Chemometrics and Intelligent Laboratory Systems 137 (2014) 162–172

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

An assessment of the jackknife and bootstrap procedures on uncertainty estimation in the variable importance in the projection metric

N.L. Afanador [a,c,*], T.N. Tran [b], L.M.C. Buydens [c]

[a] Center for Mathematical Sciences, Merck, Sharp, & Dohme, West Point, PA, USA
[b] Center forMathematical Sciences, Merck, Sharp, & Dohme, Oss, Netherlands
[c] Institute for Molecules and Materials, Analytical Chemistry, Radboud University Nijmegen, Netherlands

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form 24 May 2014
Accepted 30 May 2014
Available online 6 June 2014

Keywords:
Partial least squares
Bootstrap
Jackknife
Variable importance in the projection

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.

1. 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 >1 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.

* Corresponding author at: Center for Mathematical Sciences, Merck Sharp & Dohme Corp., WP97-B208, 770 Sumneytown Pike, PO Box 4, West Point, PA, USA.
E-mail address: nelson_afanador@merck.com (N.L. Afanador).

http://dx.doi.org/10.1016/j.chemolab.2014.05.016
0169-7439/

2. Materials and methods

2.1. 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].

$$X = TP' + E \tag{1}$$

$$y = Tc + f \tag{2}$$

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, $c(h \times 1)$ are the PLS regression coefficients, and where $E(n \times p)$ and $f(n \times 1)$ are the residuals 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, and w_k, stand for the k-th column of T, P, and W, respectively ($k = 1, 2, \ldots, h$).

$$w_k = X'_{(k)} y_{(k)} / \left\| X'_{(k)} y_{(k)} \right\| \tag{3}$$

$$t_k = X_{(k)} w_k \tag{4}$$

$$\hat{c}_k = t'_k y_{(k)} / t'_k t_k \tag{5}$$

$$p_k = X'_{(k)} t_k / t'_k t_k \tag{6}$$

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

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

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, is obtained. This step is determined by the data analyst and is often supported by the use of cross-validation.

2.2. PLS-VIP

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

$$\mathrm{VIP}_j = \sqrt{p \sum_{k=1}^{h} \left(\hat{c}_k^2 t'_k t_k \right) \left(w_{(kj)} \right)^2 / \sum_{k=1}^{h} \hat{c}_k^2 t'_k t_k} \tag{9}$$

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 kth dimension $\left(\hat{c}_k^2 t'_k t_k \right)$. 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 variable's importance and ranking. The scientific notation used for the confidence interval methods is as follows:

Notation for confidence interval methods

θ	population parameter
$\hat{\theta}$	sample estimate of θ
$\hat{\sigma}$	sample estimate of the population parameter σ
$\hat{\theta}_{(i)}$	jackknife *replicate* estimate of $\hat{\theta}$ with the ith observation removed
$\hat{\theta}_{(\bullet)}$	jackknife estimate of $\hat{\theta}$ across all jackknife replicates
$\hat{\sigma}_J$	jackknife estimate of $\hat{\sigma}$
$\hat{B}^*_J$	jackknife estimate of bias
$\hat{\theta}^*_b$	bootstrap *replicate* sample estimate of $\hat{\theta}$
$\hat{\theta}^{*\bullet}$	bootstrap estimate of $\hat{\theta}$ across all bootstrap replicates
$\hat{\sigma}_B$	bootstrap estimate of $\hat{\sigma}$
$\hat{B}^*$	bootstrap estimate of bias
α	statistical significance level

2.3. 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 ith observation from the dataset. It then calculates the overall jackknife estimate of the parameter, $\hat{\theta}_{(\bullet)}$, by taking the average of the aforementioned replicate estimates (Eq. (10)). An estimate for the standard deviation of said statistic, $\hat{\sigma}_J$, can then be calculated using both $\hat{\theta}_{(\bullet)}$ and the replicates from the re-computations, $\hat{\theta}_{(i)}$ (Eq. (11)) [7].

$$\hat{\theta}_{(\bullet)} = (1/n) \sum_{i=1}^{n} \hat{\theta}_{(i)} \tag{10}$$

$$\hat{\sigma}_J = \left[((n-1)/n) \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)} \right)^2 \right]^{1/2} \tag{11}$$

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

$$(\theta_\alpha, \theta_{1-\alpha}) = \hat{\theta} \pm t_{(1-\alpha/2, n-1)} \hat{\sigma}_J \tag{12}$$

An additional property of the jackknife procedure is that it allows for the estimation of bias between the current estimate and the target