ELSEVIER

Chemometrics and Intelligent Laboratory Systems 64 (2002) 79–89

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Model selection for partial least squares regression

Baibing Li *, Julian Morris, Elaine B. Martin

Centre for Process Analytics and Control Technology, University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK

Received 9 August 2001; accepted 24 June 2002

Abstract

Partial least squares (PLS) regression is a powerful and frequently applied technique in multivariate statistical process control when the process variables are highly correlated. Selection of the number of latent variables to build a representative model is an important issue. A metric frequently used by chemometricians for the determination of the number of latent variables is that of Wold's *R* criterion, whilst more recently a number of statisticians have advocated the use of Akaike Information Criterion (AIC). In this paper, a comparison between Wold's *R* criterion and AIC for the selection of the number of latent variables to include in a PLS model that will form the basis of a multivariate statistical process control representation is undertaken based on a simulation study. It is shown that neither Wold's *R* criterion nor AIC exhibit satisfactory performance. This is in contrast to the adjusted Wold's *R* criteria which is shown to demonstrate satisfactory performance in terms of the number of times the known true model is selected. Two industrial applications are then used to demonstrate the methodology. The first relates to the modelling of a product quality using data from an industrial fluidised bed reactor and the second focuses on an industrial NIR data set. The results are consistent with those of the simulation studies.

Keywords: Akaike Information Criterion (AIC); Cross-validation; Model selection; Multivariate statistical process control; Partial least squares regression; Wold's *R* criterion

1. Introduction

For industrial-scale processes where a large number of process variables are monitored, one of the more commonly applied modeling techniques for the development of a multivariate statistical process control (MSPC) nominal representation is partial least squares (PLS) [1,2]. The objective of PLS is to define a set of latent variables through the projection of the process and quality spaces onto new orthogonal subspaces, by maximising the covariance between the two spaces [3,4]. The latent variables are defined as linear combinations of the original variables. Although as many latent variables as min(*N*,*M*) can be calculated, where *N* is the sample size and *M* is the number of process variables, it is conjectured that the lower order latent variables are associated with process noise and should be excluded from the model. Therefore to remove the noise, a criterion is required for selecting the number of latent variables to include in the PLS model.

A wealth of approaches have been cited in the literature to select the number of variables to

* Corresponding author. Fax: +44-1912225748.
E-mail address: baibing.li@ncl.ac.uk (B. Li).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(02)00051-5

include in a model, including Akaike Information Criterion (AIC), Final Prediction Error Criterion (FPE), Bayesian Information Criterion (BIC), Law of Iterated Logarithms Criterion (LILC), Normalised Residuals Sum of Squares (NRSS), Multiple Correlation Coefficient (R^2), Adjusted Multiple Correlation Coefficient, (${R_a}^2$), Overall F-test of the Loss Function (OVF) and Mallow's statistic (C_p). A review of these model order selection criteria can be found in Haber and Unbenhauen [5].

Within the chemometric literature, Wold's R criterion, which is based on cross-validation [6,7], has been the typical approach used to select the number of latent variables. Wold's R criterion is based on calculating the ratio of the PRedicted Error Sum of Squares (PRESS) for latent variable $(m+1)$ and latent variable, m. Inclusion of latent variables into the model terminates at m when the ratio exceeds unity [7]. More recently statisticians [8–10] have advocated the use of Akaike Information Criterion (AIC) [11]. AIC is based on a trade-off between accuracy and model parsimony. One advantage of AIC is that of computational costs since, as noted by [12,13], cross-validation is a time-consuming procedure. The question therefore arises, which approach is more appropriate for PLS model selection in terms of its application in multivariate statistical process control (MSPC)?

To evaluate the capabilities of different criteria, simulated models allow the underlying structures of the models to be known [14,15]. Practical case studies, as described in Refs. [12,16], although important, cannot provide quantitative assessments. In this paper a comparison is performed between Wold's R criterion and AIC for PLS model selection using simulation models that generate multi-collinear process data. A further criteria examined is that of the adjusted Wold's R criteria which adopts different thresholds from the Wold's R criterion.

In Section 2, the PLS algorithm and a number of criteria for PLS model selection are briefly summarized. In Section 3, the simulation models are described before an analysis of the simulation results is undertaken in Section 4. In Sections 5 and 6, two practical examples using industrial data are examined and finally a number of conclusions are drawn in Section 7.

2. PLS algorithm and model selection criteria

2.1. Summary of the PLS algorithm

Consider a data set representing the "normal" operating conditions of a process. $\mathbf{X}_{N\times M}$ represents the data matrix of process variables and $\mathbf{Y}_{N\times K}$ the data matrix of quality (response) variables, which are recorded for N time points. The objective of linear PLS is to project the data down onto a number of latent variables, say $\mathbf{t}_j$ and $\mathbf{u}_j$ ($j=1,\ldots,A$), where A is the number of the latent variables, and then to develop a regression model between $\mathbf{t}_j$ and $\mathbf{u}_j$:

$$\mathbf{u}_j = b_j\mathbf{t}_j + \mathbf{e}_j \qquad j = 1,\ldots,A \tag{1}$$

where $\mathbf{e}_j$ is a vector of errors and b_j is an unknown parameter estimated by $\hat{b}_j = (\mathbf{t}_j^T\mathbf{t}_j)^{-1}\ \mathbf{t}_j^T\mathbf{u}_j$. The latent variables are computed by $\mathbf{t}_j=\mathbf{X}_j\mathbf{w}_j$ and $\mathbf{u}_j=\mathbf{Y}_j\mathbf{q}_j$, where both $\mathbf{w}_j$ and $\mathbf{q}_j$ have unit length and are determined by maximizing the covariance between $\mathbf{t}_j$ and $\mathbf{u}_j$. $\mathbf{X}_{j+1}=\mathbf{X}_j-\mathbf{t}_j\mathbf{p}_j^T$ where $\mathbf{X}_1=\mathbf{X}$ and $\mathbf{p}_j=\mathbf{X}_j^T\mathbf{t}_j/(\mathbf{t}_j^T\mathbf{t}_j)$ and $\mathbf{Y}_{j+1}=\mathbf{Y}_j-\hat{b}_j\mathbf{t}_j\mathbf{q}_j^T$ where $\mathbf{Y}_1=\mathbf{Y}$.

Letting $\hat{\mathbf{u}}_j=\hat{b}_j\mathbf{t}_j$ be the prediction of $\mathbf{u}_j$, the matrices $\mathbf{X}$ and $\mathbf{Y}$ can be decomposed as the sum of the following outer products:

$$\mathbf{X} = \sum_{j=1}^{A}\mathbf{t}_j\mathbf{p}_j^T + \mathbf{E} \text{ and } \mathbf{Y} = \sum_{j=1}^{A}\hat{\mathbf{u}}_j\mathbf{q}_j^T + \mathbf{F} \tag{2}$$

where $\mathbf{E}$ and $\mathbf{F}$ are the residuals of $\mathbf{X}$ and $\mathbf{Y}$ after extracting the first A pairs of latent variables.

In PLS regression, each pair of latent variables, $\mathbf{t}_j$ and $\mathbf{u}_j$ ($j=1,\ldots,A$) is sequentially extracted through an iterative procedure, the basis of which is the NIPALS algorithm [3,4]. The only issue remaining to be addressed is how to determine the number of latent variables, A.

2.2. Cross-validation based criteria

Wold's R criterion is based on cross-validation. In cross-validation the data, $\mathbf{X}$ and $\mathbf{Y}$, is split into a number of blocks, k, and a one latent variable model is built from $(k-1)$ blocks of data. Based on this one latent variable model, the excluded block is used for testing and an individual PRESS is calculated. This procedure is repeated excluding each block, once and only once, and then the total PRESS is calculated for one latent variable by summing the individual PRESS values. This procedure is then repeated for 2, 3, ..., min(N,M) latent variables and a series of PRESS values are obtained [6,7].

Several criteria have been proposed to identify the appropriate number of latent variables to include in the model. Two of the more commonly applied approaches include using that number of latent variables that give the minimum PRESS, or alternatively Wold's R criterion [7]. Wold's R criterion is given as $R = \mathrm{PRESS}(m+1)/\mathrm{PRESS}(m)$ where $\mathrm{PRESS}(m)$ denotes the PRESS after including the first m latent variables. Wold's R criterion terminates when R is greater than the given threshold, unity [7], and hence $A = m$. Osten [12] showed that selecting the absolute minimum PRESS had particularly poor statistical properties and suggested using a criterion based on the first local minimum in the PRESS, cf. Wold's R criterion. In addition, Osten [12] proposed an F-test based criterion, where the F value is given by:

$$F = \frac{\mathrm{PRESS}(m) - \mathrm{PRESS}(m+1)}{K} \bigg/ \frac{\mathrm{PRESS}(m+1)}{NK - (m+1)K} \quad (3)$$

The above criterion is compared against an F value, $F_{K,NK-(m+1)K,\ 0.95}$.

Besides Wold's R criterion and Osten's F criteria, the adjusted Wold's R criterion is also considered in the subsequent comparison. Instead of adopting unity as a threshold as in Wold's R criterion, the adjusted Wold's R criterion uses 0.95 and 0.90 as thresholds due to sampling variability as stated in Krzanowski [17]. These are denoted by R(0.95) and R(0.90), respectively. The adjusted R criteria states that an additional latent variable will not be included in the PLS model unless it provides significantly better predictions. A similar threshold was adopted by Krzanowski [17]. For the W statistic, Eastment and Krzanowski [14] originally selected unity as a threshold, whilst Krzanowski [17] proposed adopting 0.9 as a threshold due to sampling variability.

2.3. Akaike Information Criterion

One of the most frequently used criteria in system modelling and system identification is that of Akaike's Information Criterion (AIC) [11]. For problems associated with a single response variable, i.e. $K = 1$, the information criterion is given by:

$$\mathrm{AIC}(m) = N\log(\hat{\sigma}^2) + 2m \quad (4)$$

where m is the number of model parameters, N is the sample size and $\hat{\sigma}^2$ is the maximum likelihood estimate of the variance of the response variable, Σ. The right-hand side of Eq. (4) consists of two terms, the first, $N\ \log(\hat{\sigma}^2)$, represents model accuracy and the second, $2m$, relates to model parsimony. It should be noted that an arbitrary constant may be added to the right-hand of Eq. (4). For example since $\hat{\sigma}^2 = \mathrm{RSS}/N$, where RSS is the residual sum of squares, Eq. (4) can be re-written as:

$$\mathrm{AIC}(m) = -N\log(N) + N\log(\mathrm{RSS}) + 2m \quad (5)$$

or else:

$$\mathrm{AIC}(m) = N\log(\mathrm{RSS}) + 2m \quad (6)$$

Both versions of AIC, Eqs. (5) and (6), are equivalent in terms of model selection since selection is based on the difference between the values of AIC between two candidate models [11,18]. The multivariate version of AIC was given by Bedrick and Tsai [15]:

$$\begin{aligned}\mathrm{MAIC}(m) = {} & N(\log|\hat{\Sigma}| + K) \\ & + 2d[Km + K(K+1)/2] \end{aligned} \quad (7)$$

where $d = N/[N - (m + K + 1)]$ and $\hat{\Sigma}$ is the maximum likelihood estimate of Σ. This is a corrected version of the multivariate AIC for the small sample case. When sample sizes are large, $d = N/[N - (m + K + 1)] \approx 1$, and thus Eq. (7) may be further simplified.

Akaike [19] investigated an alternative multivariate version of AIC for factor analysis. Here the assumption of the 'response variables' in the linear model in Ref. [19] was not based on conditional normality as had previously been the case. The derived multivariate version of AIC [19] therefore differs from Eq. (7) as given in Ref. [15]. In this paper, the version of AIC in Ref. [19] is not considered.

Besides AIC, a number of alternative information criteria exist, such as those developed by Hannan and Quinn [20] and Schwarz [21]. However attention in this paper focuses on AIC as it has been one of the more frequently applied approaches by statisticians for PLS model selection in recent years.

3. Simulation models

A simulation study was undertaken to evaluate the performance of different criteria for PLS model selection. In the subsequent study, simulation models were first developed from which data was generated. Then different criteria were applied for model selection. The resulting models were then compared with the true models and finally an evaluation of the different criteria for PLS model selection was made through a comparison of the frequency as to which the true model was selected [14,15].

The framework for the simulation models was based on the study of Næs and Martens [22] for the problem of a single response variable. The Næs–Martens' framework is extended in this paper to the situation where there exist multiple response variables. For each A^* ($2 \leq A^* \leq 4$), where A^* is the true number of latent variables, the X-block data, **X**, with sample size N were generated as:

$$\mathbf{X} = \sum_{i=1}^{A^*} \mathbf{r}_i \boldsymbol{\xi}_i^{\mathrm{T}} + \tilde{\mathbf{E}} \tag{8}$$

where $\tilde{\mathbf{E}}=[\mathbf{e}_1, \ldots, \mathbf{e}_6]$ and $\mathbf{e}_j$ ($j=1,\ldots,6$) are simulated as mutually independent normal variables, e_j ($j=1,\ldots,6$), with zero-mean and $\mathrm{var}(e_j)=0.01$. $\mathbf{r}_i$ ($i=1,2,3,4$) were generated from mutually independent normal variables, r_i ($i=1,2,3,4$) with zero-mean and $\mathrm{var}(r_1)=10$, $\mathrm{var}(r_2)=5$, $\mathrm{var}(r_3)=2$ and $\mathrm{var}(r_4)=0.5$ and $\boldsymbol{\xi}_i$ are given by:

$$\boldsymbol{\xi}_1 = [1,1,1,1,1,1]^T/6^{1/2}, \qquad \boldsymbol{\xi}_2 = [0.5,0.5,-1,0.5,0.5,-1]^T/3^{1/2},$$

$$\boldsymbol{\xi}_3 = [1,1,1,-1,-1,-1]^T/6^{1/2}, \qquad \boldsymbol{\xi}_4 = [-1,1,0,-1,1,0]^T/2$$

It is noted that $\mathrm{var}(r_1)+\mathrm{var}(e_j)=10.01$ is the largest eigenvalue of cov(**X**) and is approximately 1000 times as large as the smallest eigenvalue. The Y-block data, **Y**, were generated as:

$$\mathbf{Y} = \sum_{i=1}^{A^*} \mathbf{z}_i \boldsymbol{\eta}_{A^*_i}^{\mathrm{T}} + \boldsymbol{\Psi} = \sum_{i=1}^{A^*} \mathbf{r}_i \boldsymbol{\eta}_{A^*_i}^{\mathrm{T}} + \tilde{\mathbf{F}}_{A^*} \tag{9}$$

where $\boldsymbol{\Psi}=[\psi_1,\ldots,\psi_{A^*}]$ was generated by an $A^* \times 1$ random vector with a multivariate normal distribution $N(\mathbf{0}, \mathbf{S})$ where $\mathbf{S}=\sigma^2[(1-\lambda)\mathbf{I}+\lambda\mathbf{1}\mathbf{1}^T]$ and $\sigma^2=0.001$ and $\lambda=0.6$. **I** is an identity matrix, **1** is a vector of unity and $\tilde{\mathbf{F}} = \sum_{i=1}^{A^*} \mathbf{f}_i \boldsymbol{\eta}_i^{\mathrm{T}} + \boldsymbol{\Psi}$ is a noise matrix and $\boldsymbol{\eta}_i$ is given by:

$$\eta_{21} = [1,2,1]^T/6^{1/2}, \quad \eta_{22} = [0,1,-2]^T/5^{1/2}$$

$$\eta_{31} = [1,2,1]^T/6^{1/2}, \quad \eta_{32} = [0,1,-2]^T/5^{1/2},$$

$$\eta_{33} = [-5,2,1]^T/30^{1/2}$$

$$\eta_{41} = [1,2,1,0]^T/6^{1/2}, \quad \eta_{42} = [0,1,-2,1]^T/6^{1/2},$$

$$\eta_{43} = [0,1,-2,-5]^T/30^{1/2},$$

$$\eta_{44} = [-5,2,1,0]^T/30^{1/2}$$

The matrix $\mathbf{Z}=[\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4]$ was constructed as:

$$\mathbf{z}_i = \mathbf{r}_i + \mathbf{f}_i \tag{10}$$

where $\mathbf{f}_i$ ($i=1, 2, 3, 4$) were generated as independent normal variables, f_i ($i=1, 2, 3, 4$), with zero-mean and $\mathrm{var}(f_1)=0.25$, $\mathrm{var}(f_2)=0.125$, $\mathrm{var}(f_3)=0.05$ and $\mathrm{var}(f_4)=0.0125$. It is noted that both $\{\boldsymbol{\xi}_i\}$ and $\{\boldsymbol{\eta}_{A^*i}\}$ are normalized orthogonal vector series, and r_i ($i=1,2,3,4$) are mutually independent random variables. Hence, comparing Eq. (2) with Eqs. (8) and (9), it can be concluded that latent variable $\mathbf{t}_i$, loading vectors $\mathbf{p}_i$ and $\mathbf{q}_i$ obtained from the PLS algorithm are approximately equal to $\mathbf{r}_i$, ξ_i and η_{A*i} $(i=1,\ldots, A^*)$ respectively. The Y-block data, **Y**, of the response variables then essentially depends on $\mathbf{r}_i$, $(i=1,\ldots, A^*)$, plus noise. This means that the theoretical value of the number of latent variable is equal to A^*.

4. Simulation results

4.1. Comparison of different criteria

For a fixed sample size, $N=1000$ and a fixed number of blocks, $k=5$, for cross-validation, the performances of a number of different criteria were investigated. **X** and **Y** were generated as 1000×6 and $1000 \times H$ data matrices respectively, where $H=3$ if $A^*=2$ or 3, and $H=4$ if $A^*=4$. To reduce random fluctuations, 1000 simulation experiments were performed.

The relative cumulative variances captured by the six latent variables for the X and Y blocks, averaged over the 1000 simulation experiments, are given in Table 1. It can be seen from the first two rows of Table 1 that on average, for $A^*=4$, the first four latent variables capture 100% and 99.94% of the variance in the **X** and **Y** data sets, respectively. The additional two latent variables do not contribute to explaining any of the variability in the data. This verifies the theoretical value of the number of latent variables, $A^*=4$.

Table 2 summarizes the frequencies of the selected numbers of latent variables. The first five rows give the frequencies of the selected models after applying different criteria to the PLS algorithm, when for the true model, the number of latent variables is $A^*=4$. It can be seen that in this case neither Wold's R criterion nor AIC give satisfactory performance. Of the 1000 experiments, in 490 cases Wold's R selects four latent variables. AIC exhibits better performance than Wold's R criterion, giving the right value in 75.8% of the time. In comparison Osten's F criterion is better than either Wold's R criterion or AIC. It gave the right number of latent variables 85.4% of the time. In contrast, both the adjusted Wold's $R(0.90)$ and $R(0.95)$ criteria gave the true value 100% of the time.

Table 1
Relative cumulative variances of **X** and **Y** for $A^*=2$, 3 and 4 (sample size, $N=1000$; number of blocks, $k=5$)

True models	Blocks	Number of latent variables					
		1	2	3	4	5	6
$A^*=4$	X-block	0.7896	0.9657	0.9979	1.0000	1.0000	1.0000
	Y-block	0.5983	0.8928	0.9961	0.9994	0.9994	0.9994
$A^*=3$	X-block	0.8159	0.9978	1.0000	1.0000	1.0000	1.0000
	Y-block	0.7819	0.9950	0.9994	0.9994	0.9994	0.9994
$A^*=2$	X-block	0.8176	1.0000	1.0000	1.0000	1.0000	1.0000
	Y-block	0.7855	0.9994	0.9994	0.9994	0.9994	0.9994

Table 2
Comparison of the frequencies of the selected number of latent variables (sample size, $N=1000$; number of blocks, $k=5$)

True models	Criteria	Number of latent variables						Mean number of latent variables
		1	2	3	4	5	6	
$A^*=4$	AIC	0	0	0	758	228	14	4.26
	Osten's F Criterion	0	0	0	854	132	14	4.16
	Wold's R Criterion	0	0	0	490	36	474	4.98
	Adjusted $R(0.95)$	0	0	0	1000	0	0	4.00
	Adjusted $R(0.90)$	0	0	0	1000	0	0	4.00
$A^*=3$	AIC	0	0	623	340	37	0	3.41
	Osten's F Criterion	0	0	861	115	23	1	3.16
	Wold's R Criterion	0	0	611	44	16	329	4.06
	Adjusted $R(0.95)$	0	0	1000	0	0	0	3.00
	Adjusted $R(0.90)$	0	0	1000	0	0	0	3.00
$A^*=2$	AIC	0	478	443	79	0	0	2.60
	Osten's F Criterion	0	895	75	30	0	0	2.14
	Wold's R Criterion	0	713	26	43	0	218	2.98
	Adjusted $R(0.95)$	0	1000	0	0	0	0	2.00
	Adjusted $R(0.90)$	0	1000	0	0	0	0	2.00

In this example the variables are highly multicollinear, thus including all six latent variables is unacceptable. From Table 2, it can be observed that there is a tendency for Wold's R criterion to include more latent variables than required. One reason for this is that the ratio of consecutive PRESS values may be influenced by random noise, thus a threshold of unity tends to encourage overfitting. Similar conclusions can be drawn for $A^*=2$ and 3. Thus in this study it can be concluded that the adjusted Wold's R criteria exhibits superior performance to both AIC and Wold's R criterion.

4.2. Comparison of different sample sizes

The simulation study was repeated for a sample size of $N=100$. The number of blocks was again taken as $k=5$ and the experiments were repeated 1000 times. For a sample size of 100, the relative cumulative variances captured by the six latent variables for the X and Y blocks, averaged over the 1000 simulation experiments, are given in Table 3. They are similar to their counterparts in Table 1 where $N=1000$.

Table 4 summarises the frequencies of the selected numbers of latent variables for a sample size of 100. It can be seen that, in general, the results of AIC are better in Table 4 than those in Table 2. This is because the multivariate AIC adjusts for small sample sizes [15]. For Osten's F, Wold's R and the adjusted R criterion, the results are slightly worse than those reported in Table 2 for a sample size of 1000. However, it is clear that the adjusted Wold's R criteria still exhibit the best performance.

Table 4
Comparison of frequencies of the selected numbers of latent variables (sample size, $N=100$; number of blocks, $k=5$)

True models	Criteria	Number of latent variables						Mean number of latent variables
		1	2	3	4	5	6	
$A^*=4$	AIC	0	0	0	840	159	1	4.16
	Osten's F Criterion	0	0	0	823	162	15	4.19
	Wold's R Criterion	0	0	0	476	59	465	4.99
	Adjusted R(0.95)	0	0	0	969	31	0	4.03
	Adjusted R(0.90)	0	0	0	1000	0	0	4.00
$A^*=3$	AIC	0	0	711	268	21	0	3.31
	Osten's F Criterion	0	0	860	123	17	0	3.16
	Wold's R Criterion	0	0	600	57	12	331	4.07
	Adjusted R(0.95)	0	0	938	61	1	0	3.06
	Adjusted R(0.90)	0	0	993	7	0	0	3.01
$A^*=2$	AIC	0	535	413	51	1	0	2.52
	Osten's F Criterion	0	880	102	17	1	0	2.14
	Wold's R Criterion	0	670	60	48	4	218	3.04
	Adjusted R(0.95)	0	961	38	1	0	0	2.04
	Adjusted R(0.90)	0	997	3	0	0	0	2.00

4.3. Comparison of different block-sizes in cross-validation

Selecting a different number of blocks in cross-validation can affect the results of model selection. Theoretically, the method of 'leave-one-out' can extract the maximum possible information [14] and thus is the best in theory. This method, however, has very high computational costs, especially when sample sizes are large. This is the main reason for many researchers proposing that the data are split into four to six blocks and it has been argued that the impact of different numbers of partitions has limited impact on the final results [7,12,16].

Table 3
Relative cumulative variances of **X** and **Y** for $A^*=2$, 3 and 4 (sample size, $N=100$; number of blocks, $k=5$)

True models	Blocks	Number of latent variables					
		1	2	3	4	5	6
$A^*=4$	X-block	0.7887	0.9656	0.9979	1.0000	1.0000	1.0000
	Y-block	0.6002	0.8941	0.9961	0.9994	0.9994	0.9994
$A^*=3$	X-block	0.8155	0.9978	1.0000	1.0000	1.0000	1.0000
	Y-block	0.8412	0.9983	0.9994	0.9994	0.9994	0.9994
$A^*=2$	X-block	0.8194	1.0000	1.0000	1.0000	1.0000	1.0000
	Y-block	0.7871	0.9994	0.9994	0.9994	0.9994	0.9994

For a sample size of 100, i.e. $N=100$, Table 5 summarises the frequencies as to the number of latent variables selected where the number of blocks in the cross-validation is taken as $k=5$, 10 and 100 (i.e. leave-one-out). It can be seen that when the number of blocks is large, the percentage of choosing the appropriate number of latent variables to give the true model is slightly higher. However in general these improve-

Table 5
Comparison of frequencies of the selected numbers of latent variables for different numbers of blocks in cross-validation with sample size, $N=100$

True models	Criteria	Number of blocks	Number of latent variables						Mean number of latent variables
			1	2	3	4	5	6	
$A^*=4$	Osten's F Criterion	k = 5	0	0	0	823	162	15	4.19
		k = 10	0	0	0	886	108	6	4.12
		leave-one-out	0	0	0	916	75	9	4.09
	Wold's R Criterion	$k=5$	0	0	0	476	59	465	4.99
		$k=10$	0	0	0	440	32	528	5.09
		leave-one-out	0	0	0	394	0	606	5.21
	Adjusted $R(0.95)$	$k=5$	0	0	0	969	31	0	4.03
		$k=10$	0	0	0	981	19	0	4.02
		leave-one-out	0	0	0	986	14	0	4.01
	Adjusted $R(0.90)$	$k=5$	0	0	0	1000	0	0	4.00
		$k=10$	0	0	0	1000	0	0	4.00
		leave-one-out	0	0	0	999	1	0	4.00
$A^*=3$	Osten's F Criterion	k = 5	0	0	860	123	17	0	3.16
		k = 10	0	0	885	98	17	0	3.13
		leave-one-out	0	0	910	73	16	1	3.11
	Wold's R Criterion	$k=5$	0	0	600	57	12	331	4.07
		$k=10$	0	0	630	35	18	317	4.02
		leave-one-out	0	0	641	31	6	322	4.01
	Adjusted $R(0.95)$	$k=5$	0	0	938	61	1	0	3.06
		$k=10$	0	0	961	38	1	0	3.04
		leave-one-out	0	0	968	32	0	0	3.03
	Adjusted $R(0.90)$	$k=5$	0	0	993	7	0	0	3.01
		$k=10$	0	0	997	3	0	0	3.00
		leave-one-out	0	0	998	2	0	0	3.00
$A^*=2$	Osten's F Criterion	k = 5	0	880	102	17	1	0	2.14
		k = 10	0	895	79	25	1	0	2.13
		leave-one-out	0	913	60	27	0	0	2.11
	Wold's R Criterion	$k=5$	0	670	60	48	4	218	3.04
		k = 10	0	690	49	53	2	206	2.99
		leave-one-out	0	694	46	91	0	169	2.90
	Adjusted $R(0.95)$	$k=5$	0	961	38	1	0	0	2.04
		$k=10$	0	966	32	2	0	0	2.04
		leave-one-out	0	979	21	0	0	0	2.02
	Adjusted $R(0.90)$	$k=5$	0	997	3	0	0	0	2.00
		$k=10$	0	1000	0	0	0	0	2.00
		leave-one-out	0	999	1	0	0	0	2.00

ments, where they exist, are marginal. Therefore, the simulation results in this paper support the conclusions of Refs. [7,12], i.e. partitioning the data into four to six blocks is appropriate for cross-validation.

5. An industrial application of chemical process modeling

An industrial fluidised-bed catalytic reactor is now considered [23]. Conventional process controllers maintain the flows and temperatures at the desired levels. The reactor is cooled by a number of internal cooling coils in the bed of the reactor with the coil coolant flow rates being fixed. As the coils foul, their thermal efficiency is reduced and the flow of raw materials into the reactor changes through the action of the plant control systems. The aims of the plant operational staff in operating the process are to maximize reactor performance for a given production rate and to minimize product losses to by-products. This has a consequential impact on equipment fouling, catalyst efficacy, waste treatment costs, energy usage per unit product, inventory and responsiveness to scheduled grade changes. During reactor operation, the cooling coils are switched in and out as part of a de-fouling recycling procedure to maintain optimum bed cooling and good production. This results in clustering in the data due to different operation conditions.

Data on the process variables is collected as 5-min averages, with data for the quality variables (yields and conversion of the process feed) being available from an on-line gas chromatograph every 30 min. Seven 'chemical quality' variables and 36 process variables are measured on-line. A total of 1335 observations are included for process modeling. The reactor is complex, multivariable, highly instrumented and data rich. Because of the many competing reactions in the process, there is no mechanistic model that provides the kinetic distribution of products from a specific set of process conditions. The objectives of the analysis are therefore to build a representation of the process for the monitoring of process performance, providing on-line diagnostic support through the early warning of process malfunctions and the identification of changes in process operation.

Fig. 1 shows a plot of the relative cumulative variances captured by the PLS latent variables. Two different phases of behaviour are observed. In the first phase, including an additional latent variable results in the capture of a significant level of variability in the quality variables, whilst in the second phase, only marginal improvements are observed as more latent variables are included in the model. The cut-off point of the two phases is located between 7 and 10 latent variables. Similar behavior can be observed in Fig. 2 where during phase I, the PRESS decreases rapidly. The rate of decrease in the PRESS then becomes quite slow after the first 7–10 latent variables have been included in the model.

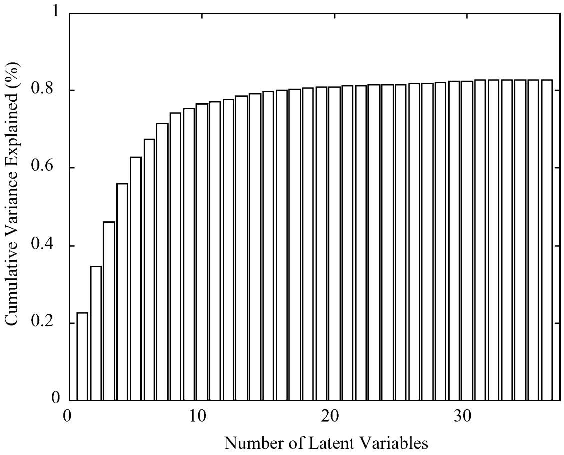

Fig. 1. Relative cumulative variance captured by latent variables.

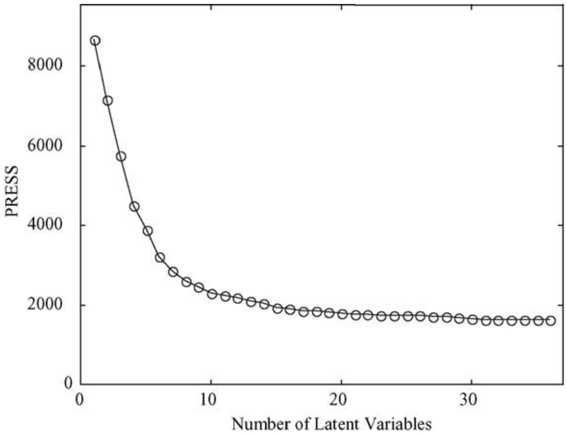

Fig. 2. PRESS versus number of latent variables (number of blocks, $k = 5$).

Applying different criteria to the data set results in quite different PLS models. Both Osten's F and Wold's R criteria include 21 latent variables in the model, whilst AIC identifies 32 latent variables. In contrast, adjusted Wold's R criteria, $R(0.90)$ and $R(0.95)$, result in more parsimonious model, including only 7 and 10 latent variables respectively in the model. From Figs. 1 and 2, it appears that both Wold's R criterion and AIC include too many latent variables when compared with the increase in captured variance or the decrease in PRESS with increasing numbers of latent variables.

6. A study using industrial NIR data

The major focus of this paper is model selection for PLS for its application to multivariate statistical process control (MSPC), where the number of observations are typically much larger than the number of process variables and the aim of the analysis is to

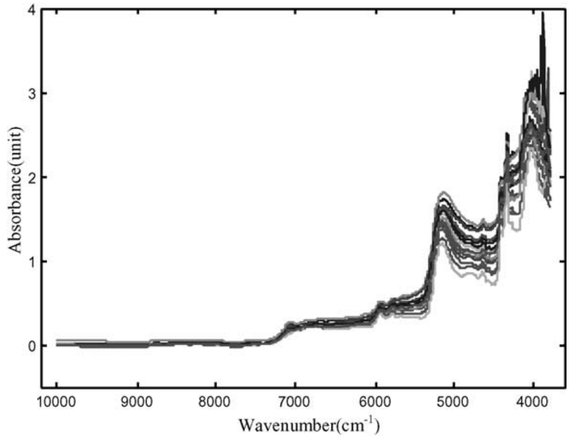

Fig. 3. NIR spectra for calibration data.

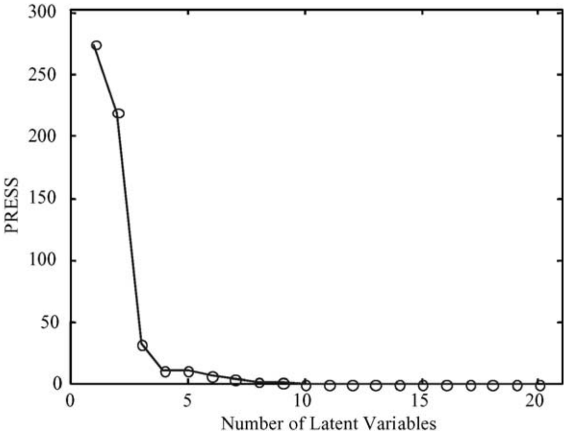

Fig. 4. PRESS versus number of latent variables (number of blocks, $k=5$).

extract latent variables to build representative models of the process. However, in contrast to this situation is where the number of observations is less than the number of explanatory variables. This is a typical scenario with spectroscopic data.

An industrial NIR spectral data set was collected to investigate the feasibility of replacing certain laboratory analyses with NIR spectroscopy. Two response variables, active and excipient, are considered. The NIR spectra with 1550 channels were measured over the range 10 000 to 3800 cm^{-1} using a resolution of 4 cm^{-1}. Spectra for sixteen samples were used for calibration (Fig. 3). A further six samples were used as unseen test data. The prediction was investigated in terms of the root mean prediction squared error (RMPSE).

A combined model for both response variables (active and excipient) was first built using PLS2. The variation captured for both **X** and **Y** by PLS for models with different numbers of latent variables is shown in Table 6. Fig. 4 gives the plot for the PRESS.

From Table 6 and Fig. 4, it is conjectured that a PLS model comprising four to six latent variables is appropriate. Model selection and prediction are summarized in Table 7 (Case I). It can be seen that by applying the adjusted Wold's *R* criteria a parsimonious model is obtained and the best prediction, i.e. the lowest value of RMPSE, is obtained with four latent variables. AIC also gives quite good prediction results. Wold's *R* criterion included too many (81) latent variables, resulting in a large value of RMPSE, clearly overfitting the data in this case. Osten's *F* is prematurely terminated for this example since only one latent variable has been included.

A model for the response variable 'active ingredient' was then built using PLS1. Model selection and prediction are summarized in Table 7 (Case II). In this case, Wold's *R* criterion has a tendency to include too many latent variables but it gives the best performance in terms of prediction. Finally, a model for the response variable 'excipient' is built using PLS1 with

Table 6
Relative cumulative variances of **X** and **Y** explained

No. of LVs		1	2	3	4	5	6	8	10	12
Case I	X-block	0.8619	0.9385	0.9589	0.9839	0.9904	0.9956	0.9980	0.9985	0.9998
$Y=[y_1, y_2]$	Y-block	0.5117	0.6950	0.9485	0.9821	0.9859	0.9903	0.9984	0.9996	1.0000
Case II	X-block	0.8628	0.8747	0.9216	0.9837	0.9872	0.9924	0.9980	0.9985	0.9998
$Y=y_1$	Y-block	0.5256	0.9722	0.9830	0.9852	0.9898	0.9929	0.9987	0.9998	1.0000
Case III	X-block	0.8608	0.9395	0.9612	0.9839	0.9871	0.9918	0.9980	0.9986	0.9998
$Y=y_2$	Y-block	0.5267	0.7514	0.9485	0.9827	0.9892	0.9933	0.9988	0.9997	1.0000

Table 7
Selected models and predictions

		AIC	*F*	*R*	Adjusted *R*(0.95)	Adjusted *R*(0.90)
Case I	No. of LVs	5	1	81	4	4
Y=[y_1, y_2]	RMPSE	1.7808	3.3648	9.1304	1.7183	1.7183
Case II	No. of LVs	3	1	41	4	4
$Y=y_1$	RMPSE	1.8522	2.1169	1.2627	1.6609	1.6609
Case III	No. of LVs	6	1	54	54	4
$Y=y_2$	RMPSE	1.5078	4.1428	1.3276	1.3276	1.7430

model selection and prediction being summarized in Table 7 (Case III). For this situation, Wold's *R* criterion and adjusted *R*(0.95) exhibit the overall best performance in terms of prediction.

It should be noted that for cases II and III very similar patterns, in terms of the captured variances, are apparent (see Table 6) and similar plots of PRESS (not displayed) as for Case I are achieved. Therefore, it can be assumed that an appropriate model should include between four and six latent variables for all three cases. However, these 'appropriate' models do not necessarily give the best prediction in terms of the RMPSE for the test data. Thus, from this example, it can be concluded that there are some very complex mechanisms that need to be taken into account during model selection and building. A model capturing a reasonable amount of variation in the response variable(s) does not necessarily give the best prediction for the unseen test data.

7. Conclusion and discussions

A comparison has been performed between Wold's *R* criterion and AIC for the selection of the number of latent variables to include in a PLS model. From the simulation results, the adjusted Wold's *R* criteria which adopts thresholds less than unity results in more parsimonious models than those selected by Wold's criterion, and achieve a higher percentage in terms of selecting the true model. This is because when using the adjusted Wold's *R* criteria, a latent variable will not be included in the PLS model unless the additional latent variable can provide significantly better prediction performance.

Some additional comments are made on the selection of the thresholds. First, since the theoretical distribution of Wold's *R* statistic is unavailable, an 'optimal threshold' is not available. Therefore, an appropriate threshold can be only be chosen empirically. In addition, the threshold depends on the preference of the analyst and their perception in terms of model parsimony and accuracy. This can be clearly seen from the practical examples discussed in Section 5, where $A=7$ and $A=10$ were respectively located at the end of phase 1 and at the beginning of the phase 2 in Figs. 1 and 2.

Finally, it is noted that according to Eastment and Kraznowski [14], model selection is also strongly dependent on the objective of the analysis. If the aim is dimensionality reduction and noise removal, such as for applications in MSPC, it is suggested from the simulation study carried out in this paper that the adjusted Wold's *R* would be an appropriate choice. If, however, the aim is prediction, for example, then Wold's *R* criterion or AIC may be preferred since accuracy, instead of parsimony, is the major concern as indicated by the example of the NIR data.

Acknowledgements

The authors acknowledge the financial support of the EU ESPRIT project PERFECT (Performance Enhancement through Factory On-line Examination of Process Data) No. 28870.

References

[1] E.B. Martin, A.J. Morris, C. Kiparrisides, Manufacturing performance enhancement through multivariate statistical process control, Annu. Rev. Control 23 (1999) 35–44.

[2] A. Höskuldsson, PLS regression methods, J. Chemom. 29 (1988) 409–412.

[3] H. Wold, Non-linear estimation by iterative least squares procedures, in: F. David (Ed.), Research Papers in Statistics, Wiley, New York, 1966, pp. 411–444.

[4] H. Wold, Model construction and evaluation when theoretical knowledge is scare. Theory and application of partial least squares, in: J. Kmenta, J.B. Ramsey (Eds.), Evaluation of Econometric Models, Academic Press, New York, 1980, pp. 47–74.

[5] R. Haber, H. Unbenhauen, Structure identification of non-linear dynamic systems—a survey on input–output approaches, Automatica 26 (4) (1990) 651–677.

[6] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. B36 (1974) 111–147.

[7] S. Wold, Cross-validation estimation of the number of components in factor and principal component analysis, Technometrics 24 (1978) 397–405.

[8] I.S. Helland, Maximum likelihood regression on relevant components, J. R. Stat. Soc. B54 (1992) 637–647.

[9] T. Næs, I.S. Helland, Relevant components in regression, Scand. J. Statist. 20 (1993) 239–250.

[10] P. Naik, C.-L. Tsai, Partial least squares estimator for single-index models, J. R. Stat. Soc. B62 (2000) 763–771.

[11] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. Stat. Math. 21 (1969) 243–247.

[12] D.W. Osten, Selection of optimal regression models via cross-validation, J. Chemom. 2 (1988) 39–48.

[13] A. Lorber, B.R. Kowalski, Alternatives to cross-validatory estimation of the number of factors in multivariate calibration, Appl. Spectrosc. 44 (1990) 1464–1470.

[14] H.T. Eastment, W.J. Krzanowski, Cross-validatory choice of the number of components from a principal component analysis, Technometrics 24 (1982) 73–77.

[15] E.J. Bedrick, C.-L. Tsai, Model selection for multivariate regression in small samples, Biometrics 50 (1994) 226–231.

[16] D.M. Himes, R.H. Storer, C. Georgakis, Determination of the number of principal components for disturbance detection and isolation, Proceeding of the American Control Conference, Baltimore, 1994, pp. 1279–1283.

[17] W.J. Krzanowski, Cross-validation in principal component analysis, Biometrics 43 (1987) 575–584.

[18] H. Akaike, A new look at statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723.

[19] H. Akaike, Factor analysis and AIC, Psychometrika 52 (1987) 317–332.

[20] E.J. Hannan, B.G. Quinn, The determination of the order of autoregression, J. R. Stat. Soc. B41 (1979) 190–195.

[21] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[22] T. Næs, H. Martens, Comparison of prediction methods for collinear data, Commun. Stat., Simul. 14 (1985) 545–576.

[23] A. Simoglou, E.B. Martin, A.J. Morris, Multivariate statistical process control of a fluidised bed reactor, Control Eng. Pract. 8 (2000) 809–893.