ELSEVIER Linear Algebra and its Applications 354 (2002) 245–253

LINEAR ALGEBRA AND ITS APPLICATIONS

www.elsevier.com/locate/laa

The asymptotic variance of the univariate PLS estimator

A. Phatak [a,*], P.M. Reilly [b], A. Penlidis [b]

[a] *CSIRO – Mathematical & Information Sciences, Private Bag 5, Wembley, WA 6913, Australia*
[b] *Department of Chemical Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1*

Received 12 March 1999; accepted 16 April 2001

Submitted by H.J. Werner

Abstract

In this short note, we derive an expression for the asymptotic covariance matrix of the univariate partial least squares (PLS) estimator. In contrast to M.C. Denham [J. Chemometrics 11 (1997) 39], who provided a locally linear approximation based on a recursive definition of the estimator, we derive a more compact expression for the asymptotic covariance matrix by combining a standard convergence result with matrix differential calculus, in particular the approach of J.R. Magnus and H. Neudecker [Matrix Differential Calculus with Applications in Statistics and Econometrics, revised ed., Wiley, Chichester, UK, 1991]. We also describe some theoretical and practical aspects of calculating the asymptotic covariance matrix, and illustrate its use on spectroscopic data.

AMS classification: 62J07; 62E20; 62H99; 15A99; 62P30

Keywords: Partial least squares regression; Matrix differential calculus; Asymptotic covariance

1. Introduction

Univariate partial least squares regression (PLS) is a biased estimation procedure that is widely used in the field of chemometrics [2,9,19,21]. It is closely related to principal component regression in the sense that the basic idea is to regress onto a subspace of the predictor variables rather than onto the range space of all the

* Corresponding author.
E-mail addresses: aloke.phatak@cmis.csiro.au (A. Phatak), pmreilly@cape.uwaterloo.ca (P.M. Reilly), penlidis@cape.uwaterloo.ca (A. Penlidis).

0024-3795/02/$ - see front matter
PII: S0024-3795(01)00357-3

predictors, as in ordinary least squares. It is different, however, in that the subspace is chosen with respect to the response variable, and hence the estimator of the vector of coefficients β in the usual linear model is a non-linear function of the vector of observations y. As a consequence, it is difficult, if not impossible, to derive the exact distribution of the estimator. Approximate distributional results would be useful for constructing, for example, confidence intervals for the parameter estimates or for predictions from PLS.

In some earlier work, Denham [5] provided a locally linear approximation to the covariance matrix based on the first derivative of the PLS vector of coefficients. His expression stems from a recursive definition of the estimator first derived by Helland [10]. The principal objective of this note is to derive a more compact expression for the asymptotic covariance matrix by combining a standard result in convergence with matrix differential calculus, in particular, the approach of Magnus and Neudecker [14]. Similar expressions were given in [17] without proof. Section 2 outlines the mechanics of PLS regression and the principal result is derived in Section 3. Section 4 briefly discusses some theoretical and practical aspects of its use and then goes on to outline its application to real data.

2. Univariate PLS regression

We begin with the usual linear model of the form

$$y = 1_n\beta_0 + X\beta + \epsilon, \tag{1}$$

where y is an $n \times 1$ vector of observations on a response variable, the vector 1_n is of length n and consists of ones; X is an $n \times p$ matrix consisting of values of p explanatory variables whose columns have been centered about their means, so that $X'1_n = 0$; ϵ is an n-vector of errors which are independently and identically distributed with zero expectation and variance σ^2; β_0 is an unknown constant; and β represents a $p \times 1$ vector of coefficients. For the purposes of the derivation below, we shall assume that $n > p$ and that X is of full rank. Furthermore, we limit the discussion to the estimation of β.

The ordinary least squares estimator of β is given by $\hat{\beta} = (X'X)^{-1}X'y$ and its variance by $(X'X)^{-1}\sigma^2$. When the columns of X are highly collinear, however, the variances of some of the estimated regression coefficients can be very large. Consequently, the mean squared error, defined as $\mathrm{MSE}(\hat{\beta}) = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)]$, can become inflated. To overcome this problem, biased methods such as ridge regression [9], principal component regression [11], and partial least squares are often used. In using biased estimation, we hope that the bias of the estimator will be offset by a corresponding reduction in variance, and hence that the overall MSE will be reduced.

A brief description of the origins of PLS may be found in [4], along with its statistical properties and its connections to Rayleigh–Ritz/Lanczos methods for find-

ing the extremal eigenvalues of a symmetric matrix. Simulation studies show that PLS performs at least as well as, and sometimes outperforms, ridge and principal component regression [1,9].

The PLS estimate of β in (1) is given by

$$\hat{b}_m = K_m(K_m'X'XK_m)^{-1}K_m'X'y = K_m(K_m'SK_m)^{-1}K_m's, \tag{2}$$

where $m = 1, 2, \ldots, p$ specifies the 'dimensionality' of the estimator and the columns of $K_m(p \times m)$ are given by the Krylov sequence $\{s, Ss, \ldots, S^{m-1}s\} \equiv \mathscr{K}(S, s, m)$, with $s = X'y$ and $S = X'X$. The value of m is not known a priori, and hence it is usually determined by cross-validation.

In (2), any matrix $V_m = K_m M$ can be used as long as its columns span the same space as $\mathscr{K}(S, s, m)$, that is, as long as $M(m \times m)$ is non-singular. In particular, two alternative bases, which we denote by W_m and R_m, are often used, and their columns w_i and r_i arise out of algorithms commonly used to calculate $\hat{b}_m$. The w_i stem from the Gram–Schmidt orthogonalization of $\mathscr{K}(S, s, m)$ in its natural order; as a result, $W'X'XW$ is tridiagonal [15]. Alternatively, we can calculate $V_m = R_m$ such that $V_m'X'XV_m$ is a diagonal matrix.

There is currently a lively debate going on about the nature of PLS regression. On the one hand, most statisticians think of it much as we have outlined it above, as a biased estimator of the parameters of the linear model in (1) [2,9,19]. On the other hand, others, especially in the chemometric literature, sometimes prefer to think of it as arising out of a latent variable model [3]. It is not clear, however, that this philosophical difference makes any *practical* difference to how PLS is used. The interested reader is referred to the papers by Frank and Friedman [9] and Sundberg [19] and the discussions contained therein for an exposition and analysis of both points of view.

3. Asymptotic variance of PLS estimator

The basic result which allows us to calculate the asymptotic covariance matrix of $\hat{b}_m$ is given in the following theorem; it forms the basis for the approximation method known as the 'delta' method.

Theorem 1 ([18, p. 388] and [12]). *Let $\{y_{(n)}\}$ be a sequence of random vectors $y_{(n)}$ and μ a compatible fixed vector. Assume that $\sqrt{n}[y_{(n)} - \mu] \xrightarrow{\mathrm{D}} N(0, \Psi)$, or equivalently that $\sqrt{n}\,y_{(n)}$ is asymptotically normally distributed with mean $\sqrt{n}\mu$ and variance Ψ.*

Furthermore, let $g(z)$ be a vector function of a vector z with first and second derivatives existing in a neighbourhood of $z = \mu$. Then

$$\sqrt{n}[g(y_{(n)}) - g(\mu)] \xrightarrow{\mathrm{D}} N(0, T\Psi T'),$$

where the matrix T is given by

$$T = \left.\frac{\partial g(z)}{\partial z'}\right|_{z=\mu}$$

and consists of the derivatives of the elements of $g(z)$ with respect to the elements of z.

Hence, what we require now to obtain the asymptotic variance of $\hat{b}_m$ is the Jacobian matrix $J = \partial\hat{b}_m(y)/\partial y'$. The result is set out below.

Theorem 2. *Let $\hat{b}_m$ be defined as in* (2). *Then*

$$J \equiv \frac{\partial\hat{b}_m(y)}{\partial y'} = (s' \otimes I_p)(I_{p^2} + C_p)\left[K_m(K_m'SK_m)^{-1} \otimes (I_p - H_mS)\right]U_m' + H_mX', \tag{3}$$

where I_p denotes the identity matrix of order p, C_p denotes the $p^2 \times p^2$ commutation matrix, $H_m = K_m(K_m'SK_m)^{-1}K_m'$, and $U_m = \{X, XS, \ldots, XS^{m-1}\}$.

Proof. In the derivation that follows, we drop the subscript m in intermediate steps for the sake of clarity and compactness. Taking differentials of (2) yields

$$\begin{aligned}\mathrm{d}\hat{b}_m = {} & (\mathrm{d}K)(K'SK)^{-1}K's + K\Big[-(K'SK)^{-1}(\mathrm{d}K')SK(K'SK)^{-1} \\ & -(K'SK)^{-1}K'S(\mathrm{d}K)(K'SK)^{-1}\Big]K's + K(K'SK)^{-1}(\mathrm{d}K')s \\ & + HX'\mathrm{d}y,\end{aligned} \tag{4}$$

where the expression inside the square brackets $[\cdot]$ represents $\mathrm{d}(K'SK)^{-1}$. Expanding (4) and using the definition of H gives

$$\begin{aligned}\mathrm{d}\hat{b}_m = {} & (\mathrm{d}K)(K'SK)^{-1}K's - K(K'SK)^{-1}(\mathrm{d}K')SHs \\ & -HS(\mathrm{d}K)(K'SK)^{-1}K's + K(K'SK)^{-1}(\mathrm{d}K')s \\ & + HX'\mathrm{d}y,\end{aligned} \tag{5}$$

Taking vecs and using the result that $\mathrm{vec}(ABC) = (C' \otimes A)\,\mathrm{vec}\,B$ for conformable matrices A, B, and C leads to

$$\begin{aligned}\mathrm{d}\hat{b}_m = {} & \Big[s'K(K'SK)^{-1} \otimes I_p\Big]\,\mathrm{vec}\,\mathrm{d}K - \Big[s'K(K'SK)^{-1} \otimes HS\Big]\,\mathrm{vec}\,\mathrm{d}K \\ & + \Big[s' \otimes K(K'SK)^{-1}\Big]\,\mathrm{vec}\,\mathrm{d}K' - \Big[s'HS \otimes K(K'SK)^{-1}\Big]\,\mathrm{vec}\,\mathrm{d}K' \\ & + HX'\mathrm{d}y\end{aligned}$$

$$= \left[s'K(K'SK)^{-1} \otimes (I_p - HS) \right] \text{vec } \mathrm{d}K + \left[s'(I_p - HS) \otimes K(K'SK)^{-1} \right] \text{vec } \mathrm{d}K' + HX'\mathrm{d}y. \quad (6)$$

To combine the terms in $\mathrm{d}K$ and $\mathrm{d}K'$ we need to use the following result. For $c(p \times 1)$, $B(p \times m)$, and $A(p \times m)$ say,

$$(c' \otimes B)\text{vec } A' = (B \otimes c')\text{vec } A \quad (7)$$

since both are equal to the vector $BA'c$. Using (7) in (6) and rearranging further yields

$$\mathrm{d}\hat{b}_m = \left(s' \otimes I_p + I_p \otimes s'\right) \left[K(K'SK)^{-1} \otimes (I_p - HS) \right] \mathrm{d} \text{ vec } K + HX'\mathrm{d}y. \quad (8)$$

Now from the definition of K_m, we can write $\text{vec} K_m = U_m' y$, where U_m is the $n \times pm$ matrix $\{X, XS, \ldots, XS^{m-1}\}$. Inserting this into (8) and then simplifying and rearranging leads to

$$\mathrm{d}\hat{b}_m = (s' \otimes I_p)(I_{p^2} + C_p) \left[K(K'SK)^{-1} \otimes (I_p - HS) \right] U'\mathrm{d}y + HX'\mathrm{d}y, \quad (9)$$

where, for $A(p \times p)$, C_p is the $p^2 \times p^2$ commutation matrix [14] that transforms vec A into vec A', i.e., C_p vec A = vec A'. Hence, the Jacobian matrix is given by

$$J = (s' \otimes I_p)(I_{p^2} + C_p) \left[K_m(K_m'SK_m)^{-1} \otimes (I_p - H_m S) \right] U_m' + H_m X', \quad (10)$$

where the subscript m has been restored to emphasize that it is the Jacobian of the m-dimensional estimator, $\hat{b}_m$. □

For large n, therefore, and under the assumptions of the linear model in (1), the variance of the PLS estimator $\hat{b}_m$ is $JJ'\sigma^2$ evaluated about some point y_0. Note also that any matrix $V_m = K_m M$ can be used in the expression above as long as $M(m \times m)$ is non-singular. In the resulting expression, V_m is inserted in place of K_m and U_m' replaced by $(M' \otimes I_p)U_m'$.

A slightly more economical representation of (10) is possible if we recognize that $I_p - H_m S$ is an oblique projector. $H_m S$ projects onto K_m along the direction orthogonal to SK_m, and we write it as $\mathscr{H}_{K_m|SK_m^{\perp}}$. Consequently, $I_p - H_m S$ projects onto $SK_m^{\perp}$ along K_m and can be written as $\mathscr{H}_{SK_m^{\perp}|K_m}$. Hence, an alternative to (10) is

$$J = (s' \otimes I_p)(I_{p^2} + C_p) \left[K_m(K_m'SK_m)^{-1} \otimes \mathscr{H}_{SK_m^{\perp}|K_m} \right] U_m' + H_m X'. \quad (11)$$

Because both (10) and (11) involve Kronecker products, large, sparse matrices will be generated. An alternative form that avoids such large matrices can be derived using (8) as a starting point, and it can be written as follows. If we let $D_m = K_m(K_m' \times SK_m)^{-1}$, then

$$J = \left\{ \sum_{i=1}^{m} \left[(s'D_m')_i \cdot \mathscr{H}_{SK_m^{\perp}|K_m} + (D_m)_i s' \mathscr{H}_{SK_m^{\perp}|K_m} \right] S^{i-1} \right\} X' + H_m X', \tag{12}$$

where the notation $(\cdot)_i$ indicates the ith element of a vector or the ith column of a matrix.

4. Practical issues

In practice, the locally linear approximation $JJ'\sigma^2$ must be evaluated about some point y_0 and a suitable estimate of σ^2 must be obtained. Denham [5,6] and Phatak [16] consider these issues in greater detail, and the interested reader is referred to those works for a more comprehensive discussion. The meta-parameter m must also be estimated.

Two natural points about which to linearize the PLS estimator $\hat{b}_m$ are $y_0 = E(y)$ and the observed data y. In the former instance, some plug-in estimate of $E(y)$ will be required; Denham [6] suggests some alternatives, but linearizing about the observed data has been found to work well in practice [6,16].

A reasonable estimate of the error variance σ^2 can be obtained by calculating the residual sum of squares (RSS) divided by some appropriate number of degrees of freedom. If $\hat{y} = 1_n\bar{y} + X\hat{b}_m$ represents the fitted values from PLS, the residuals are $r = y - \hat{y}$. Hence $\mathrm{RSS} = r'r$ and $\hat{\sigma}^2 = \mathrm{RSS}/(d-1)$. But what should the value of d be? Following on from ordinary regression, $d = n - m$ has been suggested, but it has been pointed out [9,20] that PLS uses up more than m degrees of freedom because it is a non-linear function of y. Hence, both Denham [5] and Phatak [16] suggest using

$$d = \mathrm{tr}(I_n - J'X')(I_n - XJ), \tag{13}$$

where I_n is the identity matrix of order n and J must be evaluated at y_0. The matrix $(I_n - J'X')(I_n - XJ)$ can be thought of as analogous to the 'hat' matrix $I_n - X(X'X)^{-1}X'$ in ordinary regression, but although the former is symmetric it is not idempotent. We should note in passing that PLS is often used in instances where $n - 1 < p$. In such circumstances, no estimate of σ^2 can be obtained, and the local linearization procedure must be replaced by, for example, bootstrapping [5].

For confidence intervals constructed from the linear approximation to have good coverage properties, it is important to choose the correct value of the meta-parameter m [5]. As with ridge-regression, where the ridge parameter must also be estimated from the data [9], cross-validation is commonly used to determine m (see, for

example, [13]). Fortunately, in many of the applications for which PLS is used, the minimum cross-validation sum of squares occurs over a range of values of m. Also, extensive simulations [7,8,16] show that at or near the correct value of m, bias is negligible. Alternatively, Denham [5,6] provides expressions for bias based on different plug-in estimates of $E(y)$.

An example. The example we consider—calibration of an infrared spectrometer—illustrates a common application of PLS. The observed response y consists of concentrations between 0.5% and 5% of a particular constituent of $n = 229$ samples of oil-bearing shales. The explanatory variables consist of near-infrared spectra measured at $p = 82$ equally spaced wavelengths in the range (w_1, w_p) nanometres.[1] Thus, each explanatory variable corresponds to the reflectance of the sample when the incident radiation has wavelength w_i, $i = 1, 2, \ldots, p$. The columns of X are highly collinear and thus the eigenvalue spectrum of the cross-product matrix $\lambda(X'X)$ is broad and consists of a few large eigenvalues and many small ones:

$$(\lambda_1, \ldots, \lambda_{82}) = (101.1, 37.5, 18.8, 2.3, 1.2, \ldots, 2.5 \times 10^{-8}, 2.4 \times 10^{-8}).$$

Here, the ordinary least squares estimator, though unbiased, will have large variance $((X'X)^{-1}\sigma^2)$ because of the presence of very small eigenvalues. In circumstances such as these, however, the use of biased estimation procedures generally leads to a substantial reduction in variance at the cost of a comparatively small bias; hence, mean squared error will be reduced [11, Chapter 8].

Cross-validation was used to estimate the dimensionality m of the PLS estimator. Fig. 1 shows the cross-validation sum-of-squares as a function of m, and we can see that there is a broad trough between $m = 6$ and $m = 8$. To illustrate the calculations, we shall use $m = 6$ and evaluate the Jacobian at the observed data.

The elements of the parameter estimate $\hat{b}_{m=6}$ are plotted in Fig. 2. It has the appearance of a smooth curve because the estimates have been joined. The residual sum-of-squares is RSS $= 7.33$, and the value of d calculated using (13) is $d = 219.9$. Hence, PLS uses up approximately $n - d = 229 - 219.9 = 9.1 (> m = 6)$ degrees of freedom, and the corresponding estimate of σ^2 is $\hat{\sigma}^2 = 7.33/(219.9 - 1) = 0.034$.

From (10)–(12), we can see that the first term of the Jacobian provides the basis of a correction to the zeroth order approximation $H_m X'X H_m \hat{\sigma}^2 = H_m \hat{\sigma}^2$ of the covariance matrix of $\hat{b}_m$. To compare the two estimates of variance, we have used the diagonal elements of $H_m \hat{\sigma}^2$ and $J J' \hat{\sigma}^2$ to construct approximate pointwise 95% confidence bounds for the elements of $\hat{b}_m$. The results are shown in Fig. 2, where we see that the zeroth order approximation underestimates variability and produces smaller intervals. Regression coefficients from spectroscopic data such as these are sometimes used by analysts to provide qualitative information about regions of the spectrum that have greater predictive utility. Hence, by underestimating the

[1] To protect the confidentiality of the data, the wavelengths have not been specified.

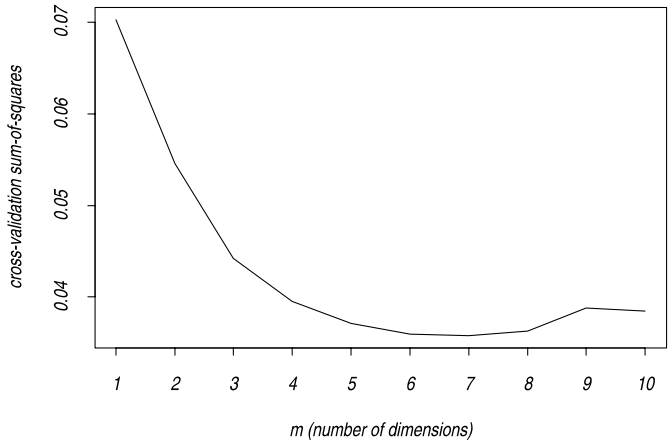

Fig. 1. Cross-validation plot for spectroscopic data.

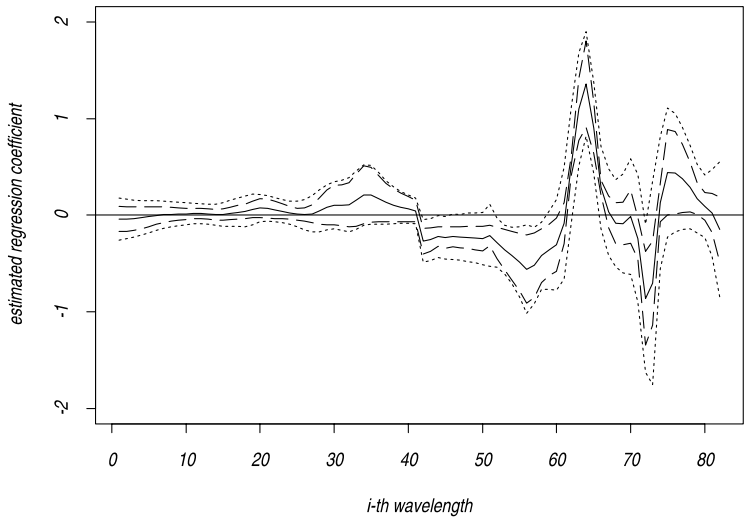

Fig. 2. Estimated regression coefficient $\hat{b}_{m=6}$ (—) plotted along with approximate pointwise 95% confidence intervals derived from (a) zeroth order approximation $H_m\hat{\sigma}^2$ (— — —) and (b) linear approximation $JJ'\hat{\sigma}^2$ (------).

variability, as the zeroth order approximation does, regions can be incorrectly identified as being useful. In Fig. 2, for example, both approximations lead to the same conclusions about the region between w_1 and w_{40}. By contrast, using the zeroth order approximation may lead to the conclusion that the region between w_{42} and w_{50} is useful for prediction, whereas the linear approximation shows that it is likely not useful. In a prediction context, the fact that $H_m\sigma^2$ underestimates variability leads to

optimistic coverage rates of prediction intervals, whereas the linear approximation provides better coverage rates [5].

Acknowledgement

The authors would like to thank Heinz Neudecker and the anonymous referees for providing suggestions that have helped to clarify the presentation.

References

[1] T. Almøy, A simulation study on comparison of prediction methods when only a few components are relevant, Comput. Statist. Data Anal. 21 (1996) 87–107.
[2] P.J. Brown, Measurement, Regression and Calibration, Oxford University Press, Oxford, 1993.
[3] A.J. Burnham, J.F. MacGregor, R. Viveros, Latent variable multivariate regression modeling, Chemometrics Intellig. Lab. Sys. 48 (1999) 167–180.
[4] S. de Jong, A. Phatak, Partial least squares regression, in: S. Van Huffel (Ed.), Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling, SIAM, Philadelphia, PA, 1997, pp. 25–36.
[5] M.C. Denham, Prediction intervals in partial least squares, J. Chemometrics 11 (1997) 39–52.
[6] M.C. Denham, Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction, J. Chemometrics 14 (2000) 351–361.
[7] N.M. Faber, A closer look at the bias-variance trade-of in multivariate calibration, J. Chemometrics 13 (1999) 185–192.
[8] N.M. Faber, Response to Comments on construction of confidence intervals in conection to partial least squares, J. Chemometrics 14 (2000) 363–369.
[9] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–135.
[10] I.S. Helland, On the structure of partial least squares regression, Comm. Statist. B Simulation Comput. 17 (1988) 581–607.
[11] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[12] T. Kollo, H. Neudecker, Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices, J. Multivariate Anal. 47 (1993) 283–300.
[13] F. Lindgren, P. Geladi, S. Wold, Kernel-based PLS regression: Cross-validation and applications to spectral data, J. Chemometrics 8 (1994) 377–389.
[14] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, revised ed., Wiley, Chichester, UK, 1991.
[15] R. Manne, Analysis of two partial-least-squares algorithms for multivariate calibration, Chemometrics Intellig. Lab. Sys. 2 (1987) 187–197.
[16] A. Phatak, Evaluation of some multivariate methods and their applications in chemical engineering, Ph.D. Thesis, University of Waterloo, Waterloo, 1993.
[17] A. Phatak, P.M. Reilly, A. Penlidis, An approach to interval estimation in partial least squares regression, Anal. Chim. Acta 277 (1992) 495–501.
[18] C.R. Rao, Linear Statistical Inference Applications, 2nd ed., Wiley, New York, 1973.
[19] R. Sundberg, Multivariate calibration–Direct and indirect regression methodology, Scand. J. Statist. 26 (1999) 161–207.
[20] H. van der Voet, Pseudo-degrees of freedom for complex predictive models: the example of partial least squares, J. Chemometrics 13 (1999) 195–208.
[21] H. Wold, Soft modeling. The basic design and some extensions, in: K.G. Jöreskog, H. Wold (Eds.), Systems Under Indirect Observation, Part II, North-Holland, Amsterdam, pp. 1–53.