring that only those variables that explain a significant proportion of the current process variability are considered by the process engineer as potential important explanatory variables. The performance of each method was then assessed in terms of its cross-validated R$^2$, final number of important variables chosen and model complexity (number of latent variables). If all methods chose the same important variables in the same importance rank, then the analysis results, with respect to the cross-validated R$^2$, would be identical.

The authors would like to note that the observed CV- R$^2$ results in Table 2 align with cross-validated results observed when modeling actual process data, where the signal-to-noise ratio is known to be much lower than in other simpler, and more well characterized, biological and chemical processes.

With the exception of the SPIRA dataset, where all the results appear to be uniform in terms of the CV-R$^2$, the Normal bootstrap confidence interval was favored in two of the four datasets analyzed, followed by the Basic which was favored by one of the datasets analyzed. It should be noted that the bias-corrected jackknife confidence interval approach also fared well in two of the four datasets analyzed, followed by the PLS-VIP >1 rule in which only the magnitude of the VIP score was assessed. For ADPN Figure 3 shows the important variable results of the confidence interval approaches, whereas Figure 4 displays the important variables selected via the PLS-VIP >1. Although the PLS-VIP >1 rule fared well in identifying the most important variables in their correct sequence, it can be observed from Figure 4 that it tends to initially identify a much larger number of important variables, as compared to the other methods. Application of the forward selection process to the selected important variables in this instance, as well as in
the other methods, may provide an avenue for significantly decreasing the initial number of explanatory variables associated with the current process performance. It can also be noted that the bias-corrected jackknife confidence interval approach also did well in the four process datasets when compared to the uncorrected jackknife analogue.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Final No. of Important Variables Selected</th>
<th>Final Model Complexity (LVs)</th>
<th>Final CV-R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADPN</td>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>2</td>
<td>2</td>
<td>51%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Jackknife CI</td>
<td>2</td>
<td>2</td>
<td>51%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Jackknife BC CI</td>
<td>3</td>
<td>3</td>
<td>62%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Bootstrap Normal CI</td>
<td>3</td>
<td>3</td>
<td>62%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Bootstrap Basic CI</td>
<td>3</td>
<td>3</td>
<td>62%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Bootstrap Student CI</td>
<td>2</td>
<td>2</td>
<td>51%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Bootstrap Percentile CI</td>
<td>5</td>
<td>3</td>
<td>52%</td>
</tr>
<tr>
<td>ADPN</td>
<td>Bootstrap BCa CI</td>
<td>6</td>
<td>4</td>
<td>57%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>3</td>
<td>1</td>
<td>71%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Jackknife CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Jackknife BC CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Bootstrap Normal CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Bootstrap Basic CI</td>
<td>3</td>
<td>1</td>
<td>71%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Bootstrap Student CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Bootstrap Percentile CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>VACCINE</td>
<td>Bootstrap BCa CI</td>
<td>2</td>
<td>1</td>
<td>66%</td>
</tr>
<tr>
<td>LATEX</td>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>5</td>
<td>5</td>
<td>66%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Jackknife CI</td>
<td>8</td>
<td>5</td>
<td>72%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Jackknife BC CI</td>
<td>6</td>
<td>6</td>
<td>71%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Bootstrap Normal CI</td>
<td>6</td>
<td>6</td>
<td>71%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Bootstrap Basic CI</td>
<td>6</td>
<td>6</td>
<td>66%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Bootstrap Student CI</td>
<td>6</td>
<td>6</td>
<td>66%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Bootstrap Percentile CI</td>
<td>6</td>
<td>6</td>
<td>66%</td>
</tr>
<tr>
<td>LATEX</td>
<td>Bootstrap BCa CI</td>
<td>6</td>
<td>6</td>
<td>66%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>PLS-VIP &gt; 1.0 Rule</td>
<td>3</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Jackknife CI</td>
<td>6</td>
<td>2</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Jackknife BC CI</td>
<td>5</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Bootstrap Normal CI</td>
<td>5</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Bootstrap Basic CI</td>
<td>5</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Bootstrap Student CI</td>
<td>3</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Bootstrap Percentile CI</td>
<td>3</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>SPIRA</td>
<td>Bootstrap BCa CI</td>
<td>5</td>
<td>1</td>
<td>47%</td>
</tr>
</tbody>
</table>

Table 2 – Results from the Analysis of Manufacturing Process Data: Methods highlighted in bold font are considered to have performed best within each of the actual manufacturing process datasets that were modeled.
Figure 3 – Important Variable Results via the Confidence Interval Method: The x-axis lists all the variables that were deemed important across all the confidence interval approach methods. Vertical lines represent the width of the calculated confidence intervals within each method. Where a confidence interval is present within a method, the corresponding predictor variable on the x-axis was deemed important. Where a confidence interval is omitted, the corresponding predictor variable on the x-axis was deemed not important by that method.

Figure 4 – Important Variables Selected via the PLS-VIP > 1 Rule: The x-axis lists all the variables that were deemed important using the magnitude of the PLS-VIP > 1 rule. Vertical lines are representative of the magnitude of the PLS-VIP.
Discussion

In both the simulation study and actual process datasets the bootstrap Normal and Basic confidence intervals appear to be the most consistent methods for identifying the variables most correlated with the response. An aspect of the PLS-VIP that appears to be mitigated against by the bootstrap Normal, and to some extent the Basic and bias-corrected Jackknife, is the degree to which a variables importance may be either upgraded or downgraded as model complexity increases. This downgrading is inherent in the way the PLS-VIP is calculated wherein variability present in an explanatory variable that may no longer be meaningful in explaining \( y \) is incorporated into that variables final PLS-VIP estimate. This phenomenon is inherent in latent variable modeling methods such as PLS wherein the X-loadings used to decompose \( X(k) \) result in outer-product matrices, \( X(k+1) \), that are rotated towards the largest eigenvector of \( X'(k)X(k) \) (6). Consequently, given that the VIP incorporates the weight vectors \((w_{jk})\), calculated from \( X(k+1) \), \( w_{jk} \) it may contain irrelevant information pertaining solely to \( X(k) \), and not its relationship to \( y \). This rotation away from the explained variance in \( y \) could result in biasing the variances in favor of the explained variance in \( X(k) \), and may lead to misinterpretation of a variables importance. In this case, as latent variables are incorporated into the PLS-VIP, the rotation due to the subsequent component loadings may result in an explanatory variable that was important in the first latent variable being down-weighted in the PLS-VIP as subsequent information regarding that variable is added to the model (Figure 5). A much more detailed discussion of this phenomenon is provided in [23].

As observed in this study the bias-correction properties of the aforementioned methods appear to provide a sufficient adjustment to overcome this inherent downgrading in PLS-VIP calculation, as is further validated when one compares the jackknife confidence intervals to its bias-corrected analogue.

We can also note that if the realizations from the VIP can be assumed to be independent and identically distributed, with \( E(X_i) = \mu \) and \( Var(X_i) = \sigma^2 \), and where \( \sigma^2 \) is estimated via the bootstrap procedure, then the distribution function, \( U_n \), will converge to a standard normal distribution as \( n \to \infty \) (Eq. (30)) [30],

\[
U_n = \sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right)
\]

where \( \overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i \). Given that \( \frac{U_n}{\sqrt{n}} \) constitutes a pivotal quantity, the bootstrap confidence estimation method (22) appears to works well in large sample size datasets, resulting bootstrap distributions are approximately normal (Figure 6). Consequently, the estimate of the standard error obtained, coupled with the bias correction on \( \hat{\theta} \), suitably provides an unbiased estimate of the VIP.

Page 45 of 101
The performance of the PLS-VIP >1 rule, which is currently treated as a popular method for important variable selection, demonstrated that while this approach is capable of providing a reasonable estimate of the most important variables, its inherent high false positive rate (Figure 4) makes it a sub-optimal approach for obtaining an initial parsimonious list of the most important variables. In the absence of any confidence interval driven method, the application of the forward selection process used in this work may help assuage this result by helping the engineer focus on the first few important variables. We would like to also note that a pooled estimate of the overall standard deviation across all the VIP scores for all variables is not recommended in that it does not represent the individual uncertainty around a particular variable. Both the bootstrap and jackknife show that the uncertainty in the estimation of the VIP of each variable is different. When there is no agreed upon analytical solution for the standard deviation of a weighted mean [25], wherein the VIP is simply a weighted average, the bootstrap can be safely used, especially in small sample sizes. Though the calculation of bootstrapped estimates may at times be computationally intense, they forgo having to come up with analytical approximations, or incorrect assumptions, for standard errors.

Figure 5 – ADPN Bootstrap Normal, Bias-Corrected Jackknife, Jackknife: In this Figure the vertical bars represent the confidence intervals associated with each identified important variable via the aforementioned methods; horizontal panels represent increasing model complexity for a three latent variable model. The x-axis lists all the variables that were deemed important across the above confidence interval approach methods. The smooth fit within each panel is provided as a reference line to highlight the on-average decrease in the PLS-VIP estimate as model complexity increases.
Conclusion

The PLS-VIP '>1' rule along with seven different uncertainty estimation methods were presented, along with their resulting ability for detecting important variables in simulated and actual process data sets. The performance of the PLS-VIP >1 rule demonstrated that while this approach is capable of providing a reasonable estimate of the most important variables, it has an inherent high false positive rate. As shown in this work, in most instances, the bootstrap methods for estimating confidence intervals resulted in more positive metrics, signaling that a general bootstrap approach is to be recommended for determining the uncertainty around the PLS-VIP estimate. However, not all the bootstrap approaches had the same performance. In this study both the Normal and Basic confidence interval approaches performed best. As stated, this result may be due to the large sample sizes used in this study, in conjunction with the inherent bias-correcting properties within these methods. This is further validated by the results of the bias-corrected jackknife when compared to its un-corrected estimate. Although there is an inherent bias correction in the calculation of the BCa confidence intervals, as can be seen from Figure 3, this bias correction appears to not be sufficiently large to mitigate against the inherent down-weighting that occurs in the PLS-VIP calculation.
Given these results, and in the absence of the ability to assess the bootstrap distribution in each circumstance, it is the recommendation of the authors that when working with large datasets, such as those often encountered in industrial processes, the implementation of either the Normal or Basic bootstrap confidence interval for the PLS-VIP will result in a more consistent determination of the important variables currently driving a manufacturing process.

If computation speed is a concern, the authors recommend the use of the bias-corrected jackknife confidence interval in place of its un-corrected jackknife. Future work examining bootstrap confidence intervals when dealing with sample sizes much smaller than the ones presented here, and in conjunction with varying cut-off values for the PLS-VIP, should be undertaken.

References


Chapter 4: (SMC) Significant Multivariate Correlation – Evaluating Variable Importance Selection in PLS Regression and Classification

Abstract

Identifying important variables for PLS regression has been a difficult task due to the complexity of PLS where there are varying combinations of rotations and projections between different spaces. By better understanding the relationship between the regression (coefficients), the orthogonally decomposed variances, and the behavior of biased regression, the new method, so called significant multivariate correlation (SMC), has been developed for statistically assessing variable importance for PLS regression and classification. This article summarizes the SMC method, with a focus on the evaluation of the method on simulated and real data sets in order to illustrate its performance over several commonly used method; i.e. Variable importance in the projection (VIP) and Selectivity Ratio (SR).
Introduction

Partial least square (PLS) has become a common regression method for data with highly collinear variables. The application of this method has allowed the analytical chemist to get a better hold of increasingly complex data and extract the relevant information embedded within hundreds of spectra or chromatograms. However, the crucial interpretation step is often hampered by the challenging task of identifying the important variables and providing a ranking of their contribution to the PLS regression model. This is because in taking into account the degree of collinearity between predictor variables, these same variables are summarized and projected onto a reduced orthogonal space defined by a number of limited latent variables that forms the basis within which the regression is performed.

The decomposition of orthogonal variances is at the heart of latent modeling methods such as PLS regression and PCA [2]. This orthogonalization has been recently used as a basis for model interpretation in Orthogonal-PLS and Target Projection [5] and variable selection in Selectivity Ratio [5] using the concept of Orthogonal Filtering Method [4]. Therefore it is extremely important to understand the orthogonalization mechanism and its impact to PLS modeling and related methods.

The appearance of these methods can be traced back to the early 20th century beginning via the use of the power sequence [1] or Krylov sequence which can be seen as an iterative method for determining the dominant eigenvector of a large symmetric matrix. The sequence collects vectors with the property that the vector element is in turn rotated toward and converges to the dominant eigenvector approximated by the last vector. When the Krylov sequence is applied to the cross-product matrix ($X'X$), the sequence vector property is extended with the additional property of orthogonally decompose variances; and it is therefore used for geometric interpretation of PCA and PLS [2, 3]. In this work, we call this as a basic sequence to distinguish it from the general power sequence and highlight its additional property. Here, the basic vector is not only rotated toward the dominant eigenvector but also satisfies the conditions for orthogonality; the total data variance is decomposed into two orthogonal variances, explained and residual.

We show in our work that this orthogonalization can result in not only regression bias in PLS but also biased estimation of orthogonal variances in filtering based methods such as Target Projection which directly influences the interpretation of the PLS model. In particular, the bias is dependent on the degree of rotation away from the regression vector to the first basic vector, in which the resultant rotation can result in misinterpreting a predictor variables importance due to the biased estimates of both the explained and residual variances that are used for assessing statistical significance. Readers can find detail discussion on this topic in [3].

For this reason, we developed in this article and in [3] the method, so called significant multivariate correlation (SMC) for the purpose of statistically assessing a predictor variables importance for PLS regression, and classification, taking into account a better understanding of the basic rotation effect.
We will show that the SMC has the following two required features for variable selection in PLS: it provides a complete list of important variables and ranks them according to their F-distribution test statistic, while taking into account the degree of bias present in the regression model.

Readers are referred to [3] for more detail theoretical discussion of the SMC method, the focus of this article is on the evaluation, illustration of the methods properties on simulated and real data sets and a comparison with Variable Importance in the Projection (VIP) and Selectivity Ratio (SR).

**Methods**

Significance multivariate correlation – (SMC)

The concept underlying SMC is briefly described below; a more detailed methodological and theoretical discussion will be presented in [3].

The decomposition of orthogonal variances is at the heart of latent modeling methods such as PLS regression, PCA and the extensions toward the better interpretation and variable selection TP [5].

In general, given a PCA or PLS model, orthogonal filtering method [4] (OFM) can be seen as a rotation step applied to a vector \( \mathbf{w} \) representing the relevant information of the data; i.e. the 1st eigenvector in PCA or the normalized regression coefficients in PLS. The irrelevant or orthogonal information estimated by OFM is defined as the variation \( \| E_{new} \|^2 \) according to the following:

\[
\begin{align*}
(1) & \quad t_{orth} = Xw \\
(2) & \quad p_{orth} = X't_{orth} / (t_{orth}'t_{orth}) \\
(3) & \quad E_{new} = X - t_{orth}p_{orth}'
\end{align*}
\]

Here, \( t_{orth} \) and \( p_{orth} \) are the new (orthogonal) score and loading estimated by OFM.

In can be seen that OFM - after the normalization - is the first element of the basic sequence defined as follows [3]:

\[
(4) \quad p_{i} = \frac{(x'x)p_{i-1}}{\|(x'x)p_{i-1}\|}
\]

With \( p_0 = w \) and \( p_1 = p_{orth} \).
This is a common algorithm that can be related back to the power method [1], the Krylov sequence, NIPALS [6], SIMPLS [7], and the discussion found in [2]. Here, the cross-product matrix $X'X$ is used instead of a general symmetric matrix in the power sequence to introduce the orthogonality of variances property for every sequence vectors where the residual variance $\|E_{new}\|^2$ will be orthogonal to the explained variance component represented by $t_{ortho}$ and $p_{new}$ except at the initial vector $p_0 = w$ [3].

As shown by (1) and (2) the weight vector, $w$, is rotated to the loading $p_{new}$, basic vector, and hence towards largest eigenvector of $X$ via a basic rotation. This is actually the purpose of the basic sequence in PCA to find eigenvectors and eigenvalues and to allow orthogonal decomposition in PLS. However, this rotation produces an undesired effect for the interpretation of variable importance when the explained variance and the residual in $X$ obtained from (4) is used. This $X$ variances from PLS model are not necessarily relevant because the rotation of the loading toward the dominant eigenvector. Accepting these irrelevant variances can bias the results and leads to in an increased false positive rate in the determination of variable importance. In other terms, any evaluation of the variable importance based directly or indirectly on the explained $X$-variance and the $X$-residual would be negatively affected by a large irrelevant variance in $X$, taken accidentally into account in the rotation step. Indirect use of the loadings corresponds to situations where the explained variance in $X$ is underlying the assessment of the importance of the variables. More illustration of this feature is provided in the simulations in the next section.

We propose a new method called Significance Multivariate Correlation (SMC) for variable selection in PLS by simply discard the basic rotation in orthogonal filtering and target projection. SMC uses directly the normalized regression coefficients so that the right source of variances are used and hence SMC is not a part of the basic sequence which is known in this work that creates variance bias and increase false positive in the model interpretation.

\[
(5) \quad p_{SMC} = \frac{\hat{b}_{pls}}{\|\hat{b}_{pls}\|}
\]

With $\hat{b}_{pls} = RP'(X'X)^{-1}X'y = RP'\hat{b}_{OLS}$ and $R = (P'W)^{-1}$

\[
(6) \quad t_{SMC} = Xp_{SMC} = X \frac{\hat{b}_{pls}}{\|\hat{b}_{pls}\|} \quad \text{as such} \quad \frac{\hat{y}}{\|\hat{b}_{pls}\|}
\]

Therefore the reconstruction of $X$ can be represented in the regression form (without actual regression of $X$ onto $t_{SMC}$) within the score space constructed using the PLS regression coefficients $\hat{b}_{pls}$.

\[
(7) \quad X = t_{SMC}p_{SMC}' + E_{SMC} = \frac{(\hat{y}\hat{b}'_{pls})}{\|\hat{b}_{pls}\|^2} + E_{SMC}
\]

Or for each variable in a form of regression of $x_l$ on $t_{SMC}$:
Lacking the actual regression step (of X on \( t_{SMC} \)) in equations (7)-(8) as the SR method, SMC is not a complete basic rotation and the explained variance (or regression variance) \( \| t_{SMC} p_{SMC} \|^2 \) in equation (9) may not be orthogonal to the estimated residual variance \( \| e_{SMC} \|^2 \) in equation (10) however it reflects the relevant variation in the predicted response projected back onto the original X-variable space via the PLS regression vector.

\[
 SS_{i,PLS,\text{regression}} = \| t_{SMC} p_{SMC} \|^2 = \left\| \frac{(\hat{y}_p)_{PLS}}{\| b_{PLS} \|^2} \right\|^2
\]

\[
 SS_{i,PLS,\text{residual}} = \| e_{SMC} \|^2 = \left\| x_l - \left( \frac{\hat{y}_p)_{PLS}}{\| b_{PLS} \|^2} \right) \right\|^2
\]

\[
 MS_{i,PLS,\text{regression}} = SS_{i,PLS,\text{regression}} / 1
\]

\[
 MS_{i,PLS,\text{residual}} = SS_{i,PLS,\text{residual}} / (n - 2)
\]

\[
 SMC_i = \frac{MS_{i,PLS,\text{regression}}}{MS_{i,PLS,\text{residual}}} \frac{\left\| \frac{(\hat{y}_p)_{PLS}}{\| b_{PLS} \|^2} \right\|^2}{\left\| x_l - \left( \frac{\hat{y}_p)_{PLS}}{\| b_{PLS} \|^2} \right) \right\|^2 / (n-2)}
\]

We should note that the impact to the threshold value for the SMC test is dependent on the actual rotation effect. As stated above, due to the lack of the regression step (orthogonal variance decomposition) in equations (16)-(17), the variance decomposition is biased in that the \( SS_{i,PLS,\text{regression}} + SS_{i,PLS,\text{residual}} \) may not always be equal to the total sum of squares \( \| x_l \|^2 \) and will only do so when the basic rotation effect is negligible with \( p_{SMC} \approx p_{TP} \).

**Variable importance in the projection (VIP)**

With respect to the VIP [8], variable importance is established using projection information from X and y as follows:

\[
 VIP_j = \sqrt{\frac{m \sum_{k=1}^{h} v_k (w_{jk})^2}{\sum_{k=1}^{h} v_k}}
\]
VIP is the proportion of the fraction of the explained variance of $X$ expressed by $v_k = (\hat{e}_k^2 \ e_k \ t_k)$ weighted by the covariance between $X$ and $y$, represented by $w_{jk}$, for each variable $j$ over all latent variables.

The average of the squared PLS-VIP scores is equal to one, hence the cutoff threshold for the VIP, "VIP scores >1", rule is generally used as the criterion for important variable selection.

**Selectivity ratio (SR)**

Similar to the concept of orthogonal filtering [4], non-relevant (orthogonal) variance with respect to $y$ is estimated in TP by performing an eigenvalue decomposition using the normalized regression coefficients, producing the single score vector $t_{TP}$ and loading vector $p_{TP}$ via (1)-(2). As stated before, the rotation step involving the loadings might be unfavorably influence by irrelevant variance present in $X$. The sums of squares of the explained variance and the residual variance are calculated accordingly:

\[
SS_{i,explained} = \|t_{TP}p_{TP,i}'\|^2
\]

\[
SS_{i,residual} = \|E_{TPi}\|^2
\]

The Selectivity Ratio (SR) [5] can be calculated as a F-test with respect to the columns of (15, 16).

\[
Selectivity \ ratio = F_{SR} = \frac{SS_{explained}}{SS_{residual}}
\]

In terms of the degrees of freedom associated with the F-test in (17), this can be improved, but these aspects are beyond the scope of this paper and will be discussed with more detail in [3].

**Results and Discussion**

**SIMULATION**

The simulated data was constructed around a simple structure based on two $X$ blocks. The first one $X_1$ is related to the response $y$ with the first 19 variables ordered in function of their correlation to $y$, starting from 0.95 until 0.05, 11 random variables are then appended. This data was generated using Approximation of a DIstribution for a given COVariance (ADICOV) [10]. The second block, is $X_2$, is constructed in a similar fashion but in relation to another response uncorrelated to the one of interest. These two data blocks contain roughly the same amount of variance. The relative variance contribution of the blocks is set by multiplying $X_2$ by a scaling factor, $\gamma$, ranging from 0.01 to 5. The two blocks, $X_1$ and $X_2$, are then concatenated into a single data matrix $X$. The obtained
correlation structure of 1000 simulated samples is represented in Figure 1A. As stated before, PLS aims at explaining the covariance structure between X and y, therefore one could expect that the performance of the model to explain Y will drop when X contains more unrelated variance.

Figure 1: A) Correlation structure of the simulated data. B). Comparison of the evaluation of variable importance by the VIP, SR and SMC.
FAVORABLE SITUATION Y RELATED INFORMATION DOMINATES IN X

All generated data sets were then analyzed using PLS and the variable importance in each model was evaluated using the VIP, SR and finally the newly introduced SMC. All results for the simulation are regrouped in Figure 1B. The number of true and false positives are summarized in Table 1 for the most favorable (γ = 0.01) and less favorable case (γ = 5). In a favorable case where the data block related to the response (X1) dominates the overall variance (γ = 0.01) all methods appropriately selects the variables most correlated to y. However one can already notice that the SR is very conservative and only selects the 5 most correlated variables whereas the VIP selects the 12 most correlated ones. Finally SMC picks up 19 variables related to Y and the two variables from X2 (false positives). The number of false positives obtained via SMC is the highest. This might sounds alarming but it only reflects the significance level chosen here (0.05). Interestingly SR was set with the same significance level and yet not false positive arise. This illustrates again the conservativeness of SR. The explanation for this conservativeness of the SR lies in the application of the F-test with the incorrect use of the degrees of freedom which is explored in greater detail in [3]. In the case of the VIP, the situation is more complex: the number of false positive might greatly depend on the fixed cutoff value of 1. This fixed threshold does not reflect the change of situation from a clean to a noisy data set. As shown by Wehrens et al. the correct cutoff value is very case dependent [13].

Table1: Number of true positives and false positives selected by the different methods.

<table>
<thead>
<tr>
<th></th>
<th>True positive</th>
<th>False positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ = 0.01γ = 5</td>
<td>γ = 0.01γ = 5</td>
<td>γ = 0.01γ = 5</td>
</tr>
<tr>
<td>VIP</td>
<td>12 12</td>
<td>2 0</td>
</tr>
<tr>
<td>SR</td>
<td>5 4</td>
<td>0 0</td>
</tr>
<tr>
<td>SMC</td>
<td>19 19</td>
<td>2 3</td>
</tr>
</tbody>
</table>

UNFAVORABLE SITUATION Y UNRELATED INFORMATION DOMINATES IN X

When γ is increased the situation becomes less ideal for the PLS algorithm. The variance in X is dominated by unrelated structure and therefore the covariance between X and y is less optimal in terms of regression. In this instance the regression becomes more biased for the true positive variables due to the rotation of the regression coefficients toward the expanded variance in X2.

Hence increased variance (unrelated to y) should lead to a higher false positive rate in most methods. The SR distinguishes itself by not selecting any false positive but this is due to its conservativeness which can be seen by the fact no true positives are selected. The bias effect leading to higher false positive rate is increased with a smaller sample size.

The results obtained on this simple simulation provide some insight on the sensibility (or lack thereof) of the different methods. We’ve shown that the SR tends to be too conservative. The classical VIP and SMC resist well the addition of extra variance. The
performance of the VIP is however dependent on the sample size and reproducing these calculations with a smaller population (of 100 samples from 1000 in this case) – see supplementary figure S1 - leads to the higher selection rate of false positive by the VIP going from (maximally) 2 false positives when using 1000 samples to 8 false positives when using only 100 samples. This may explain the unstable cutoff value of the VIP when compared with the other methods.

The next sections will prove that similar results can be obtained on real data.

NEAR INFRARED DATA SETS

OCTANE DATA

The behavior of all methods was further evaluated using a well-known benchmark data set. The octane data set consist of 60 Near Infrared spectra of gasoline measured between 900 and 1700 nm at a 2 nm interval. A complete description of the experimental settings can be found in [11].

To further investigate the influence of unrelated variance in $X$ we add 400 noise variables with high and low noise variances to these spectra, as represented in Figure 2. All calculations were then performed on both the original octane data and the augmented version including the extra noise variables.

![Figure 2: Octane data set used with and without addition of noise variables.](image)
The variable importance was evaluated using the three methods and is shown in Figure 3 for the augmented octane data. The SR is once again the most conservative approach. The selection of too few variables by the SR leads to a poor predictive power. The VIP and SMC in contrast select more variables and both methods focus on similar set of bands. Note that the VIP selects extra variables around wavelength 1680 which is not selected by SMC. This may indicate the false positive variables due to the bias correction procedure in the inclusion of \(X\)-variance in the VIP.

![VIP SR SMC](image)

**Figure 3**: Variable importance evaluated by the VIP, SR and SMC on the octane data set augmented by noise variables. Note that the selected wavelengths left of the red dashed lines are similar to the results are obtained on the original octane data set, see supplementary Figure S2. The black dashed lines represented the critical values above which variables are selected as important.

Further assessing the quality of these variable selections is difficult without entering into chemical interpretation, it is however possible to compare them according to the prediction performance obtained using PLS on selected variables. Table 2 reports the prediction error for each case. Clearly the PLS model based on variables selected by SMC out-performs the other ones, which indicates that the relevant variables were better determined by the SMC. Interestingly SMC leads to a RMSECV of 0.33 using only 85 variables.

**Table 2**: Prediction performance evaluated by cross validation of PLS models (2 LVs) based on selected variables on the real and augmented octane data sets.

<table>
<thead>
<tr>
<th></th>
<th>Octane data</th>
<th>Augmented data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSECV</td>
<td>Number of variables</td>
</tr>
<tr>
<td>All variables</td>
<td>0.46</td>
<td>401</td>
</tr>
<tr>
<td>VIP</td>
<td>0.38</td>
<td>88</td>
</tr>
<tr>
<td>SR</td>
<td>0.59</td>
<td>11</td>
</tr>
<tr>
<td>sMC</td>
<td>0.33</td>
<td>85</td>
</tr>
</tbody>
</table>
FALSE POSITIVE RATE

The rate of false positives, *i.e.* the selection of non-important variables, can be assessed using the octane data set augmented with noise variables. The noise variables are themselves divided into two blocks: the first one corresponds to noise with a high standard deviation and the second one with low standard deviation; as can be seen in Figure 2. Given the significance level chosen here, one can expect all methods to select approximately 20 false positives (5% of 400 noise variables). First, as expected, one can notice that all approaches are negatively affected by the presence of these extra noise variables. Interestingly the VIP only selects false positive variables if they correspond to high standard deviations; approximately 100 out of 200 “high” noise variables. This indicates that the VIP is influenced by fluctuations of the loadings themselves caused by unrelated sources of variance in \(X\). Once again the SR does not select any noise variable, but also very few true positives. Finally SMC is also affected by the presence of these extra variables and selects 18 of the random variables, but note that this is in line with the significance level (5%) chosen.

The results obtained on the octane data and its augmented version demonstrates that SMC evaluates most adequately the variables importance, leading to a subset of variables providing the best regression model. The evaluation of the false positive rate also validates that only SMC is consistent with the significance level chosen here (5%).

CLASSIFICATION PROBLEM

Finally the applicability of our approach is generalized to a classification problem. PLS is indeed commonly applied as PLS Discriminant Analysis (PLS-DA) where the response \(y\) is not a continuous one but a dummy variable consisting of 0s and 1s. The important variables in such a PLS model are then the one allowing for the separation of two (or more) groups of samples from each other. Translated to *omics* fields such variables become biomarkers allowing for classification of subjects, e.g. healthy individuals from sick ones. To illustrate this aspect we compare the results obtained using subset of variables selected by the different methods mentioned earlier on a NMR metabolomics data set described in [12]. The data used here contains two groups: healthy and neurological affected animals. Table 3 summarizes these results in term of number of selected variables and RMSECV.

Table 3: Prediction performance evaluated by cross validation of PLS models (2 LVs) based on selected variables of the classification data.

<table>
<thead>
<tr>
<th></th>
<th>RMSECV</th>
<th>Number of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>All variables</td>
<td>0.33</td>
<td>153</td>
</tr>
<tr>
<td>VIP</td>
<td>0.29</td>
<td>46</td>
</tr>
<tr>
<td>SR</td>
<td>0.40</td>
<td>20</td>
</tr>
<tr>
<td>SMC</td>
<td>0.28</td>
<td>32</td>
</tr>
</tbody>
</table>
The lowest error is achieved using variables selected by SMC. This indicates that the SMC selected variables are the most relevant set for this classification problem. This performance is achieved based on only 32 variables out of the 153 initially available. The SR is selecting relevant variables (data not shown) yet is too conservative, leading to the exclusion of information relevant to the classification problem and increasing the error in prediction. The performance obtained based on the VIP selected variables is comparable to the ones obtained using SMC, yet 46 variables are included which suggest that some false positives are selected. A possible explanation can be proposed based on the cutoff value used with the VIP. Given that the average of the sums-of-squares of the VIP is equal to 1, the cutoff threshold in the VIP, "VIP scores >1" rule is generally used as the criterion for important variable selection. This is not a statistically justified limit and can be shown to be very sensitive to the presence of non-relevant information pertaining to \( \mathbf{X} \). The results obtained here are consistent with previous observations [13] that the VIP is negatively influence by the variance of unrelated variables when the population size is small.

**Conclusions**

The decomposition of orthogonal variances is at the heart of latent modeling methods such as PLS regression and PCA and it is demonstrated in our work that the orthogonal variances can been seen as the additional property of the basic sequence when moving from the old and well-known power method for finding eigenvectors and eigenvalues to the basic sequence which is the special case of the power method associated with the basic data matrix \( \mathbf{X}'\mathbf{X} \) [2, 3]. Although the sequence decomposition is particularly useful for decomposition and analysis of very large sparse matrices for the variable correlation information, this basic rotation results, however, in a biased regression in the case of PLS due to the oblique projections of the OLS regression coefficients [2, 3]. Understanding this concept of basic rotation is very important because, as shown in this work, it can create a bias in the estimation of the orthogonal residual variance with respect to the response in orthogonal correction methods, such as Target Projection in variable interpretation and selection [3]. The evaluation of the variable importance using the loadings can be negatively affected by a large irrelevant variance in \( \mathbf{X} \).

Given the above, we proposed the SMC variable selection method to maximize the use of the rotation information as the source of information for identifying important variables in the presence of a biased regression method such as PLS, and advancing them for final variable selection in the case of PLS regression and classification. Variables exhibiting minimal bias with respect to their parameter estimation, and statistically significant in the model, are highlighted by the SMC method and ranked based on their respective test statistics relative to the F-distribution.

Simulation and application of SMC to two real data sets (NIR and NMR metabolomics) illustrates the outstanding properties of SMC compared with several representatives of commonly used variable selection methods such as the VIP and Selectivity Ratio (SR). SMC can be expected to provide sufficient answers in the identification of important variables in the PLS model. In all cases, SMC provides the most optimal variable list.
(minimal false negative and false positive errors) and improves the predictive performance of the PLS model due to the bias reduction property of the variable selection method.

**Supporting Information:** Additional results on simulation and NIR data are provided as supplementary information; this material is available free of charge via the Internet at http://pubs.acs.org.

# References


Chapter 5: Interpretation of Variable Importance in Partial Least Squares with Significant Multivariate Correlation (SMC)

Abstract

Despite gaining popularity and success in many modeling applications, Partial Least Squares (PLS) regression continues to provide challenges in the evaluation of important variables. This article describes the relationship between the regression coefficients and orthogonally decomposed variances in PLS. The relation between prediction, model interpretation, and important variable determination is described using the theory of the basic sequence presented here as a special case of the famous Krylov sequence (or the power method).

Variable selection methods e.g. Selectivity Ratio (SR) and Variable Importance in the Projection (VIP) are also described in this framework. We show that the interpretation can be affected by unnecessary rotation towards the main source of variance in the $X$-block. Significance Multivariate Correlation (SMC) is developed using the knowledge obtained from the basic sequence to minimize the effect of irrelevant $X$-structures. Simultaneously SMC highlights the variables most correlated to the response. The performance of SMC is demonstrated, using simulated and real data sets, against commonly used variable selection methods, such as the Variable Importance in the Projection and Selectivity Ratio.

T.N. Tran, N.L. Afanador, and L.M.C. Buydens, Chemometrics and Intelligent Laboratory Systems, (n.d.).
Introduction

Model interpretation is an important task in most applications of Partial Least Squares (PLS), and researchers are continuously looking for optimal interpretation tools. These interpretational features in both sample and variable spaces provide additional information that allows for a better understanding of the specificity of the PLS model in quantifying the contribution of predictor variables.

From its nature as a latent regression method, PLS has the attractive feature of being able to deal with a large number of predictor variables exhibiting a high degree of multicollinearity. However, the use of latent variables instead of the original variables also creates difficulty in model interpretation. This complexity is due to the fact that PLS constructs latent variables that not only maximize the correlation of $X$ to the response $y$, but concurrently tries to maximize the explained variance in $X$ [1]. As such, one cannot directly interpret the model using model parameters such as the weights and loadings. This is especially the case in analytical data influenced by multiple sources of variation. When the dominant source of variation is not related to $y$, the maximization of the explained $X$-variance is likely to bring irrelevant information into the PLS model. The interpretation of PLS model parameters and variable importance based on these parameters is then not a straightforward exercise. The literature largely discuss this issue and multiple approaches have been proposed to tackle this problem [2].

Recently, Target Projection with Selectivity Ratio (SR) [3,4] were proposed to re-quantify the captured $X$-variance to improve interpretation of variable importance via the target rotation or the orthogonal filtering strategy. The objective is to allocate information proportional to the co-variance between the $X$-variables and the response, and at the same time isolate orthogonal irrelevant variation. Using the property of the basic sequence theory proposed in this work, we illustrate that these methods can lead to misinterpretation due to biased estimation of the relevant variances. The basic sequence, as an extension of the Krylov sequence (or the power method) [5], describes the relationship between the regression coefficients and the explained $X$-variance via orthogonally decomposed variances. It also provides new insights into the causes of biased estimates and explains the limitation of directly using the PLS model parameters, such as the weights, loadings, or variance adjustment methods via orthogonal filtering, as exemplified by target projection. Based on this new knowledge, we developed a method, the Significant Multivariate Correlation (SMC), for statistically assessing variable importance using the correct sources of variation.

The focus on this article, the interpretation of variable importance carried out using only parameters calculated from the original PLS model, is referred to in the literature as ‘filter’ variable selection methods [6, 7]. The SMC method, proposed in this work, together with Variable Importance in the Projection (VIP) [8], SR and PLS regression coefficients (Beta CI) [9] are in this class of filter methods. This variable selection category does not require re-fitting the model on another sub-set of variables or modifying the original PLS calculation as in other classes of variable selection methods.
However, the optimal use of the existing model parameters for variable selection and interpretation continues to provide challenges due to the complexity of the PLS model.

This article presents in detail the methodology of SMC and a brief performance illustration of the method on both an omics related simulated dataset and a real dataset of NIR spectra of pharmaceutical tablets. Several commonly used methods in the filter variable selection category, VIP, SR, Beta CI are compared to the SMC. Note that the comparison to methods from other categories i.e. wrapper and embedded, are not in the scope of this article. Readers are referred to [10] for more discussion on this topic. Furthermore, one may also have a different variable selection strategy for a different data type such as spectral or process data. For example in the case of Interval PLS [11] for spectral data, the difference can be seen as an extra pre-processing step to split the wavelength range into smaller equal distance subintervals and in terms of core modeling, the PLS model with the number of input variables is unchanged. The SMC proposed in this work can be used as a general interpretation strategy for a PLS model as applied to different data types.

**Methods**

**Theory**

We will begin in this section with a review of the current ‘filter’ variable selection methods studied in this work. The most popular filter method is most-likely the VIP, which was proposed in 1993 by Wold et al. [8]. As such, this method will be used as our benchmark. More recently the SR [12] was introduced in which the ratio between explained and unexplained variances is used to help focus the data analyst on the most relevant variables.

**Variable importance in the projection (VIP)**

With respect to the VIP [8], variable importance is established using projection information from $X$ and $y$ as follows:

\[
VIP_j = \sqrt{d \sum_{k=1}^{h} v_k (w_{kj})^2 / \sum_{k=1}^{h} v_k}
\]

Where $d$ is the number of variables and $h$ the number of latent variables in the PLS model. VIP is the proportion of the fraction of the explained variance of $X$ expressed by $v_k = c_k^2 t_k' t_k$ weighted by the covariance between $X$ and $y$, represented by $w_{kj}$, for each variable $j$ over all latent variables. The term $c_k$ is obtained for each column of the PLS scores $T$ and for the predicted response $y_k$ in equation (2).
Furthermore, given that the average of the sums-of-squares of the VIP is equal to 1, the
cutoff threshold value in VIP, "VIP scores >1" rule is typically used as the criterion for
important variable selection. This is not a statistically justified limit and can be shown to
be very sensitive to the presence of non-relevant information pertaining to $X$.

**Selectivity Ratio**

Given the PLS regression coefficient vector, $\hat{b}_{PLS}$, Target Projection is performed via the
projection of the rows of $X$ onto the normalized regression coefficients vector in equation
(3); as such $t_{TP}$ is proportional to the predicted values, $\hat{y}$. The loadings, $p_{TP}$, are
obtained by projecting the columns of $X$ onto the obtained score vector, $t_{TP}$, which again
is proportional to $\hat{y}$ in equation (3) [4].

$$t_{TP} = X\hat{b}_{PLS}/\|\hat{b}_{PLS}\|$$

$$p_{TP} = X't_{TP}/(t_{TP}'t_{TP})$$

The ratio of the explained variance ($SS_{l,explained}$) and the residual variance for each
variable ($SS_{l,residual}$) as the sums-of-squares in equation (6) and (7), respectively, is used
in the Selectivity Ratio (SR) to determine the variable importance equation (8).

$$x_i = t_{TP,i}p_{TP,i} + e_{TP,i}$$

$$SS_{l,explained} = \|t_{TP}p_{TP,i}\|^2$$

$$SS_{l,residual} = \|e_{TP,i}\|^2$$

$$SR_i = \frac{SS_{l,explained}}{SS_{l,residual}}$$

A critical threshold value for determining variable importance was suggested in [13]
wherein the $SR_i$ is assessed against the F-distribution with $n - 2$ and $n - 3$ degrees of
freedom.

**Beta_CI**

For the selection of important variables, the straightforward strategy is to quantify the
confidence intervals around the regression coefficients, $\hat{b}_{PLS,i}$ for each variable, $i$. 
Unfortunately, given the dependence on $y$ for formulating the PLS hat-matrix, a closed
analytical form for the uncertainty is not available for PLS regression coefficients; hence
resampling techniques such as the Jackknife [14] or bootstrap [9], are often used to determine the confidence intervals.

The variables are considered important in the Beta_CI approach if \( CI_{bi} \) does not cover zero; aligning with the null hypothesis of not enough evidence to conclude a linear relationship. Various resampling techniques can be used for Beta_CI, but none offer a straightforward ranking of variable importance in the model. The use of the absolute value of the obtained estimate can be used as a guide wherein those variables with higher \( \hat{b}_{pls,i} \) that also exhibit the narrowest confidence intervals are deemed important. A similar approach exists wherein \( \hat{b}_{pls,i} \) is tested against its corresponding bootstrap estimate of the standard error, and sorting on the resulting \( p \)-values.

In this section, we will introduce the theory of basic sequence and demonstrate how it is a special case of the Krylov sequence with the vector rotation properties. This topic is further developed by relating it to the Target Projection and Selectivity Ratio methods. We show that these methods may perform sub-optimally in important variable selection. We address this issue through the SMC method, which we have formulated to correct for the rotation effect that introduces a bias in the variability estimates produced by Target Projection and used in Selectivity Ratio for assessing variable importance.

### Basic sequence

The theory of basic sequence is introduced and used throughout this article. The basic sequence \( \{z^{[0]}, z^{[1]}, ..., z^{[k]}\} \) of a data matrix \( X \) from an initial vector \( z^{[0]} \) is defined in this work as an iterative procedure used to update the vector \( z^{[0]} \) with the cross-product matrix \( X^TX \) in equation (9). The scalar \( k \) represents the number of rotation steps required to reach stability. As such, the basic sequence is as a special case of a Krylov sequence (or the power method) for finding the dominant eigenvector and has been used to interpret some of the behavior of PLS in [5,15].

\[
(9) \quad z^{[k]} = (X'X)^k z^{[0]}, \text{ which is also denoted as } B^{[k]}(z^{[0]}, X)
\]

The basic sequence has the property of the power method in that the basic vector \( z^{[k]} \) is rotated towards the dominant eigenvector of \( X^TX \). Indeed, let \( \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m \) be \( m \) eigenvectors with their eigenvalues ordered as \( \lambda_1 > \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_m \). The initial vector \( z^{[0]} \) and \( z^{[k]} \) can be expressed as in equations (10) and (11), respectively, with scalars \( a_1, ..., a_m \).

\[
(10) \quad z^{[0]} = \sum_{i=1}^{m} a_i \mathbf{v}_i \\
(11) \quad z^{[k]} = (X'X)^k z^{[0]} = (X'X)^k \sum_{i=1}^{m} a_i \mathbf{v}_i = \sum_{i=1}^{m} a_i (X'X)^k \mathbf{v}_i 
\]

Because \( (X'X)^k \mathbf{v}_i = \lambda_i^k \mathbf{v}_i \) holds for each eigenvector \( \mathbf{v}_i \), Equation (11) can be written as
\[ z^{[k]} = \sum_{i=1}^{m} a_i \lambda_i^k v_i \text{, and by a simple reordering becomes} \]
\[ z^{[k]} = a_1 \lambda_1^k \left( v_1 + \sum_{j=2}^{m} \frac{a_j}{\lambda_1} \left( \frac{\lambda_j}{\lambda_1} \right)^k v_j \right) \]

Hence, as \( k \to \infty \), \( z^{[k]} \) aligns along the direction of the dominant eigenvector \( v_1 \).

For enhancing the convergence, the normalization \( z^{[k]} = \frac{z^{[k]}}{\|z^{[k]}\|} \) is applied after each step \( k \). Moving from \( z^{[k-1]} \) to \( z^{[k]} \) is also referred to as a rotation with its speed dependent on the ratio of the dominant eigenvalue to the \( 2^{nd} \) largest eigenvalue, \( \frac{\lambda_1}{\lambda_2} \).

Basic sequence and relation to SR

We are convinced that the concept behind SR is appropriate. However, we show in this work that the (normalized) Target Projection loading, \( p_{TP} \) is basically the basic sequence \( B^{[1]}( \tilde{b}_{PLS}, X) \) in equation (12) determined by substituting equation (3) in equation (4). As the first property of the basic sequence, the TP loading is the rotation of the regression coefficients vector towards the dominant eigenvector of \( \tilde{X}^T X \), which may be independent from the response. For this reason, depending on the actual magnitude of the rotation, \( p_{TP} \) may be less proportional the co-variance of the \( X \)-variables and the response variable \( y \).

\[ \begin{align*}
(12) \quad p_{TP}/\|p_{TP}\| &= \frac{X^t t_{TP}}{(t_{TP}^t t_{TP})} \left/ \frac{\|X^t t_{TP}/(t_{TP}^t t_{TP})\|}{\|X^t \tilde{b}_{PLS}\|} = B^{[1]}( \tilde{b}_{PLS}, X) \right. 
\end{align*} \]

The idea behind SR is the use of a Target Projection (TP) [2,4,12] in an attempt to quantify for each variable a variance that is proportional to the co-variance between the \( X \)-variables and the response variable \( y \), while at the same time separating the orthogonal variation to the response variable. The ratio of these two variances can then be used to determine variable importance in the PLS regression model [4].

Target Projection is, in fact, the application of the orthogonal filtering method (OFM) proposed in [16], but for a single component and with the normalized regression coefficient used as the weight vector. OFM is the application of the projection and regression steps in equations (3) and (4) to obtain the score of each component. Hence OFM attempts to include variability originating from \( X \) while at the same time remaining orthogonal to \( y \) using the weight vectors corresponding to the eigenvectors of \( MX^t X \) with \( M = I - X^t y(y^t X X y)^{-1} y^t X \); this is actually the deflating of \( X^t X \) with information correlated to the response \( y \) via \( X^t y \). Readers are referred to [16] for additional information on OFM.

Hence, the application of an OFM procedure, as is the case in SR and Target Projection, may result in biased variance estimates. This bias is a direct consequence of the rotation, particularly the regression in equation (4) for the updated TP loading. The orthogonal
variance decomposition (via the regression) is necessary for OFM to ensure the orthogonal property of the weights and the scores in multiple OFM orthogonal components. However, the orthogonal variance decomposition via the regression is not necessary for the assessment of variable importance, and as demonstrated earlier, it introduces a bias with respect to the obtained variances as a consequence of the rotation property of the basic sequence.

Basic sequence and relation to VIP

Since the VIP in equation (23) includes the weight vectors \( w_{jk} \), we can explore the weight vector of a PLS component, \( a \). The weight vector \( w_a \) can indeed be expressed as using the basic sequence \( B_a^{[1]}(\hat{b}^{OLS}_a, X_a) \) by a simple substitution of the OLS regression coefficients \( \hat{b}^{OLS}_a = (X'_a X_a)^{-1}X'_a y_a \) in equation (13).

\[
B_a^{[1]}(\hat{b}^{OLS}_a, X_a) = \frac{(X'_a X_a)\hat{b}^{OLS}_a}{\|X'_a X_a\|\|\hat{b}^{OLS}_a\|} = \frac{x'_a y_a}{\|x'_a y_a\|} = w_a
\]

This indicates that a similar effect in orthogonal filtering can be observed for the VIP where the weight vector can rotate towards the dominant eigenvector of \( X^T X \). Hence, \( w_{jk} \) may contain irrelevant information due to the basic rotation effect. In VIP, the rotation has a higher impact compared with SR since the weight vectors of all components are used and hence multiple rotations are applied in VIP. As indicated earlier, a lower signal to noise ratio for a higher component leads to a faster rotation speed and a higher rotational impact to the VIP. Thus the VIP can concurrently result in both higher false positive and false negative rates in cases where irrelevant variation is dominant in the dataset or in a complex model where many latent variables are needed.

Significance multivariate correlation – (SMC)

With a better understanding of the current variable selection methods in light of the basic sequence, we developed a new method called SMC to correct for the rotation effect in Target Projection for variable selection purposes. The key points in SMC are to estimate for each variable the correct sources of variability resulting from the PLS regression (i.e. regression variance and residual variance), and use them for statistically determining a variables importance with respect to the regression model. For the estimation of variances, SMC uses the combination of the vector of the predicted values, \( \hat{y} \), as a new latent (score) vector of the PLS model in equation (15) and the regression coefficient vector, see equation (15). However, dissimilar to with the Target Projection procedure, SMC discards the orthogonal variance decomposition in equation (4) to prevent the influence of non-relevant information contained in \( X \). Hence without this rotation, the normalized regression vector is used as the loading vector as shown in equation (14). The reconstruction of \( X \) can be represented in equation (16), without the actual regression, within the score (or predicted response) space.
\[ p_{SMC} = \frac{\hat{b}_{p_{LS}}}{\|\hat{b}_{p_{LS}}\|} \]

\[ t_{SMC} = Xp_{SMC} = X \frac{\hat{b}_{p_{LS}}}{\|\hat{b}_{p_{LS}}\|} \quad \text{as such} \quad \hat{y} = \frac{\hat{y} \hat{b}_{p_{LS}}}{\|\hat{b}_{p_{LS}}\|} \]

\[ X = t_{SMC}p_{SMC}^{t} + E_{SMC} = \frac{(\hat{y} \hat{b}_{p_{LS}})}{\|\hat{b}_{p_{LS}}\|^2} + E_{SMC} \]

Or for each variable in a form of regression of \( x_l \) on \( t_{SMC} \):

\[ x_l = t_{SMC}p_{SMC}^{t} + e_{SMC} = \frac{(\hat{y} \hat{b}_{p_{LS}})}{\|\hat{b}_{p_{LS}}\|^2} + e_{SMC} \]

Lacking the actual regression step (of \( X \) on \( t_{SMC} \)) in equations (16)-(17) as the SR method in equation (4), SMC is not a complete basic rotation and the explained variance (or regression variance) \( \|t_{SMC}p_{SMC}^{t}\|^2 \) in equation (18) may not be orthogonal to the estimated residual variance \( \|e_{SMC}\|^2 \) in equation (19) however it reflects the relevant variation in the predicted response projected back onto the original X-variable space via the PLS regression vector.

\[ SS_{t_{PLS, regression}} = \|t_{SMC}p_{SMC}^{t}\|^2 = \left\| \frac{(\hat{y} \hat{b}_{p_{LS}})}{\|\hat{b}_{p_{LS}}\|^2} \right\|^2 \]

\[ SS_{t_{PLS, residual}} = \|e_{SMC}\|^2 = \left\| x_l - \frac{(\hat{y} \hat{b}_{p_{LS}})}{\|\hat{b}_{p_{LS}}\|^2} \right\|^2 \]

In SR the ratio of the two variance terms, the explained variance and residual variance in the form of sums-of-squares (SS), is compared with an F-distribution with \( n - 2 \) and \( n - 3 \) degrees of freedom. Those variables with F-values that exceed the F-test critical threshold value (as determined via the choice of significance level) are considered important variables. This choice is referred as the test of equality of variance of two (independent) sample populations, chapter 5.4 in [17]. However, in both SR and SMC, these two variance terms are obtained in the form of individual regressions of each X-variable onto the common score vector, with the loadings as the regression coefficients in equation (5) for SR and equation (17) for SMC, respectively. Therefore the Analysis of Variance (ANOVA) test for the significance of the regression is most appropriate, (see chapter 8 in [17]) and as such, the proper degrees freedom associated with the regression sums-of-squares in (6) and (18) is 1; whereas for the residual variance in equation (19) it is \( n - 2 \) with \( n \) being the number of samples. We should note that in the ANOVA the F-test is carried out using the mean-squared error, which are the raw sums of squares divided by the appropriate degrees of freedom.

Hence, SR values can be affected not only by the rotation towards the dominant eigenvector of \( X^tX \) mentioned earlier, but also due to the improper use of the F-test via the SR defined degrees of freedom. The latter issue seems dominant and results in very
low SR values that can result in very conservative estimates of variable importance, wherein a short list of selected variables is obtained.

In SMC, an F-test is also used to assess variables which are statistically significant with respect to their relationship (regression) to $y$, but for an $sMC_t$ test value an F-distribution with 1 numerator and n-2 denominator degrees of freedom is used, $F(1-\alpha, 1,n-2)$, where $\alpha$ is the chosen significance level.

\begin{align}
MS_{l,PLS\_regression} &= SS_{l,PLS\_regression} / 1 \\
MS_{l,PLS\_residual} &= SS_{l,PLS\_residual} / (n - 2) \\
SMC_t &= \frac{MS_{l,PLS\_regression}}{MS_{l,PLS\_residual}} \left( \frac{\|\tilde{y}_{p,PLS}\|^2}{\|p_{PLS}\|^2} \right)^2 \left( \frac{\|x_i - \tilde{y}_{p,PLS}\|^2}{\|x_i\|^2} \right) / (n-2)
\end{align}

We should note that the impact to the threshold value for the SMC test is dependent on the actual rotation effect. As stated above, due to the lack of the regression step (orthogonal variance decomposition) in equations (16)-(17), the variance decomposition is biased in that the $\tilde{y}_{p,PLS} + SS_{l,PLS\_residual}$ may not always be equal to the total sum of squares $\|x_i\|^2$ and will only do so when the basic rotation effect is negligible with $p_{SMC} \approx p_{TP}$.

When the basic rotation effect is negligible, the residual variance is orthogonal to the explained variance and the false positive rate associated with a non-parametric Null distribution should approximate those of the theoretical Null, 0.05, for the stated degrees of freedom. Otherwise, when the rotation effect is non-negligible in the presence of large irrelevant variation from non-informative variables in the dataset, the regression in equation (4) for SR (or in equation (13) for the weight vector in VIP) by equalizing the decomposed variances with the total variance in $X'X$ may positively bias the explained variance estimated from PLS and SMC for non-informative variables. This rotation impact is removed in SMC. Since in this case the estimated residual variance is not orthogonal to the explained variance the ratio of both mean squares $SMC_t$ may not follow an F-distribution, the false positive rate in SMC can potentially be lower than the theoretical Null at an $\alpha$ of 0.05 due to the correctly estimated variances being applied to non-informative variables.

Note that the SMC method can be directly applied to PLS2 with multiple response variables by simply applying SMC to each regression coefficient vector for each response variable.

The Matlab implementation of the proposed SMC method is provided as supplementary material. The algorithm assumes that the data is provided centered and scaled (optional).
Experimental

Permutation experiments

To further demonstrate the impact to the rotation when the basic rotation effect is negligible vs. non-negligible, and quantify the false positive rate via the use of permutation tests with respect to the SMC, additional simulations were performed. The simulated datasets consisted of one response variable and 30 predictors ($X$) variables with 15 non-informative variables and 15 important variables with varying degrees of signal-to-noise ratios (SNR). A total of 1000 simulation run replicates were run in order to estimate the false positive rate. The PLS model was applied using cross-validation for determination of model rank.

The first simulation was run with moderate irrelevant variation in the correlated non-informative variables and a SNR = 0.25 with respect to the important variables. Table 1 illustrates the angle between the different vectors of interest. One can observe that the TP loading, $p_{TP}$, is rotated towards the largest eigenvector of $X'X$, which is dominated by irrelevant variation, as compared to the SMC (which is proportional to PLS regression coefficients), while at the same time moving away from the OLS regression coefficients, which itself quantifies the largest degree of correlation of the $X$ variables to the response. Given the lower magnitude of rotation towards the irrelevant variation in the largest eigenvector of $X'X$ and a smaller angle towards the OLS regression coefficients (Beta OLS), we can conclude that the SMC loading is capturing the most relevant variation in $X$ with respect to $y$, which corresponds to the performance of the PLS model.

Table 1: Angles between the studied vectors (1st PCA loading, TP loading, SMC loading, PLS regression coefficients, and OLS)

<table>
<thead>
<tr>
<th></th>
<th>1st PCA Loading</th>
<th>TP loading ($p_{TP}$)</th>
<th>SMC loading (or Beta PLS)</th>
<th>Beta OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st PCA Loading</td>
<td>83</td>
<td>85</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>TP loading ($p_{TP}$)</td>
<td>20</td>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMC loading (or Beta PLS)</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The second simulation varied the number of correlated variables from 0 to 15 in increments of 5, while also varying the signal-to-noise ratio from .25 to 1 in increments of .25. The observed permutation test false positive rate with respect to the SMC is shown in Figure 1. We can see that the false positive rate for non-informative variables (as per the non-parametric Null distribution) ranges between 0.04 – 0.06. In the moderate signal-to-noise level (of 25%), and without the correlation between non-informative
variables, the false positive rate for non-informative variables (as per the non-parametric Null distribution) approximates 0.05. It is interesting to note that there is a downward trend (lower false positive rate) as the number of correlated (non-informative) variables is increased. This positive effect is due to the rotation effect in SMC as highlighted in Table 1, wherein more information pertaining to $X'X$ is being captured in the residual vector of non-informative correlated variables. Simultaneously the explained variance is relatively stable for these variables. Hence the ratio of explained over unexplained variances is lowered, helping reduce the number of false positives. The trend is less at higher signal-to-noise ratios as the noise level is reduced and simultaneously more relevant information being captured in the explained variance.

![Figure 1: Permutation test false positive rate with respect to the SMC](image)

**DESCRIPTION OF DATASETS**

**SIMULATION**

The synthetic data used to compare the variable selection methods has two objectives. It must be built in a simple way, yet it must also reproduce characteristics encountered in modern measurements, such as in metabolomics. Various blocks composed this. Each block relates to different (independent) responses. One response $y$ corresponds to a smaller part of the $X$ variance than other sources of influence and is only related to the first variables. The objective is to model only that particular response. This situation would correspond to a case where one would want to model e.g. the impact of a disease in omics data, which is principally influenced by numerous other parameters such as age, genetic background or diet. Moreover, each block can be decomposed into purely uninformative variables and variables correlated to the different responses. Figure 2 presents the correlation structure used to simulate data matrix $X$. Only the first block of 19 variables is related to the response of interest, $y$. A second block of variables (30 to 49) follows a similar structure but related to another (orthogonal) response. Finally, 30 additional variables containing only Gaussian noise are added to the data. This covariance structure is then used as the input for the ADICOV method [18] to generate 1000 samples.
Figure 2: Correlation of each variable to the response of interest $y$, and correlation matrix of the data $X$. The 19 first variables related to $y$ while the rest of the data contain one unrelated structure and purely random noise.

NIR SHOOT OUT DATA

The data set used in this study is known as the Chambersburg Shoot-out 2002 data set. This set contains transmission spectra of pharmaceutical tablets recorded on Foss/NIRSystems Multitab Spectrometers instrument over the spectral region from 600 to 1898 nm with 2 nm increments on the wavelength scale. In this study, the 155 NIR spectra from Calibrate1 were used and the spectra are showed in Figure 4. More details about these spectra are available [19,20].

Results

SIMULATION DATASET

Based on the covariance structure of the synthetic data, we produced a thousand samples and divided them between training and test sets. The optimization of the global PLS model relied on cross validation, resulting in a PLS model based on 9 LVs. We can notice that the model is relatively complex for this simulated data. This complexity can be explained by the rotation of latent variable toward the non-relevant group of information.
at each PLS component rather than focusing on the designed correlated component structure in $X$, hence additional components are needed to compensate for this.

The application of the 4 variable selection methods permitted us to generate 4 new PLS models.

The selection obtained using all approaches is provided in Figure 3 together with the list of variables that should have a relation to the predicted response $y$. Selected variables are represented in white while non-selected ones are left black. Here variables 1 to 19 were expected to be selected. Note that the correlation of variable 1 to $y$ is equal to 0.95 while variable 19 is only correlated to $y$ with a correlation coefficient of 0.05.

![Variables selected by different methods](image)

**Figure 3:** Correlation of the variables to the response of interest $y$ compared with the variables selection obtained using VIP, SR, cSR, SMC and Beta CI.

All methods efficiently detect the most relevant variables, however some interesting discrepancies can be observed. First VIP selected a larger number of variables including variables unrelated to $y$. These false positives were selected both in the structured part of the data (but unrelated to $y$) and in the purely noisy variables (30 last variables) indicating that VIP tends to focus on the non-informative $X$-variance due to the effect of multiple rotations applied to the weight vectors from a model with multiple components, nine in this case. Beta CI exhibited a similar behavior here with 17 false positives. The main disadvantage of Beta CI is that it is more intensive in terms of computation however this results seems to indicate that the false (positive) discovery rate of Beta CI can be higher than expected given the level of significance (0.05) chosen. It may due to the additional
uncertainty of bootstrapping method; however, a further investigation of this issue is not within the scope of this article.

In terms of false positive, SR and SMC were clearly the best methods: none were selected. For the case of SMC, as stated earlier in the theory section, this was expected due to the positive rotation effect where the basic rotation impact is removed in SMC and hence the rates of false positive in SMC can be lower than the theoretical Null for the stated degrees of freedom at an $\alpha = 0.05$. This agrees with the results obtained from the permutation experiments in the theory section. In term of the rotation issue in SR, as discussed earlier, SR and SMC are relatively close to each other form a theoretical point of view. The solution proposed by SR is a rotation of the SMC solution towards the dominant source of variance. In this simulation the dominant sources of variation were in the unrelated structures represented in variables 30 to 90. It is however hard to visualize the effect in the final SR values due to the same reason of the degrees of freedom. Instead, the effect of this rotation can be seen in the explained variance with SMC and SR, as shown in Figure 4. The rotation for SMC to the SR solution induces the inclusion of irrelevant $X$-variance. Given this observation one would expect SR to select all relevant variables and a number of false positives. The result displayed in Figure 2 contradicts this intuition. Indeed a second factor comes into play. The number of degrees of freedom in the F-test used in SR leads to a very low SR values and hence a very conservative decision. The false positives are thereby excluded from the selection yet this come with the cost of excluding also relevant variables. The number of false negatives is de facto the main disadvantage of SR.

![Figure 4: Comparison of the sums-of-squares of the explained variance in SMC and SR.](image)
The advantage of the simulation was that we could assess the quality of the selection using a gold standard. However, one could wonder how these results translate when the SMC is applied to real data.

In term of model predictive performance comparison with the models on the selected variables from different methods, Table 2 regroups the numbers of selected variables, the complexity of the models and their Root Mean Square Error of Prediction (RMSEP). SR was very conservative but led to some improvement of the prediction. However VIP and Beta CI were all selecting a larger subset of variables, but often associated with more complex models (larger number of latent variables). In this simulation, being relatively simple, the error terms were very comparable: SMC yielded the best performance (and the simplest model) here followed by SR and Beta CI. SMC. However one could estimate that this simulation is too simplistic, thereby the differences in RMSEP are not meaningful. The gain in prediction on a real dataset is demonstrated in the second example. The main objective of the simulation is to evaluate the variable(s) selected from a qualitative point of view. In other words, are we selecting the correct variables?

Table 2: Complexity and performance of the different PLS models respectively constructed on the full simulated data or on selected subsets of variables

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of variables</th>
<th>Number of LVs</th>
<th>RMSEP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All variables</td>
<td>90</td>
<td>9</td>
<td>7.1</td>
</tr>
<tr>
<td>VIP</td>
<td>26</td>
<td>6</td>
<td>3.8</td>
</tr>
<tr>
<td>SR</td>
<td>4</td>
<td>3</td>
<td>3.1</td>
</tr>
<tr>
<td>SMC</td>
<td>17</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td>Beta CI</td>
<td>34</td>
<td>5</td>
<td>4.2</td>
</tr>
</tbody>
</table>

NIR SHOOT OUT DATA

The complete data was first pre-processed using Savitsky Golay smoothing (windows of 11 points, 2nd derivative) and Standard Normal Variate (SNV). Subsequently 155 NIR spectra were used to construct PLS models able to predict the concentration in the active product ingredient (API), the active compound of pharmaceutical tablets.

The five approaches discussed earlier were applied to this model, and 5 new PLS models were constructed. The models were then compared using an independent test set consisting of 460 spectra of the same type of tablets measured on the same spectrophotometer. The number of LVs used in each model was determined by cross-validation. Here all subsets led to models based on 4 LVs which simplify the comparison.

As the model predictive performance comparison (Table 3) using models built on the selected variables from different methods shows, the VIP selected a relatively small number of variables but the prediction power remains similar as the one obtained using the complete dataset. SR and SMC all led to a better prediction on the test set yet SMC
distinguish itself with an error twice as low. This indicates that the most complete important variable list is obtained by SMC by maintaining the best prediction in PLS.

Table 3: Complexity and performance of the different PLS models respectively constructed on the full NIR data or on selected subsets of variables

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of variables</th>
<th>Number of LVs</th>
<th>RMSEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>All variables</td>
<td>650</td>
<td>7</td>
<td>5.10</td>
</tr>
<tr>
<td>VIP</td>
<td>118</td>
<td>4</td>
<td>5.35</td>
</tr>
<tr>
<td>SR</td>
<td>297</td>
<td>4</td>
<td>2.58</td>
</tr>
<tr>
<td>SMC</td>
<td>418</td>
<td>4</td>
<td>2.39</td>
</tr>
<tr>
<td>Beta CI</td>
<td>421</td>
<td>4</td>
<td>4.71</td>
</tr>
</tbody>
</table>

The results can also be examined from a qualitative point of view. The NIR spectra (Figure 5A), used here contain a region typically characterized as noisy and uninformative, especially the region above 1800 nm. The application of SMC eliminates “only” 232 variables out of 650. However, it is interesting to note that the region above 1800 nm was completely discarded. From a visual inspection one can easily establish that the excluded region contains mostly irrelevant and noisy information. SR eliminated also this region but selected also fewer variables in the other spectral regions. This is in line with our previous results showing that SR tends to lead to false negative i.e. to discard relevant variables.

For illustration of the rotation effect, in this NIR data the dominant sources of variation are clearly allocated in the region above 1800 nm. The effect of this rotation can be seen in the explained variance with SMC and SR, as shown in Figure 5B where the explained variance for variables in this region is much higher for SR due to the regression in equation (4). However, this information is not properly captured in SR value with the use of n-2 degrees of freedom instead of 1. Therefore, in most cases, SR tends to underestimate the number of important variables. The rotation has more impact to VIP which confirms the behavior obtained with the simulation data. Both VIP and Beta CI methods select many (false positive) variables from this region. In the case of Beta CI, it may due to the extra uncertainty added by the bootstrapping method.
Figure 5: A) NIR calibration data, the variables selected by SMC, VIP, SR and Beta CI are respectively represented by black crosses, red circles, blue right pointing triangles and purple left pointing triangles. B) Explained variance associated with SR and SMC.

This second example demonstrates on real data that the application of SMC permits the selection of the relevant subset of variables by minimizing both false negative and false positive. This would obviously have an impact of the interpretation of the results. In some applications, the interpretation is less relevant than the predictive power; here SMC also distinguishes itself since the model constructed on the SMC selected the subset of
relevant variables that led to the best performing model indicating the completeness of the selected informative variables.

**Discussion and Conclusion**

The interpretation of variable importance using only parameters calculated from the original PLS model has the advantages of being fast and easy to compute. However, the interpretation of the variables based on the existing model parameters remains challenging due to the complexity of the PLS model and the lack of complete understanding of the properties of PLS.

Besides the advantages of the filter methods such as VIP and SR, it has been seen that defining a reliable threshold is the main limitation of this variable selection category. This limitation is formally illustrated and explained by the basic sequence theory for PLS in this work.

We demonstrate that orthogonal filtering based methods, such as Target Projection, can be fully explained using the basic sequence with the properties of the Krylov sequence (or power method) i.e. the rotation towards the dominant eigenvector. The consequence of the resulting rotation is the incorporation of irrelevant variation from $X$ that can have an adverse impact in the evaluation of a variable’s importance. This impact is not trivial and with a high rotation magnitude it can heavily influence either the false positive or false negative rates. Our conclusion is that this variance orthogonal decomposition as the essential element of the basic rotation is not necessary for the assessment of variable importance.

Given the above, we developed a novel approach: Significant Multivariate Correlation (SMC). The underlying concept of SMC permits the maximum use of the information obtained from the basic sequence proposed in this work for identifying important variables in PLS. Variables exhibiting minimal bias (with respect to their parameter estimation), and statistically significant in the model, are best highlighted by the SMC method. Moreover the variables are ranked based on their respective F-values with a defined significant threshold value.

Hence, in the framework of filter methods category, SMC method expects to be the best-in-class in providing a more reliable list and ranking of variable importance. For more discussion and comparison of this class to other variable selection methods from other categories i.e. wrapper and embedded, readers are referred to [6,7].

Application of SMC to simulated and real data (NIR shootout) illustrates the outstanding properties of SMC as compared to several commonly used variable selection methods such as a VIP, Selectivity Ratio (SR) and Beta CI. In all tested cases, SMC provides the most optimal variable list, with minimal false negative and false positive errors, and improves the predictive performance of the PLS model due to the bias reduction property of the variable selection method.
Chapter 5

It has also been demonstrated that both SMC and SR are more reliable for datasets with noisy variables when compared to the VIP. However, SR is often too conservative due to the improper use of the F-test on the regression and residual variances. In all cases, both VIP and SR may suffer from non-relevant variance with respect to $y$ due to the unwanted rotation effect in the calculation of the weight vectors in VIP and the TP loading.
Chapter 6: Variable Importance in PLS in the Presence of Autocorrelated Data

Abstract

An integral part of interpreting atypical process performance in manufacturing processes is a multivariate understanding of process parameters and their relationship to a product’s critical quality attributes. In this endeavor, Partial Least Squares (PLS) has greatly advanced the analysis of data that exhibits a high level of multicollinearity, but has not fully explored the impact to important variable selection in the presence of autocorrelation, particularly in the residuals, wherein a current observation is correlated to some degree with the previous observation(s). This autocorrelation provides an additional challenge to understanding model performance and important variable selection. This paper introduces an autocorrelation correction factor formulation to PLS in an attempt to address this concern and illustrates its application to the recently proposed Significant Multivariate Correlation (SMC) variable selection method. Our results demonstrate that the correction factor formulation presented in this paper has the desired effect of driving down the false positive rate when applied to the SMC.

N.L. Afanador, T.N. Tran, and L.M.C. Buydens, Chemometrics and Intelligent Laboratory Systems, (n.d.).
Introduction

Partial Least Squares (PLS) has gained popularity within the manufacturing industry for its ability to relate a large number of correlated explanatory variables to a response via a multivariate model. This has proved helpful in determining the important variables responsible for product quality changes. However, different manufacturing processes will exhibit varying degrees of autocorrelation wherein a current process measurement is correlated with previous process measurements. This generally occurs in manufacturing due to similar operation conditions taking place, such as the use of the same operators and raw materials. Within a modeling context, this autocorrelation can manifest itself to some degree in the residuals resulting from a PLS analysis. When this occurs the assumption of independence is violated and any statistical hypothesis tests will be biased. Consequently, the identification of important variables responsible for changes in process performance can prove challenging. While the problem of autocorrelated data has been explored with respect to process monitoring and control [1, 2, 3], our proposed strategy varies from these methods in two ways. The first is that we correct for autocorrelation after the PLS model has been applied to the original raw data and focus on correcting for any autocorrelation in the \( x_i \)-residuals, via the \( MSE_{x_i} \). Secondly, we attempt to quantify the degree of autocorrelation for each variable’s \( x_i \)-residuals separately, and make an adjustment based on the degree of statistically significant autocorrelation that is present. Recently, the Significant Multivariate Correlation (SMC) methodology was introduced [4] and showed favorable results as compared to other commonly used important variable determination methods: Variable Importance in the Projection (VIP), Selectivity Ratio and the use of bootstrap confidence intervals for the PLS regression coefficients (Beta CI). The goal of this study is to introduce and illustrate the impact of an autocorrelation correction factor in the important variable selection properties of the SMC, both with and without autocorrelation, while varying signal-to-noise ratio. Our results demonstrate that the use of the correction factor formulation presented in this paper as an autocorrelation adjustment applied to the \( MSE_{x_i} \) obtained from the SMC has the desired effect of driving down the false positive rate. We would like to note that the autocorrelation factor formulation presented in this paper can also be applied to other methods that employ a mean-squared-error term for either hypothesis tests or calculating confidence intervals.

Methods

Autocorrelation

As previously stated, often-times manufacturing processes will exhibit varying degrees of autocorrelation wherein a current process measurement is correlated to some degree with previous process measurements. This autocorrelation is not limited to processes measurements themselves, but can also be transmitted to the \( x_i \)-residuals that result from
a PLS model. The structure of this autocorrelation can oftentimes be classified as a first order autoregressive process (AR1), which presupposes that the contribution of process disturbances that occurred in the past are small in comparison to current process disturbances [5]. When this condition is reasonably met we can assume an infinite set of weights in descending order that are similar in form to an exponential decay pattern [5].

The weights of these disturbances are then noted as $\phi^k$, for $k = 0, 1, 2, \ldots$, where $k$ stands for the $k^{th}$ lag. Given this general description, an AR(1) process in terms of an $x_t$-residual, $u_t$, can then be formally defined per Eq. (1).

\[(1) \quad u_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \cdots = \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-k}\]

From Eq. (1), we can also have Eq. (2) for the variable at time point $i-k$.

\[(2) \quad u_{i-1} = \varepsilon_{i-1} + \phi \varepsilon_{i-2} + \phi^2 \varepsilon_{i-3} \ldots\]

Equation’s (1) and (2) can be combined for the error at time-point $i$ as the function of the error time-point $i-k$ (3)

\[(3) \quad u_t = \varepsilon_t + \phi u_{t-1}\]

from which we can see that an AR(1) process is simply the regression of $u_t$ on $u_{t-1}$ [5]. The assumption of independence can be assumed to be violated if the correlation resulting from this regression is statistically significant for some pre-stated confidence level.

For manufacturing process data the determination of whether the assumption of independence has been violated is generally made via the use of the autocorrelation function (ACF) [5] wherein the correlation between observations at varying time lags, $k$, is both calculated, Eq. (4), and displayed graphically (Figure 1).

\[(4) \quad \rho_u = \text{Cor}(u_i, u_{i+k})\]

Hence, the assumption of independence can be assumed to be violated if any of the calculated correlations is significantly different from zero, where the statistical limits displayed in Figure 1 are determined via Eq. (5).

\[(5) \quad (LCI, UCI) = 0 \pm \left( \frac{z_{0.975}}{\sqrt{N}} \right)\]
Autocorrelation Correction Factor

There have been several approaches recommended for addressing statistically significant autocorrelation that generally involve some transformation of the original $X$-variables [6], [7], [8]. For this study we have chosen the Corrected OLS correction factor [9] due to its simplicity in application and demonstrably good results in improving the coverage error for confidence intervals in the presence of autocorrelation [9], without the need to resort to a transformation of the original variables. A good approximation for this correction factor, $cf$, for moderate to large sample sizes is shown in terms of the population parameters in Eq. (6) as applied to an estimate of the $x_i$-residual variance, where $MSE_{x_i} = Var(u_i)$, resulting from some PLS model (please refer to Appendix A for the derivation of Eq. (6) [5], [10]),

$$
\sigma_i^2 = var(u_i) \left( \frac{1+\rho}{1-\rho} \right), \quad cf = \left( \frac{1+\rho}{1-\rho} \right)
$$

This correction helps approximate the true $x_i$-residual variance, $\sigma^2$, for a process that exhibits an AR(1) autocorrelation structure.

In this study we have opted for maximizing the information obtained from the ACF by calculating the plug-in estimate for the population parameter, $\rho$, as a weighted average, $\hat{\rho}^*$, of the statistically significant autocorrelations (7).

$$
\hat{\rho}^* = \frac{\sum_{k=1}^{n-1} w_k \hat{\rho}_k}{\sum_{k=1}^{n-1} w_k}, \quad w_k = \left( 1 - \frac{k}{n} \right)
$$
We should note that the above correction factor is not recommended in combination with resampling approaches, such as the bootstrap and jackknife, for estimation of standard errors for either the PLS regression coefficients or the VIP. A naïve bootstrapping approach disrupts the underlying autocorrelation structure so that the ability to calculate the correction factor via resampling in conjunction with a resampled ACF becomes extremely difficult in practice.

**Significant multivariate correlation (SMC)**

The concept underlying SMC is briefly described below with a more detailed methodological and theoretical exposition in [4]. The key points in SMC are to estimate for each variable the correct sources of variability resulting from the PLS regression, and use them for statistically determining a variables importance with respect to the PLS regression model. For the estimation of variances, SMC uses the regression coefficient vector Eq. (8) to define the co-variance between the $X$-variables and the response variable $y$ in combination with the predicted values, $\hat{y}$, to define a new latent score vector of the PLS model in Eq. (9) and Eq. (10).

(8) $\hat{b}_{pls} = RP'(X'X)^{-1}X'y = RP\hat{b}_{OLS}$, $R = (P'W)^{-1}$

(9) $p_{SMC} = \|\hat{b}_{pls}\|

(10) $t_{SMC} = Xp_{SMC} = X\frac{\hat{b}_{pls}}{\|\hat{b}_{pls}\|} = \frac{\hat{y}}{\|\hat{b}_{pls}\|}$

The reconstruction of $X$ can be represented in (11), without the actual regression, within the score (or predicted response) space [4].

(14) $X = t_{SMC}p_{SMC}^t + E_{SMC} = \left(\frac{\hat{y}b_{pls}^t}{\|\hat{b}_{pls}\|^2}\right) + E_{SMC}$

An F-test is then used to assess variables that are statistically significant with respect to their relationship to $y$, Eq. (15) – Eq. (19) [4].

(15) $SS_{t,PLS\_regression} = \|t_{SMC}p_{SMC}^t\|^2 = \left(\frac{\hat{y}b_{pls}^t}{\|\hat{b}_{pls}\|^2}\right)^2$

(16) $SS_{t,PLS\_residual} = \|e_{SMC}\|^2 = \left\|x_i - \left(\frac{\hat{y}b_{pls}^t}{\|\hat{b}_{pls}\|^2}\right)\right\|^2$

(17) $MS_{t,PLS\_regression} = SS_{t,PLS\_regression}/1$

(18) $MS_{t,PLS\_residual} = SS_{t,PLS\_residual}/(n - 2)$
Experimental

Experimental Structure

This study focuses on quantifying the sensitivity and specificity of each method via simulation experiments. Two additional datasets, obtained from an actual manufacturing process, will also be included. General agreement in the results of both the simulation experiments and actual process dataset will provide guidance on which approach is most suitable for assigning variable importance in a manufacturing process in the presence of autocorrelation. In order to maintain continuity with the work referenced in [4], the performance of the adjusted SMC was assessed against the following: unadjusted SMC [4], Selectivity Ratio [11, 12], Variable Importance in the Projection (VIP) [13], Bootstrap VIP [14], and Bootstrap $\tilde{B}_{pls}$ [4].

Simulated Manufacturing Process Experimental Section

For the purpose of this study a first order autoregressive process that exhibits positive autocorrelation was simulated; as such, we will be correcting for a downward bias in the $MSE_x$ with respect to the SMC. The simulation experiment consists of a manufacturing process composed of 70 process variables, $X$, 60 process cases, and a single response variable, $y$. Among the 70 process variables, 40 variables were designated as being highly cross-correlated noise variables that are not correlated to $y$, 15 variables were designated as independent noise variables that are not cross-correlated to each other or to $y$, and 15 variables were designated as important variables cross-correlated to each other and to $y$. All of the 70 process variables, $X$, and $y$ exhibit one of the following three levels of autocorrelation generated from a first order autoregressive process: 0.1 (low autocorrelation), 0.5 (moderate autocorrelation), and 0.9 (high autocorrelation). We can see from Figure 2 the on-average level of autocorrelation that is being simulated for each of the aforementioned autoregressive levels. Within each of these autocorrelation levels, the signal-to-noise ratio between the response and the important variables was varied as follows: 0.25, 0.50, 0.75, and 1.00. This reveals if any important variable selection method was heavily influenced by a particular autocorrelation structure in combination with a specific signal-to-noise ratio. One-thousand experimental replications of each factorial combination were performed, and the average of these results was used to assess the performance of each method. The assessment was based on the approach outlined in [13] wherein the geometric mean of both sensitivity and specificity is assessed via the metric $G$ (25). Hence, the greater a methods $G$ value, the better that method did in identifying the variables induced to correlate with the response. For the calculation of (20) Sensitivity is defined as the number of true important variables identified by each method, divided by the total number of actual important variables. Specificity was
measured as the number of true unimportant variables identified by each method, divided by the total number of actual unimportant variables.

\[ G = \sqrt{\text{Sensitivity} \times \text{Specificity}} \]
ADPN data set with $n = 57$, $y = 1$, $p = 100$ (training dataset)
ADPN data set with $n = 14$, $y = 1$, $p = 100$ (test dataset)

VACCINE

The VACCINE dataset results from the manufacture of a vaccine where the response is specified as “yield”, which shows an increase as time passes. The explanatory variables in this dataset represent variables known to potentially impact cellular growth that viral propagation and can result in the observed yield increase. It can be assumed that there is a root cause for the observed increase in yield and that the identification of this root cause can provide increased process knowledge that may be directly linked to better process control. The test set was selected as a random sample of $n = 16$ from the overall VACCINE dataset. An example of the autocorrelation structure for one randomly sampled predictor variable and the response variable are shown in Figures 3 and 4, respectively. In the training set approximately the same auto-correlation structure is maintained as that observed in the overall dataset. In addition, the observed shift in the overall dataset is adequately represented. In the test set, due to the much smaller sample size, the original autocorrelation structure is dampened, but the observed shift in the overall dataset is adequately represented. This approach allows us to monitor the impact of the correction factor in the adjusted SMC via the training set RMSEP, wherein the adjusted SMC is predicted to perform much better than the unadjusted SMC. In the case of the test set, we are able to assess the performance of the correction factor when the auto-correlation structure is dampened, but the original process shift is retained. In this latter approach, the adjusted SMC is predicted to also perform better than the unadjusted SMC given the reduced number of instances where the correction is applied, but in which the process shift is still maintained.

- VACCINE dataset with $n = 50$, $y = 1$, $p = 67$ (training dataset)
- VACCINE dataset with $n = 16$, $y = 1$, $p = 67$ (test dataset)

Figure 3 – ACF of two randomly sample X-variables (First 10 lags) with 95% Confidence Limits: An example of the autocorrelation structure for one randomly sampled predictor variable from both the ADPN and VACCINE dataset, respectively.
Figure 4 – ACF of y variables for the manufacturing process dataset (First 10 lags) with 95% Confidence Limits: The autocorrelation structure for the response variable for both the ADPN and VACCINE dataset, respectively.

Results

MANUFACTURING PROCESS SIMULATION

The results from the manufacturing process simulation are outlined in Table 1 for a signal-to-noise ratio equal to 0.75, wherein the results for $G$, Sensitivity, and Specificity represent the rounded average across the 1000 experimental replicates for each autocorrelation and signal-to-noise ratio combination. It can be observed from this table that the Significant Multivariate Correlation (SMC), adjusted for autocorrelation, is consistently favored among all the variable importance methods assessed. A similar trend was observed for other autocorrelation and SNR combinations, wherein as the level of autocorrelation increased the adjusted SMC was constantly favored.

One thing to note with respect to the VIP is that it can be expected to perform well at the highest signal-to-noise ratio and low auto-correlation (AR = 0.1) combination. In this case the VIP will simply reduce to the weights from the first LV and as such will do a good job of correctly assigning important variables. At a lower signal-to-noise ratio and low auto-correlation (AR = 0.1) combination the $\hat{b}_{pls}$ can be expected to retain the favorable property of being a good estimate of the strength of the correlation between the important variables and the response, $y$, whereas the VIP, along with the Selectivity Ratio, will tend to incorporate more information pertaining to $X$. In all instances of low autocorrelation both the unadjusted SMC and adjusted SMC performed as well, or better, as the other methods assessed. These results are aligned with the findings from our previous work as detailed in [4].
Table 1 – Comparison between different methods for SNR = 0.75

<table>
<thead>
<tr>
<th>Method</th>
<th>AR = .1</th>
<th>AR = 0.5</th>
<th>AR = .9</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjusted SMC</td>
<td>0.99</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>SMC</td>
<td>0.98</td>
<td>0.97</td>
<td>0.88</td>
</tr>
<tr>
<td>Selectivity Ratio</td>
<td>0.10</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>VIP</td>
<td>0.97</td>
<td>0.94</td>
<td>0.86</td>
</tr>
<tr>
<td>VIP Bootstrap</td>
<td>0.93</td>
<td>0.91</td>
<td>0.76</td>
</tr>
<tr>
<td>Bootstrap $\hat{b}_{pls}$</td>
<td>0.94</td>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 2 – Results from the Analysis of Manufacturing Process Data

**PROCESS DATASET**

The results from the analysis of two actual process datasets are outlined in Table 2. The results of this analysis demonstrate that the SMC, adjusted for autocorrelation, was the favored method in terms of the number of important variables chosen in combination with its corresponding RMSEP and Test Set RMSEP. It is important to note that although the RMSEP and Test Set RMSEP are in some instances are very close to each other across the various methods, the adjusted SMC was able to achieve lower values with a fewer number of important variables.

### ADPN

<table>
<thead>
<tr>
<th>Method</th>
<th>No of Important Variables Chosen</th>
<th>RMSEP</th>
<th>Test Set RMSEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjusted SMC</td>
<td>11</td>
<td>2.50</td>
<td>2.75</td>
</tr>
<tr>
<td>SMC</td>
<td>22</td>
<td>1.70</td>
<td>2.64</td>
</tr>
<tr>
<td>Selectivity Ratio</td>
<td>0</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>VIP</td>
<td>40</td>
<td>2.16</td>
<td>2.94</td>
</tr>
<tr>
<td>VIP Bootstrap</td>
<td>21</td>
<td>2.43</td>
<td>3.25</td>
</tr>
<tr>
<td>Bootstrap $\hat{b}_{pls}$</td>
<td>56</td>
<td>1.59</td>
<td>2.61</td>
</tr>
</tbody>
</table>

### VACCINE

<table>
<thead>
<tr>
<th>Method</th>
<th>No of Important Variables Chosen</th>
<th>CV R2</th>
<th>Test Set R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjusted SMC</td>
<td>16</td>
<td>1306.23</td>
<td>1249.41</td>
</tr>
<tr>
<td>SMC</td>
<td>18</td>
<td>1352.68</td>
<td>1341.44</td>
</tr>
<tr>
<td>Selectivity Ratio</td>
<td>0</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>VIP</td>
<td>24</td>
<td>1473.24</td>
<td>1253.16</td>
</tr>
<tr>
<td>VIP Bootstrap</td>
<td>16</td>
<td>1381.73</td>
<td>1445.44</td>
</tr>
<tr>
<td>Bootstrap $\hat{b}_{pls}$</td>
<td>32</td>
<td>1365.58</td>
<td>1321.23</td>
</tr>
</tbody>
</table>
Discussion and Conclusion

In both the simulation study and actual process datasets the evaluation metrics point to the adjusted SMC as the most consistent method for identifying the most important variables. These results help validate the joint use of the weighted average of the autocorrelation coefficients and the correction factor as an adjustment formulation for data known to be first order autocorrelated. Furthermore, although not detailed in this work, aside from improving the overall final list of important variables, it can be expected that a more robust ranking of important variables will result due to the autocorrelation adjustment via the resulting F-statistic for each variable. We should further note that the findings from the work in [4], as applied to new data sets, further validate the predicted performance improvement in important variable determination via the use of the SMC.

In the end, it is the conclusion of the authors that when analyzing data known to be highly to moderately first order autocorrelated, as is common in manufacturing processes, the SMC, in conjunction with the correction factor formulation presented in this work, will result in a more consistent determination of the important variables driving a process.
Appendix A
This appendix presents the theoretical derivation of the autocorrelation factor, \( \left( \frac{1+\rho}{1-\rho} \right) \), presented in this work, as it applies to the SMC.

(1) \[ x = t_{SMC} p'_{SMC} + u_{SMC_i} = \frac{\gamma \hat{b}'_{PLS_i}}{\| \hat{b}_{PLS_i} \|^2} + u_{SMC_i} \]

(2) \[ u_{SMC_i} = \epsilon_{SMC_i} + \phi u_{SMC_{i-1}} \]

(3) \[ MSE_{x_i} = Var(u_{SMC_i}) = \sigma^2_{SMC} = \sigma^2_{\epsilon_{SMC}} \left( \frac{1}{1-\phi^2} \right) \]

(4) \[ Cov(u_{SMC_i}, u_{SMC_{i+k}}) = \sigma^2_{\epsilon_{SMC}} \phi^k \left( \frac{1}{1-\phi^2} \right) \]

(5) \[ \rho^k = \frac{Cov(u_{SMC_i}, u_{SMC_{i+k}})}{\sqrt{Var(u_{SMC_i})} \sqrt{Var(u_{SMC_{i+k}})}} = \phi, \text{ for } k = 1 \]

Where \( x \) is a process variable, \( i \) the \( i^{th} \) residual, and \( k = k^{th} \) lag.

It can now be shown that the variance, with respect to the Central Limit Theorem for dependent processes, converges to an unbiased estimate of \( \sigma^2_{\epsilon_{SMC}} \) when the correction factor \( \left( \frac{1+\rho}{1-\rho} \right) \) is taken into account.

(6) \[ \sigma^2_{\epsilon_{SMC}} = Var(u_{SMC_i}) + \frac{2}{n} \sum_{k=1}^{n-1} (n-k) Cov(u_{SMC_i}, u_{SMC_{i+k}}) \]

(7) \[ \rho^k Var(u_{SMC_i}) = Cov(u_{SMC_i}, u_{SMC_{i+k}}) \]

(8) \[ \sigma^2_{\epsilon_{SMC}} = Var(u_{SMC_i}) + \frac{2}{n} \sum_{k=1}^{n-1} (n-k) \rho^k Var(u_{SMC_i}) \]

(9) \[ \sigma^2_{\epsilon_{SMC}} = Var(u_{SMC_i}) \left( 1 + 2 \sum_{k=1}^{n-1} \rho^k - \frac{2}{n} \sum_{k=1}^{n-1} k \rho^k \right) \]

where \[ \frac{2}{n} \sum_{k=1}^{n-1} k \rho^k = \frac{\rho \left( 1 + (n-1) \rho - n \rho^{n-1} \right)}{(1-\rho)^2} \rightarrow 0 \text{ and } 2 \sum_{k=1}^{n-1} \rho^k \sim \frac{2\rho}{1-\rho} \text{ as } k \rightarrow \infty, \text{ as } k \rightarrow \infty \]

(10) \[ \sigma^2_{\epsilon_{SMC}} = Var(u_{SMC_i}) \left( 1 + \frac{2\rho}{1-\rho} \right) \]

(11) \[ \sigma^2_{\epsilon_{SMC}} = Var(u_{SMC_i}) \left( \frac{1+\rho}{1-\rho} \right) \]
Chapter 6

References


Chapter 7: Conclusion

The objective of this research was the statistical evaluation of variable selection methodologies in Partial Least Squares regression, with the goal of identifying a method that can be generally applied across a multitude of problems, in the presence of autocorrelation.

This research began with the investigation of the use of the 95% lower-bound on the jackknife confidence interval for the VIP, versus the combination of bootstrap and permutation methods for determining variable importance [Chapter 2]. Although not a statistically derived cut-off value, the PLS-VIP '>1' rule was used to assess which approach performed best in detecting important variables in actual manufacturing process data sets. In this study we found that the use of the 95% lower-bound on the jackknife confidence interval for the VIP resulted in a less optimal subset of important variables when compared to a bootstrapping approach that incorporated permutation tests. In spite of this result, we were able to conclude that the use of the jackknife confidence interval approach is sufficient, though not optimal, for ensuring that the important subset of important variables is being selected.

Given the approximating nature of the Jackknife with respect to the bootstrap procedure the question arose as to which uncertainty estimation method provided the best list of important variables currently driving a process [Chapter 3]. Once again, the PLS-VIP '>1' rule was adopted and used to assess seven different uncertainty estimation methods, along with their resulting ability for detecting important variables in simulated and actual process data sets. Our work showed that a general bootstrap approach for estimating confidence intervals resulted in more positive metrics, in the absence of permutation tests. However, not all the bootstrap approaches had the same positive performance. In this study both the Normal and Basic confidence interval approaches performed best. It should be noted, that the resampling strategy presented in this research may be valid for other PLS parameters, such as loadings, scores, and regression coefficients. Additional research should be ensued in order to verify which bootstrapping uncertainty approach best estimates the sampling distribution of the aforementioned PLS parameters.

This favorability of the Normal and Basic confidence interval approaches was predominantly attributed to the inherent bias-correcting properties within these methods. This bias-correction appeared to mitigate against the incorporation into the VIP of variability present in an explanatory variable that may no longer be meaningful in explaining \( y \). Since the VIP includes the weight vectors \( \mathbf{w}_{jk} \) a similar effect to orthogonal filtering can be observed in the VIP. As such, as more information from \( \mathbf{X} \) is incorporated into the VIP, via the weight vectors, there is an increased false negative rate in the important variables, and an increased false positive rate in the un-important variables.
In light of this knowledge, a novel approach named Significant Multivariate Correlation (SMC), was developed. The underlying concept of SMC permits the maximum use of the information obtained from the basic rotation for identifying important variables in PLS [Chapter 4]. An F-test can then be used to assess a variables importance and ranking. Variables exhibiting minimal bias with respect to their parameter estimates, and statistically significant in the model, are correctly highlighted by the SMC method. Simulation and application of the SMC to two real data sets (NIR and NMR metabolomics) illustrated the outstanding properties of the SMC as compared with several commonly used variable selection methods such as the VIP and SR [Chapter 5].

Given that the SMC, and other similar methods, rely on an F-test for determining and ranking important variables, it will be very sensitive to the presence of autocorrelation in manufacturing process data. This autocorrelation can result in the violation of the independence assumption associated with the error estimate used for statistical hypothesis tests, specifically with the F-test used in the SMC for determining a variables importance. As such, as an extension of the aforementioned work, and to correct for the phenomenon of autocorrelation present in batch manufacturing data, we explored an autocorrelation correction factor that uses a weighted average of statistically significant lags, as determined from the autocorrelation function [Chapter 6]. The SMC, along with three different variable selection methods were presented, along with their resulting ability for detecting important variables in the presence of autocorrelation in both simulated and actual datasets. This work concluded that when analyzing data known to be high to moderately autocorrelated, as is common in manufacturing processes, the SMC, in conjunction with the correction factor formulation presented in this work, will result in a more consistent determination of the important variables driving a process.

In summary, this research resulted in a better understanding of resampling based methods for determining the uncertainty in the estimation of PLS parameters and the VIP, a better understanding of the orthogonal correction in the PLS algorithm, a new variable importance method (SMC) capable of providing a statistical assessment of the most important variables currently driving a manufacturing process, and an autocorrelation adjustment formulation for the SMC to better control the increased false positive rate associated with autocorrelated data.

Given the importance of properly understanding manufacturing process performance, especially in the context of pharmaceutical manufacturing where there is a vested public health interest, it is important that further research on methods that improve process understanding is continued. The Chemometrics community has done an excellent job in developing and advancing methods that deal with the problem of multicollinearity in process data. It is our hope that this research prompts more work to better understand, adjust, and correct for the impact of autocorrelation in latent based regression methods and their interpretation.
The objective of this research was the statistical evaluation of variable selection methodologies in Partial Least Squares regression, with the goal of identifying a method that can be generally applied across a multitude of problems, in the presence of autocorrelation.

This research began with a study of how the use of bootstrapping, in conjunction with permutation tests, can provide avenues for improving the selection of variables responsible for manufacturing process changes via the Variable Importance in the Projection (PLS-VIP) statistic. In this study we found that the use of the 95% lower-bound on the jackknife confidence interval for the VIP resulted in a less optimal subset of important variables when compared to a bootstrapping approach that incorporated permutation tests. In spite of this result, we were able to conclude that the use of the jackknife confidence interval approach is sufficient, though not optimal, for ensuring that the important subset of important variables is being selected.

This result lead to a performance assessment of seven resampling based methods of uncertainty estimation with the goal of assessing which method performs best in reducing the false positive rate, while at the same time not impacting the true positive rate. This research concluded that the normal bootstrap confidence intervals, followed by the basic bootstrap confidence intervals, provided the most consistent coverage for the purpose of determining a variable’s importance. A consequence of this research was a heightened awareness of the degree to which a variable’s importance may be either upgraded or downgraded as model complexity increases. This is found to be due to the VIP including the weight vectors ($w_{jk}$), calculated via a sequence of residual matrices, resulting in more information from $X$ being incorporated into the VIP. As such, there is an increased false negative rate in the important variables, and an increased false positive rate in the un-important variables.

In order to address this issue a new method was introduced, called the Significant Multivariate Correlation (SMC), for statistically assessing variable importance for PLS regression and classification. An evaluation of the performance of the SMC on simulated and real data sets demonstrating its exceptional performance over several commonly used methods is also presented. Given that the SMC, and other similar methods, rely on an F-test for determining and ranking important variables, it will be very sensitive to the presence of autocorrelation in manufacturing process data. This autocorrelation can result in the violation of the independence assumption associated with the error estimate used for statistical hypothesis tests, specifically with the F-test used in the SMC for determining a variable’s importance. As such, as an extension of the aforementioned work, and to correct for the phenomenon of autocorrelation present in batch manufacturing data, we successfully explored an autocorrelation correction formulation that improved the performance of the SMC in the presence of autocorrelation.
Samenvatting

Het doel van dit onderzoek was de statistische evaluatie van methodologieën voor selectie van variabelen in Partial Least Squares regressie, met als doel een methode te identificeren die algemeen kan worden toegepast bij een groot aantal problemen waarbij autocorrelatie aanwezig is.

Dit onderzoek begon met een studie van hoe het gebruik van bootstrapping, in combinatie met permutatietests, mogelijkheden kan bieden voor het verbeteren van de selectie van variabelen die verantwoordelijk zijn voor wijzigingen in productieprocessen door middel van de Variable Importance in Projection (PLS-VIP) schatter. In deze studie ontdekten we dat het gebruik van de 95% ondergrens van het jackknife betrouwbaarheidsinterval voor VIP resulteerde in een minder goede deelverzameling van belangrijke variabelen dan een bootstrapping-aanpak met permutatietests. Ondanks deze resultaten konden we concluderen dat de aanpak met het jackknife betrouwbaarheidsinterval, hoewel niet optimaal, volstaat om zeker te kunnen zijn dat de belangrijke subset van belangrijke variabelen wordt geselecteerd.

Dit resultaat leidde tot een prestatiebeoordeling van zeven op resampling gebaseerde methoden van onzekerheidsschatting met als doel te beoordelen welke methode het best is in het verminderen van de false positive rate, maar tegelijkertijd geen invloed heeft op de true positive rate. Uit het onderzoek bleek dat de normal bootstrap-betrouwbareihdsintervalen gevolgd door de basic bootstrap-betrouwbareihdsintervalen, de meest consistente dekking geven voor het bepalen van het belang van een variabele. Een gevolg van dit onderzoek was een toegenomen besef van de mate waarin het belang van een variabele verhoogd danwel verlaagd kan worden als de complexiteit van het model toeneemt. Dit blijkt het gevolg te zijn van het feit dat de VIP de gewichtsvectoren ($w_{jk}$), berekend door middel van een reeks residumatrices, meeneemt, waardoor meer informatie over X in de VIP wordt opgenomen. Als zodanig is er een verhoogde false negative rate in de belangrijke variabelen, en een verhoogde false positive rate in de niet-belangrijke variabelen.

Om dit probleem aan te pakken werd een nieuwe methode geïntroduceerd, genaamd Significant Multivariate Correlation (SMC), om het belang van een variabele voor PLS-regressie en -classificatie statistisch te beoordelen. Een evaluatie van de prestatie van de SMC op gesimuleerde en echte datasets die de uitzonderlijke prestaties van deze methode opzichte van een aantal veelgebruikte methoden antoonde, wordt ook gepresenteerd. Gegeven dat de SMC en andere soortgelijke methoden afhankelijk zijn van een F-test om belangrijke variabelen te bepalen en te rangschikken, zal de methode zeer gevoelig zijn voor de aanwezigheid van autocorrelatie in productieprocesdata. Deze autocorrelatie kan leiden tot schending van de veronderstelling van onafhankelijkheid die samenhangt met de schatting van de fout die gebruikt wordt voor statistische hypothesetests, in het bijzonder de F-test in de SMC voor het bepalen van het belang van een variabele. Als zodanig, in het verlengde van het voornoemde werk en voor het corrigeren van het
fenomeen van autocorrelatie aanwezig in batchproductiedata, hebben we met succes een formulering van correctie voor autocorrelatie onderzocht die de prestaties van de SMC in aanwezigheid van autocorrelatie verbetert.
Curriculum Vitae

Nelson Lee Afanador obtained his bachelor’s degree in biology from Temple University in Philadelphia, PA USA in 2001. He began work at Merck, Sharp, & Dohme that same year and currently works in its Center for Mathematical Sciences. His careers focus has predominantly been in the area of vaccine manufacturing science, technology, and data analysis. While at Merck he earned a master’s degree in Applied Statistics from West Chester University, West Chester, PA USA. From 2011 – 2014 he pursued a PhD in Chemometrics from the Analytical Chemistry Department at Radboud University Nijmegen, The Netherlands, specializing in important variable selection in Partial Least Squares regression. He has had the privilege of presenting at multiple conferences, both in the USA and Europe, on the subject of multivariate data analysis.