Stat Comput (2017) 27:757–774
DOI 10.1007/s11222-016-9651-4

A new universal resample-stable bootstrap-based stopping criterion for PLS component construction

Jérémy Magnanensi[1] · Frédéric Bertrand[2] · Myriam Maumy-Bertrand[2] · Nicolas Meyer[3]

Received: 30 April 2015 / Accepted: 22 March 2016 / Published online: 22 September 2016

Abstract We develop a new robust stopping criterion for partial least squares regression (PLSR) component construction, characterized by a high level of stability. This new criterion is universal since it is suitable both for PLSR and extensions to generalized linear regression (PLSGLR). The criterion is based on a non-parametric bootstrap technique and must be computed algorithmically. It allows the testing of each successive component at a preset significance level α. In order to assess its performance and robustness with respect to various noise levels, we perform dataset simulations in which there is a preset and known number of components. These simulations are carried out for datasets characterized both by $n > p$, with n the number of subjects and p the number of covariates, as well as for $n < p$. We then use t-tests to compare the predictive performance of our approach with other common criteria. The stability property is in particular tested through re-sampling processes on a real allelotyping dataset. An important additional conclusion is that this new criterion gives globally better predictive performances than existing ones in both the PLSR and PLSGLR (logistic and poisson) frameworks.

Keywords Bootstrap · PLSR · PLSGLR · Latent variable · Robustness

1 Introduction

Modeling relationships using traditional statistical methods like ordinary least squares regression (OLSR), between a univariate response and highly correlated covariates, is rarely recommended, and for datasets including more covariates than subjects, is not even possible. However, with recent technological and computing advances, providing consistent analysis of such datasets has become a major challenge, particularly in domains such as medicine, biology and chemometrics. For such reasons, statistical methods have been developed, including partial least squares regression (PLSR), introduced by Wold et al. (1983) and described in detail by Höskuldsson (1988), amongst others. PLSR has become a standard tool in chemometrics (Wold et al. 2001) and for dealing with genomic datasets (Boulesteix and Strimmer 2007). Indeed, due to its appealing properties, PLSR is able to efficiently deal with high-dimensional settings, and notably, resolves the collinearity problem (Wold et al. 1984).

In this paper, we focus on the PLS univariate response framework, better known as PLS1. Let n be the number of observations and p the number of covariates. Then, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ represents the response vector, with $(.)^T$ denoting the transpose, and $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_p\right) \in \mathcal{M}_{n,p}(\mathbb{R})$ the covariate matrix, with $\mathcal{M}_{n,p}(\mathbb{R})$ the set of matrices with n rows and p columns. Note that without loss of generality, $\mathbf{X}$ and $\mathbf{y}$ are supposed centered, and scaled to unit variance. PLSR consists of building $K \leqslant \mathrm{rk}(\mathbf{X})$ orthogonal latent vari-

Electronic supplementary material The online version of this article (doi:10.1007/s11222-016-9651-4) contains supplementary material, which is available to authorized users.

✉ Frédéric Bertrand
frederic.bertrand@math.unistra.fr

[1] Institut de Recherche Mathématique Avancée, LabEx IRMIA, Laboratoire de Biostatistique et Informatique Médicale, EA3430, Université de Strasbourg et CNRS, Strasbourg, France

[2] Institut de Recherche Mathématique Avancée, UMR 7501, LabEx IRMIA, Université de Strasbourg et CNRS, Strasbourg, France

[3] Laboratoire de Biostatistique et Informatique Médicale, LabEx IRMIA, EA3430, Faculté de Médecine, Université de Strasbourg, Strasbourg, France

ables $\mathbf{T}_K = (\mathbf{t}_1, \ldots, \mathbf{t}_K)$, also called components or scores vectors, in such a way that $\mathbf{T}_K$ optimally describes the common information space between $\mathbf{X}$ and $\mathbf{y}$. In order to do so, these components are built up as linear combinations of the original covariates, i.e.,

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k, \ 1 \leqslant k \leqslant K, \tag{1}$$

where $\mathbf{X}_0 = \mathbf{X}$, and $\mathbf{X}_{k-1}, \ k \geqslant 2$, represents the residual covariate matrix obtained through the OLSR of $\mathbf{X}$ on $\mathbf{T}_{k-1}$. $\mathbf{w}_k = \left(w_{1k}, \ldots, w_{pk}\right)$ is obtained as the solution of the following maximization problem (Boulesteix and Strimmer 2007):

$$\mathbf{w}_k = \underset{\mathbf{w} \in \mathbb{R}^p}{\operatorname{argmax}} \left\{\operatorname{Cov}^2\left(\mathbf{y}_{k-1}, \mathbf{t}_k\right)\right\} \tag{2}$$

$$= \underset{\mathbf{w} \in \mathbb{R}^p}{\operatorname{argmax}} \left\{\mathbf{w}^T \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \mathbf{y}_{k-1}^T \mathbf{X}_{k-1} \mathbf{w}\right\}, \tag{3}$$

with the constraint $\|\mathbf{w}_k\|_2^2 = 1$, and where $\mathbf{y}_0 = \mathbf{y}$, and $\mathbf{y}_{k-1}$ represents the residual vector obtained from the OLSR of $\mathbf{y}$ on $\mathbf{T}_{k-1}$.

These components can also be directly linked to the original covariate matrix:

$$\mathbf{t}_k = \mathbf{X}\mathbf{w}_k^* = \sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j, \ 1 \leqslant k \leqslant K, \tag{4}$$

where $\mathbf{w}_k^* = \left(w_{1k}^*, \ldots, w_{pk}^*\right)^T$ is the vector of the original covariates' weights, dependent on $\mathbf{y}$ (Wold et al. 2001). As demonstrated by Tenenhaus (1998, p. 114), by noting $\mathbf{W}_k^* = \left(\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*\right) \in \mathcal{M}_{p,k}(\mathbb{R})$, this matrix satisfies the following equation:

$$\mathbf{W}_k^* = \mathbf{W}_k \left(\mathbf{P}_k \mathbf{W}_k^T\right)^{-1}, \tag{5}$$

where $\mathbf{P}_k = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in \mathcal{M}_{p,k}(\mathbb{R})$ is the matrix containing the k vectors of regression coefficients from the OLSR of $\mathbf{X}$ on $\mathbf{T}_k$, also known as $\mathbf{X}$-loadings.

Let K be the selected number of components. The final regression model is thus:

$$\mathbf{y} = \sum_{k=1}^{K} c_k \mathbf{t}_k + \epsilon \tag{6}$$

$$= \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j\right) + \epsilon \tag{7}$$

$$= \sum_{j=1}^{p} \beta_j^{PLS} \mathbf{x}_j + \epsilon, \tag{8}$$

with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ the $n \times 1$ error vector and $(c_1, \ldots, c_K)$ the regression coefficients from the OLSR of $\mathbf{y}$ on $\mathbf{T}_K$, also known as $\mathbf{y}$-loadings.

In order to take into account specific distributions linked to the response, an extension to the generalized linear regression method, noted PLSGLR, was introduced by Marx (1996). This led to further research and developments related to the field (Nguyen and Rocke 2002; Boulesteix 2004; Ding and Gentleman 2005). Note that in this case, $\mathbf{y}$ is naturally not centered or scaled to unit variance. In this paper, the process developed by Bastien et al. (2005) and implemented in the R package *plsRglm* (Bertrand et al. 2014) is used. In this context, the regression model is the following:

$$g(\theta) = \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j\right), \tag{9}$$

with θ the conditional expected value of $\mathbf{y}$ for a continuous distribution, or the probability vector of a discrete distribution with a finite support. The link function g depends on the distribution of $\mathbf{y}$.

As mentioned above, both PLSR and its extension to generalized models rely on determining a tuning parameter: the number of components. The obtention of an optimal number of components K_{opt} is crucial to get reliable estimations of the original covariates' regression coefficients. Concluding that $K < K_{opt}$ leads to a loss of information, meaning that connections between some covariates and $\mathbf{y}$ are not correctly modeled. Concluding that $K > K_{opt}$, i.e., over-fitting, can lead to models with poor predictive ability (Wiklund et al. 2007).

Despite the fact that PLSR has become a versatile and standard tool in many domains like chemometrics, bioinformatics, medicine and social science (Rosipal and Krämer 2006), choosing well the number of components is still an open and important problem (Wiklund et al. 2007). Indeed, the relative lack of theoretical hypotheses, leading PLSR to be called a *soft-modeling* process (Manne 1987), precludes the development of typical statistical tests based on theoretical distributions for testing parameters (Wakeling and Morris 1993). Therefore, a substantial number of papers deal with this question by introducing new statistics or comparing several statistics' abilities. Most developed criteria are based on the predictive residual error sum of squares (PRESS), introduced by Allen (1971) for model selection. To be calculated, this statistic ideally needs an independent test set. However, notably due to logistical constraints, this additional set is rarely available (Efron and Tibshirani 1993, p. 240). Therefore, cross-validation (CV) techniques are usually used to obtain an estimation of PRESS-based statistics. Issues concerning CV methods for establishing prediction ability are reported in the literature, notably linked to the high variability of obtained results (Efron and Tibshirani 1993, p. 255; Hastie et al. 2009, p. 249; Wiklund et al. 2007, p. 429; Boulesteix 2014). Such issues are observed in this paper. An alternative to CV methods is the well-known bootstrap technique introduced by Efron (1979). Using this process for estimating prediction errors has already been proposed, notably by Efron and Tibshirani (1993), and also adapted to selecting the optimal number of components in PLS and principal component regression (PCR) (Wehrens and Linden 1997; Denham 2000; Amato and Vinzi 2003; Mevik and Cederkvist 2004). However, it has also been established that though the use of the bootstrap for predictive error estimation efficiently reduces the variability issue, it can also lead to large bias (Efron and Tibshirani 1993, p. 255; Kohavi 1995). Much further literature is also available, introducing new criteria or comparing criteria: Höskuldsson (1996), Van der Voet (1994), Li et al. (2002), Green and Kalivas (2002), Gourv́enec et al (2003), and Gómez-Carracedo et al. (2007) are some examples. Performing a global state-of-the-art review on this subject would be difficult due to the vast number of previous works. However, it is clear that there is not yet one precise criteria that can be considered reliable *in general*. In the PLSGLR framework, it is also notable that very few criteria adapted to this situation have been proposed, and none of them can currently be considered as a good general benchmark.

The aim of the article is twofold. First, we wish to establish a new criterion that can be considered universal, i.e., both reliable and easily adaptable to both the PLSR and PLSGLR frameworks. To the best of our knowledge, no previous criteria features this property. Second, this new criterion has to avoid CV methodology and related issues such as instability. Therefore, we develop a new bootstrap-based criterion to select the number of PLS components. The originality of the approach is due to the fact that it tests directly both the **X**- and **y**-loadings. To do this, the establishment of bootstrap-based confidence intervals (CI) is achieved. By focusing on the unknown distribution of the regression coefficients rather than predictive error-based statistics as previously proposed, we open up the possibility of directly testing the significance of successive components, which is pertinent for both the PLSR and PLSGLR frameworks. This method avoids the use of CV techniques and related issues.

In this article, we first explain the context and give theoretical details, before introducing the new algorithmic bootstrap-based criterion as pseudo-code in Sect. 2. In Sect. 3, we present existing criteria that have been chosen for comparison purposes, and then describe the simulation set-up we use to make comparisons. In Sect. 4, we analyze results obtained in the PLSR framework, followed by PLSGLR results for logistic regression (PLS-LR) and Poisson regression (PLS-PR) in Sect. 5. In Sect. 6, we focus on some real datasets and compare our new criterion to relevant existing ones. Using a real allelotyping dataset, we also compare the robustness of our new bootstrap-based criterion through resampling, approximating the distribution of the extracted number of components. Lastly, in Sect. 7, we discuss the observed advantages and disadvantages of each criterion.

2 A new boostrap based stopping criterion

2.1 Context

As mentioned in Sect. 1 and to the best of our knowledge, no criterion for tuning the number of components can currently be considered the benchmark. In addition, being derived from CV and thus linked to issues discussed in Sect. 1, known criteria are often based on arbitrary or empirical threshold values (Krzanowski 1987), or theoretical asymptotic distributions (Haaland and Thomas 1988; Osten 1988), which are not appropriate for general reliable establishment of PLS models. For such reasons, we have developed a new criterion which is not based on CV processes and does not depend on arbitrary threshold values. Futhermore, our aim is not to directly focus on predictive ability-based statistics (already well-developed in the literature), but rather on scores vectors themselves, by searching for a way to test their significance, like is done for OLSR with Student-type tests. However, as PLSR methodology is a soft-modeling process (1), no such global distribution can be used.

The bootstrap is a well-known method for approximating unknown distributions. Bootstrap techniques adapted to the regression framework have already been proposed by Efron (1979) and Freedman (1981). As a bootstrap-based criterion could be a useful way to avoid CV, it was proposed for PLS component selection, notably by Denham (2000) and Amato and Vinzi (2003). However, to the best of our knowledge, a bootstrap-based process has never been used in order to test the various loadings involved, and represents an option for choosing an optimal number of PLS components, which covers all our goals.

2.2 Bootstrapping pairs in PLSR

Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T = (\mathbf{y}, \mathbf{X}) \in \mathcal{M}_{n,p+1}(\mathbb{R})$, i.e., $\mathbf{z_i} = \left(y_i, x_{i1}, \ldots, x_{ip}\right)$, $1 \leqslant i \leqslant n$. The so-called bootstrapping pairs method was introduced by Freedman (1981) and consists of building R new datasets by re-sampling with replacement in $\mathbf{Z}$ in order to mimic the generation of the original data. This leads to an empirical approximation of the distribution linked to a statistic $\mathcal{S}(\mathbf{Z})$. This technique only relies on the assumption that the originals pairs $(y_i, \mathbf{x}_{i\bullet})$, where $\mathbf{x}_{i\bullet}$ represents the ith row of $\mathbf{X}$, are randomly sampled from some unknown $(p+1)$-dimensional distribution. It was developed to treat so-called correlation models, in which covariates are considered as random, and ϵ may be related to them. In this way, it is appropriate to "estimate the regression plane for a certain population on the basis of a simple random sample" (Freedman 1981, p. 1219).

Constructing a new component $\mathbf{t}_k$ as described in Sect. 1 implies that $\mathbf{w}_k = \frac{\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}}{\|\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}\|}$. This property leads to the following result.

Proposition 1 *Let* $\mathbf{y}_0 = \mathbf{y}$ *and* $\mathbf{X}_0 = \mathbf{X}$. *Let* $\mathbf{y}_{k-1}$ *and* $\mathbf{X}_{k-1}$, $k \geqslant 2$, *be respectively the residual response vector and covariate matrix obtained through both the OLSR of* $\mathbf{y}$ *and* $\mathbf{X}$ *on* $\mathbf{T}_{k-1}$. *Suppose that,* $\forall k \in [\![1, K]\!], \exists i \in [\![1, p]\!], \mathbf{x}_{i,(k-1)}^T \mathbf{y}_{(k-1)} \neq 0$.

Then, the PLS component building process implies that: $\forall k \in [\![1, K]\!]$, $c_k > 0$ *and, conditionally on* $\mathbf{X}$, c_k *follows a positive distribution.*

As a consequence of this result, bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ by applying PLSR to each bootstrap sample in order to test Y-loadings, is not straightforward.

Furthermore, this method does not appear to be relevant since it approximates the uncertainty of the subspace spanned by the scores vectors, though this is not the initial aim. Our goal is to test particular PLS components based on the original dataset, since these latent variables are built and used for modeling specifically these original data. In other words, a method able to test the significance of these particular random latent variables, defined as specific linear combinations obtained through the PLSR processed on the original dataset, is to be looked for. As a concrete example, let $\mathbf{t}_1^{ori} = \sum_{j=1}^{p} w_{j1}^{ori} \mathbf{x}_j$ be the first PLS component based on the original data. By bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ and applying the PLS process to a bootstrap sample $(\mathbf{y}, \mathbf{X})^b$, the obtained weights $\mathbf{w}_1^b$ are naturally different from $\mathbf{w}_1^{ori}$, so the uncertainty of the specific random variable $\sum_{j=1}^{p} w_{j1}^{ori} \mathbf{x}_j$ is not tested by this ill-adapted process, but rather uncertainty about the construction of this first component.

To succeed in testing these specific components, a bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ process has to be performed, while keeping fixed the weights $\mathbf{W}_k^{ori}$ obtained on the original data, for the construction of the components linked to each bootstrap sample. Thus, the specific uncertainty of the particular linear combination of the original variables is approximated. Performing this process is equivalent to bootstrapping pairs $\left(y_i, \mathbf{T}_{k,i\bullet}\right)$, where $\mathbf{T}_{k,i\bullet}$ represents the ith row of $\mathbf{T}_k$, i.e., sampling from an empirical distribution conditional on the scores vectors $\mathbf{T}_k$.

As the PLS components are built both for modeling the response and summarizing the original relevant information in $\mathbf{X}$, we propose to test each new component $\mathbf{t_k}$ by approximating the conditional distribution of the $\mathbf{X}$- and $\mathbf{y}$-loadings given $\mathbf{T}_k$. This is done by bootstrapping pairs $\left(y_i, \mathbf{T}_{k,i\bullet}\right)$ and $\left(\mathbf{x}_{ij}, \mathbf{T}_{k,i\bullet}\right)$, $\forall j \in [[1, p]]$. We also propose to define the significance of a new component in terms of its significance for both $\mathbf{y}$ and $\mathbf{X}$, so that the extracted number of components K is defined as the last one which is significant for both.

2.3 Adapted bootstrapping pairs as a new stopping criterion

Based on our definition of the significance of a new component, a double bootstrapping pairs algorithm was constructed. The first step consists of bootstrapping pairs $\left(\mathbf{x}_{ij}, \mathbf{T}_{k,i\bullet}\right)$, $\forall j \in [[1, p]]$. We propose that a component is considered significant for $\mathbf{X}$ if and only if it is significant for at least one of the original covariates. Components are successively tested until we reach the first non-significant one. This step leads to a maximal number of components $k_{\max}$ that can be extracted. The second step consists of bootstrapping pairs $\left(y_i, \mathbf{T}_{k,i\bullet}\right)$ to test the significance against $\mathbf{y}$ of each successive component $\mathbf{t}_k$, with $k \leqslant k_{\max}$. To avoid confusion between the number of covariates and $\mathbf{X}$-loadings, we set m as the total number of original covariates.

The algorithm of this double bootstrapping pairs implementation is thus as follows:

```
I  Bootstrapping $\left(\mathbf{X}_{i\bullet}, \mathbf{T}_{k,i\bullet}\right)$:
   Let $k = 0$.
   Repeat
     1 $k = k + 1$.
     2 Compute the $k$th component, defining $\mathbf{T}_k = (\mathbf{t}_1, \ldots, \mathbf{t}_k)$.
     3 Bootstrap pairs $\left(\mathbf{X}_{i\bullet}, \mathbf{T}_{k,i\bullet}\right)$, returning $R$ bootstrap samples:
         $(\mathbf{X}, \mathbf{T}_k)^{b_1}, \ldots, (\mathbf{X}, \mathbf{T}_k)^{b_R}$.
     4 For each $(\mathbf{X}, \mathbf{T}_k)^{b_r}$, do $m$ OLS regressions:
         $\mathbf{x}_l^{b_r} = \sum_{j=1}^{k} \left( \hat{p_{lj}}^{b_r} . \mathbf{t}_j^{b_r} \right) + \hat{\delta}_{lk}^{b_r}$.
     5 $\forall p_{lk}$, construct a $(100 \times (1 - \alpha))$ % bilateral $BC_a$ CI, noted:
         $\mathrm{CI}_l = \left[ \mathrm{CI}_{l,1}^k, \mathrm{CI}_{l,2}^k \right]$.
   Until $\forall l \in \{1, \ldots, m\}, 0 \in \mathrm{CI}_l$.
   Return $k_{\max} = k - 1$ and $\mathbf{T}_{k_{\max}}$.
II Bootstrapping $\left(y_i, \mathbf{T}_{k,i\bullet}\right)$:
   Note that for the PLSGLR case, the relevant generalized regression is performed at step 9.
   Let $k = 0$.
```

```
Repeat
  6  $k = k + 1$.
  7  Compute $\mathbf{T}_k$ by extracting the $k$ first columns from $\mathbf{T}_{k\text{max}}$.
  8  Bootstrap pairs $\left(y_i, \mathbf{T}_{k,i\bullet}\right)$, returning $R$ bootstrap samples:
       $(\mathbf{y}, \mathbf{T}_k)^{b_1}, \ldots, (\mathbf{y}, \mathbf{T}_k)^{b_R}$.
  9  For each pair $(\mathbf{y}, \mathbf{T}_k)^{b_r}$, do the OLS regression:
       $\mathbf{y}^{b_r} = \sum_{j=1}^{k} \left(\hat{c}_j^{b_r}.\mathbf{t}_j^{b_r}\right) + \hat{\epsilon}_k^{b_r}$.
  10 Since $c_k > 0$, construct a $(100 \times (1 - \alpha))$ % unilateral $BC_a$ CI:
       $\text{CI} = \left[\text{CI}_1^k \; +\infty\right[$ for $c_k$
While $\text{CI}_1^k > 0$ and $k \leqslant k_{\text{max}}$.
Return the final extracted number of components $K = k - 1$.
```

Results linked to this bootstrap-based criterion are referred to as **BootYT** in the following.

3 Simulation

3.1 Existing criteria used for comparison

To perform our benchmarking study, several existing criteria were used.

In the PLSR framework, the Q^2 criterion was selected since it represents a standard criterion, implemented notably in both the R package *plsRglm* (Bertrand et al. 2014) and the SIMCA-P software (Umetrics 2005). This criterion is based on q-fold CV methods (Breiman et al. 1984) and was computed for both $q = n$, leading to the universal standard CV method called *leave-one-out* CV (Gómez-Carracedo et al. 2007), and $q = 5$ (5-CV) following recommendations of Kohavi (1995) and Hastie et al. (2009), so as to reduce variability in the CV method. The BIC criteria, corrected with the estimated degrees of freedom (DoF) by Krämer and Sugiyama (2011), was also included since to the best of our knowledge, no published study has analyzed its performance.

In the PLSGLR framework, there are a limited number of relevant criteria available; we thus present only two here: the number of misclassified values (Meyer et al. 2010), and a criterion introduced by Bastien et al. (2005). Both are available in the R package *plsRglm*. The usual AIC and BIC criteria were also included.

– In PLSR:

1 $\mathbf{Q^2}$. For each new component $\mathbf{t}_k$, the following statistic is evaluated:

$$Q_k^2 = 1 - \frac{\text{PRESS}_k}{\text{RSS}_{k-1}},$$

where RSS_{k-1} represents the Residual Sum of Squares when the number of components is $k-1$, and PRESS_k the PRESS when the number of components is equal to k. Tenenhaus (1998) considers that a new component $\mathbf{t}_k$ improves significantly the prediction of $\mathbf{y}$ if:

$$\sqrt{\text{PRESS}_k} \leqslant 0.95\sqrt{\text{RSS}_{k-1}} \iff Q_k^2 \geqslant 0.0975.$$

Results linked to this criterion using both leave-one-out and $q = 5$ CV are referred to in the text and plots by **Q2lv1o** and **Q2K5** respectively.

2 **BIC**. The R package *plsdof*, based on the work of Krämer and Sugiyama (2011), was used to compute this criterion. It works as follows:

$$\text{BIC} = \text{RSS}/n + log(n)(\gamma/n)\widehat{\sigma}_\epsilon^2,$$

where γ represents the DoF of model (8) and $\widehat{\sigma}_\epsilon^2$ is defined by Krämer and Sugiyama (2011).
The selected model is the one which represents the first local minimum of this adapted BIC criterion; related results are referred to by **BICdof**. Results linked to models obtaining the global minimum are also returned under the acronym **BICglob**.

– In PLSGLR:

1 **AIC**. The AIC criterion (Akaike 1974) can be computed whatever the distribution involved. However, no corrected DoF have yet been suggested for the PLSGLR framework.

2 **BIC**. As in the case of AIC, the BIC (Schwarz 1978) is calculable without correcting the DoF.

3 **CV − MClassed**. This criterion can only be used for PLS-LR. Via 5-CV, it determines for each model the number of misclassified predicted values. The selected model is the one corresponding to this statistic's minimal value.

4 **p_val**. Bastien et al. (2005) define a new component $\mathbf{t}_k$ as non-significant if it contains no significant covariate. Asymptotic Wald tests are used to conclude as to the significance of the various covariates.

3.2 Simulation plan

To compare these criteria, simulations were performed by adapting the *simul_data_UniYX* function, available in the R package *plsRglm* (Bertrand et al. 2014). First, four orthonormal vectors of size p are built. Let $\mathbf{T} \in \mathcal{M}_{4\times p}(\mathbb{R})$ be the matrix containing them. Then, rows of $\mathbf{X}$ are successively obtained using $\mathbf{X}_{i\bullet} = \mathbf{R}_i\mathbf{T}+\epsilon_i$, where $\mathbf{R}_i = (r_{1i}, \ldots, r_{4i}) \in \mathbb{R}^4$ is a vector of random draws from $\mathcal{N}\left(0, \sigma_j\right)$, $j = 1, \ldots, 4$ respectively, with $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (10, 8, 6, \sigma_4)$, and ϵ_i is drawn from $\mathcal{N}(0, 10^{-2})$. Only the first three orthonormal vectors are linked to the simulated response, so that σ_4 varies during the simulations in order to understand the impact of increasing variability of this uninformative fourth component on the various criteria. Different processes were used to simulate response vectors, depending on the desired distribution. As a constant in all simulation schemes, the first three orthonormal vectors are involved, so that whatever the framework, simulations are performed to obtain a relevant subspace of dimension 3. Also, a noise parameter in $\mathbf{y}$ helps us to determine the robustness of the criteria being examined with respect to increasing values of σ_5, which characterizes its standard deviation. Fixed sets of values for σ_4 and σ_5 are given, depending on the framework, and described in the corresponding sections. For more details about these simulation processes, see Supplementary Materials 3.

Simulations were performed for two different cases, for both the PLSR and PLSGLR frameworks. The first was the $n > p$ situation with $n = 200$ and $p \in \Omega_{200} = [\![7, 50]\!]$. The second was the $n < p$ situation where $n = 20$ and $p \in \Omega_{20} = [\![25, 50]\!]$. For each fixed pair (σ_4, σ_5), which represents the standard deviation of the uninformative fourth component in $\mathbf{X}$ and the additional random noise standard deviation in $\mathbf{y}$, respectively, we simulated 100 datasets. Each dataset is based on p_l covariates, $1 \leqslant l \leqslant 100$. The p_l numbers are obtained by sampling with replacement in Ω_n. Testing these criteria on 100 different datasets allows us to calculate a mean value of the number of components for each fixed couple (σ_4, σ_5) as well as an estimated variance that represents the inherent stability of the various criteria. Lastly, the number of bootstrap replicates was fixed at $R = 500$ and CI were constructed by setting $\alpha = 0.05$. More in-depth details of the data simulation framework is available in Supplementary Materials 3.

The aim of the simulations is twofold. First, a comparison of the chosen criteria, both through their results for the number of components and their robustness against different random noise variances. Second, predictive abilities are compared using predictive normalized mean squared errors (PNMSE), calculated on 80 additional simulated samples per dataset in the $n < p$ framework, used as test sets. Normalization is performed by dividing the predictive mean squared errors (PMSE) related to the obtained model by the PMSE linked to the trivial one (constant model equal to the mean of the training data). Furthermore, as mentioned in Krämer and Sugiyama (2011, p. 702), “the large test sample size ensures a reliable estimation of the test error.” Then, for each pair of values (σ_4, σ_5), asymptotic t-tests with Welch-Satterthwaite DoF approximation (Welch 1947) are performed to compare the PNMSE averages over the 100 simulated datasets related to each criterion. All tests have been run at level $\alpha = 0.05$.

4 PLSR results

As mentioned in Sect. 3.2, the simulated subspace is spanned by three orthonormal vectors (components). By modeling using uninformative elements in $\mathbf{X}$, a model based on four components is thus overfitted. Any supplementary component will be built from random noise present in $\mathbf{X}$.

4.1 Initial selection

To select the best method, both between the Q2lv1o and the Q2K5 criteria, and between the BICdof and BICglob ones, results related to datasets with $n > p$, for the following sets of values for noise standard deviations (NSD), are considered:

$$(A) : \begin{cases} \sigma_4 \in \{0.01, 0.21, \ldots, 5.81\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}\,. \end{cases}$$

The averages of the selected numbers of components over the 100 simulated datasets per couple are calculated. These averages, denoted by *nb_comp* and related to the BIC and Q^2 criteria, are presented graphically in Figs. 1 and 2 respectively as functions of σ_4 and σ_5, respectively denoted *sigma4* and *sigma5*.

Based on results shown in Fig. 1, BICglob has stability issues. We observe that this is mainly due to the adapted DoF not necessarily increasing as the number of components rises. Therefore, since adding a component can surprisingly lead to smaller DoF, this criterion is related to both overdetermination and stability issues. The BICdof process, by searching for the first local minimum of the adapted BIC criterion, allows us to focus on the comparison between models related to k components, $1 \leqslant k \leqslant K + 1$. Based on our observations, these successive models are mainly linked to increasing DoF, avoiding issues related to the BICglob process. Therefore, the BICglob criterion should be avoided, and we retain BICdof for further comparisons.

Concerning the Q^2 criterion, results displayed in Fig. 2 point to a negligible effect of different values of q on these average numbers of components. Since reducing the value of q implies a variance decrease in related results (Hastie et al. 2009, p. 243), the Q2K5 criterion is retained here for further comparisons.

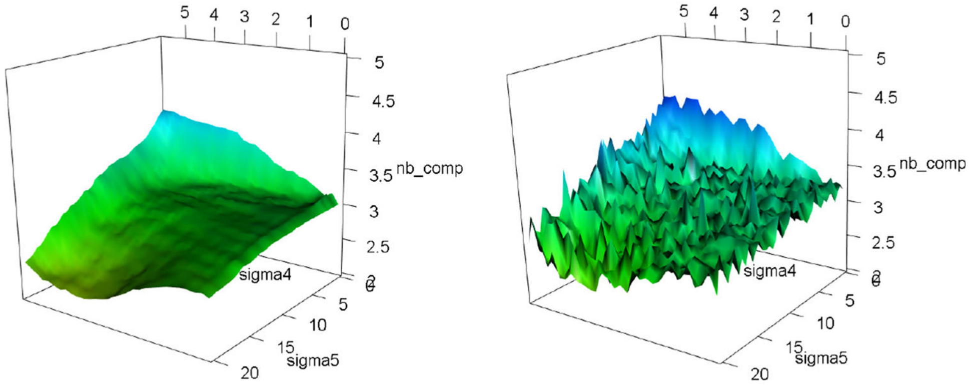


Fig. 1 PLSR, $n > p$, sets of NSD values from (A), evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5) for *BIC* based criteria, *left* BICdof, *right* BICglob

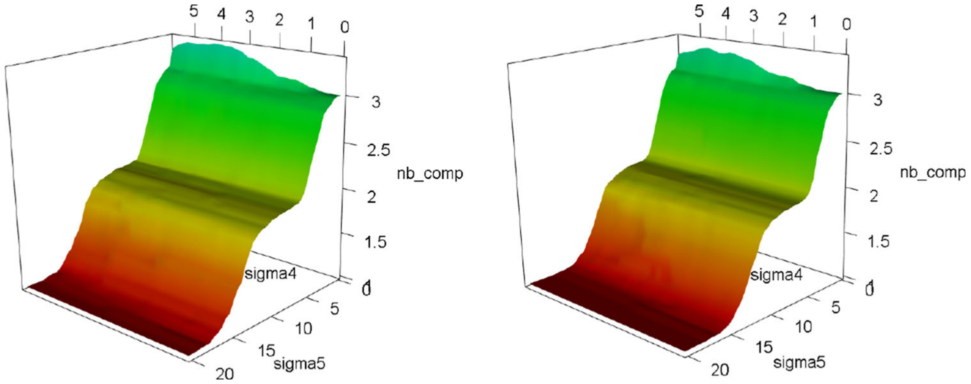


Fig. 2 PLSR, $n > p$, sets of NSD values from (A), evolution of averages of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5) for Q^2 based criteria, *left* Q2lv1o, *right* Q2K5

In light of these initial observations, only three methods are retained: Q2K5, BICdof and our new bootstrap-based criterion.

4.2 PLSR: the $n > p$ case

To compare the three retained methods when $n > p$, the following enlarged sets of values for NSD are considered:

$$(B) : \begin{cases} \sigma_4 \in \{0.01, 0.21, \ldots, 5.81\} \cup \{6.01, 7.01, \ldots, 30.01\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\} \end{cases}$$

The means of number of components over the 100 simulated datasets per pair (σ_4, σ_5) are displayed for the three criteria in Fig. 3. Variances of these numbers of components over the 100 simulated datasets per pair were also estimated, and are shown using boxplots in Fig. 4. Note that these variances approximate the inter-dataset variability for fixed values of σ_4 and σ_5, not the intra-dataset one.

In these results, we see that the Q2K5 criterion is the least robust against increasing noise variability in $\mathbf{y}$, characterized by increasing values of σ_5 (sigma5). This lack of robustness leads it to globally underestimate the number of components. BICdof has a low computational requirements and is also the most robust against increasing values of σ_5. 86.37 % of all its selected numbers of components are equal to three or four. However, as seen in Fig. 4, the BICdof features the highest global variability in number of components selected over the 100 datasets involved per pair (σ_4, σ_5). This is even more acute for datasets characterized by a fourth component standard deviation that is higher than that involved in the relevant subspace, i.e.,

$$\sigma_4 > \sqrt{\sum_{i=1}^{3} \sigma_i^2} = \sqrt{200} \simeq 14.14. \tag{10}$$

In this particular case, our new bootstrap-based criterion retains stability, while the median of the BICdof results, for instance, more than triples (0.25 to 0.79) compared to that of the whole data. Moreover, BootYT is the most robust against increasing variability of the uninformative fourth component in $\mathbf{X}$.

As a preliminary conclusion based on these initial results, advising the use of a certain choice among the BICdof or BootYT criteria is not relevant in the $n > p$ case. Due to its lack of robustness against noise variability in $\mathbf{y}$, the Q2K5 criterion should be avoided.

4.3 PLSR: the $n < p$ case

As suggested by Krämer and Sugiyama (2011), a small training sample size allows us to consider high-dimensional settings.

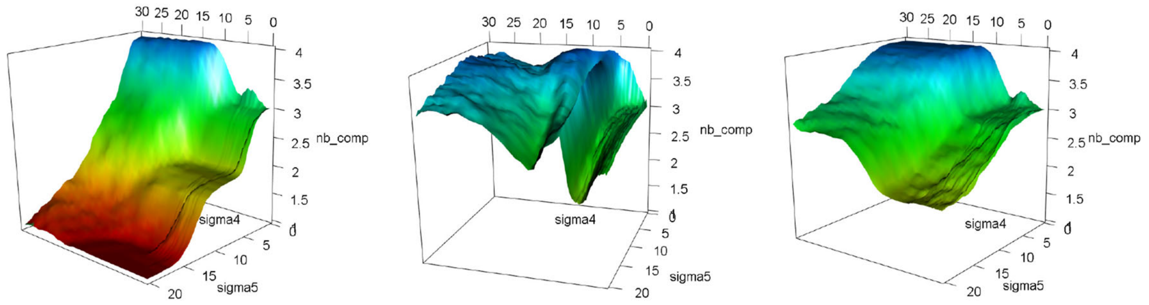
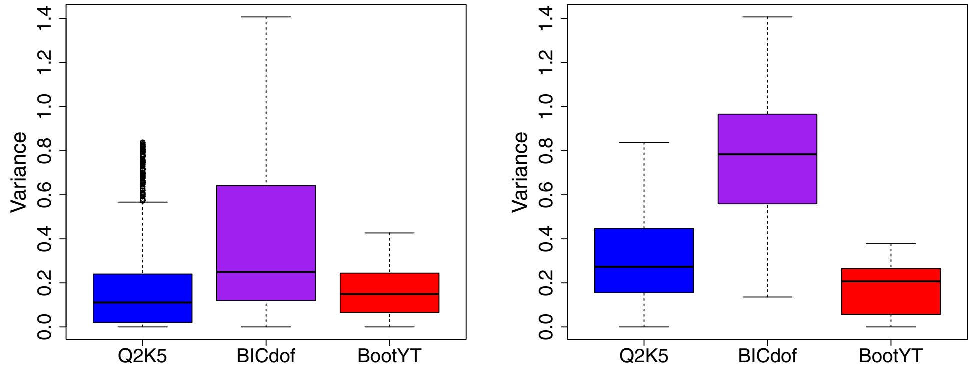


Fig. 3 PLSR, $n > p$, sets of NSD values from (B), evolution of averages of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); from *left* to *right*: Q2K5, BICdof and BootYT criteria

Fig. 4 PLSR, $n > p$, sets of NSD values from (B); *left* boxplots of estimated variance in the number of components over the 100 datasets per pair (σ_4, σ_5) for all involved values of σ_4 and σ_5, *right* boxplots of estimated variance in the number of components over the 100 datasets per pair (σ_4, σ_5) for all involved values of σ_5 and $\sigma_4 \geqslant 15.01$



4.3.1 Mean and variance analyses

In this $n < p$ framework, the following sets of values for NSD are considered for criteria comparison:

$$(C) : \begin{cases} \sigma_4 \in \{0.01, 1.01, \ldots, 6.01\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\} . \end{cases}$$

Averages of numbers of components over the 100 datasets per pair (σ_4, σ_5) are displayed in Fig. 5. Graphical representations of variances are also shown in Fig. 6.

In Fig. 5, we see that the BICdof appears to suffer from overfitting issues. Moreover, based on the results in Fig. 6, it returns results linked to out-of-range values of the variance, compared with the other two criteria. These two issues are mainly due to the extraction of 1678 (5.8 %) results equal to 19 components, whereas 26184 (91.2 %) trials give four or less components. By more carefully analyzing this phenomenon, it appears that the rate of 19 components is a globally decreasing function of σ_5. If these extreme results are considered non-representative of the criterion, the apparent lack of robustness, as well as the apparent over-fitting issues, may not be so important. However, these extreme results suggest inherent issues leading to a lack of reliability of this BIC criterion, and cannot be ignored.

Our new boostrap-based criterion underestimates the number of components but is robust to increasing noise levels in **y**, thus returning averages of number of components between 1.2 and 2.2. Moreover, the related low variance seen gives good evidence of stability. The Q2K5 criterion has comparable stability but is less robust to increasing noise levels in **y** than our criterion, meaning that in general, it is linked to significant under-fitting issues.

4.3.2 PNMSE analysis

The results of t-tests for PNMSE mean comparisons are shown in Fig. 7.

Results related to the smallest values of σ_5 require special consideration. Due to the consequent lack of noise in **y**, models related to an over-determined number of components are not linked to the usual poor predictive ability issue since these supplementary scores vectors only try to model negligible noise. This implies that PNMSE are globally subject to the same rule as the MSE, i.e., the higher the number of components, the lower the PNMSE. As a direct consequence, the BICdof, which globally leads to over-fitted models (Fig. 5), returns by far the lowest PNMSE. This fact lead to only focus on the extracted number of components when $\sigma_5 \simeq 0$, so the Q2K5 criterion is to be advised in this particular case. However, such noiseless properties are rarely satisfied in real datasets. In all other cases, the BootYT criterion returns models which are at least comparable if not better predictive performance than the other two.

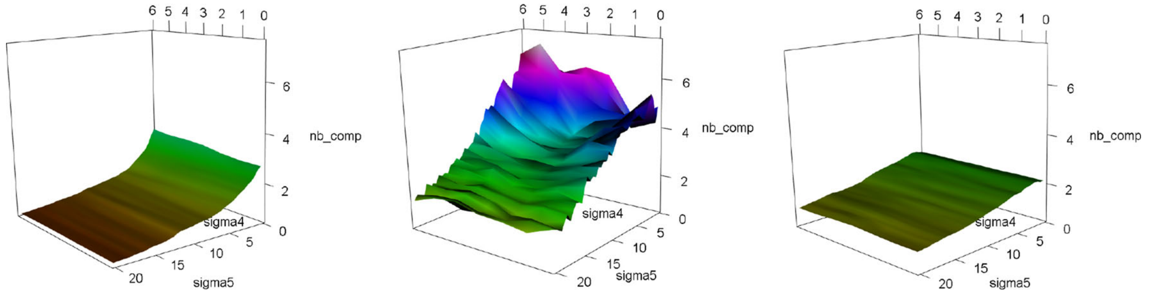
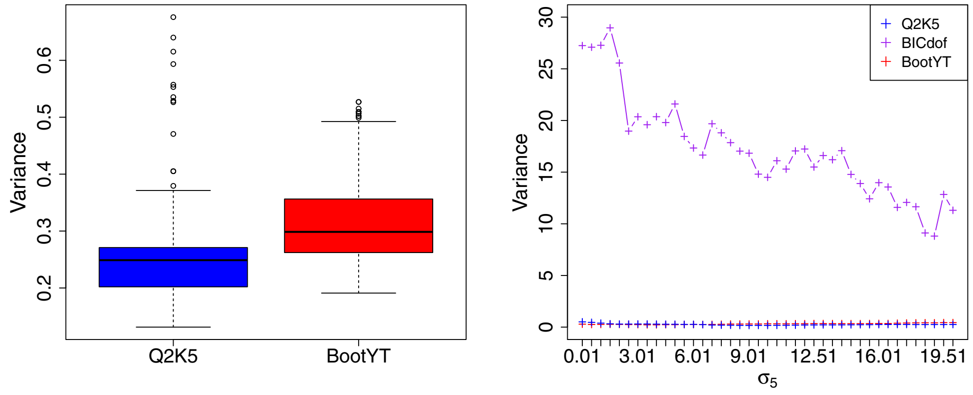


Fig. 5 PLSR, $n < p$, sets of NSD values from (C), evolution of averages of selected number of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); from *left* to *right*: Q2K5, BICdof and BootYT criteria

Fig. 6 PLSR, $n < p$, sets of NSD values from (C); *left* boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5) for both the Q2K5 and BootYT criteria, *right* evolution of estimated variances as a function of σ_5 for the three criteria studied



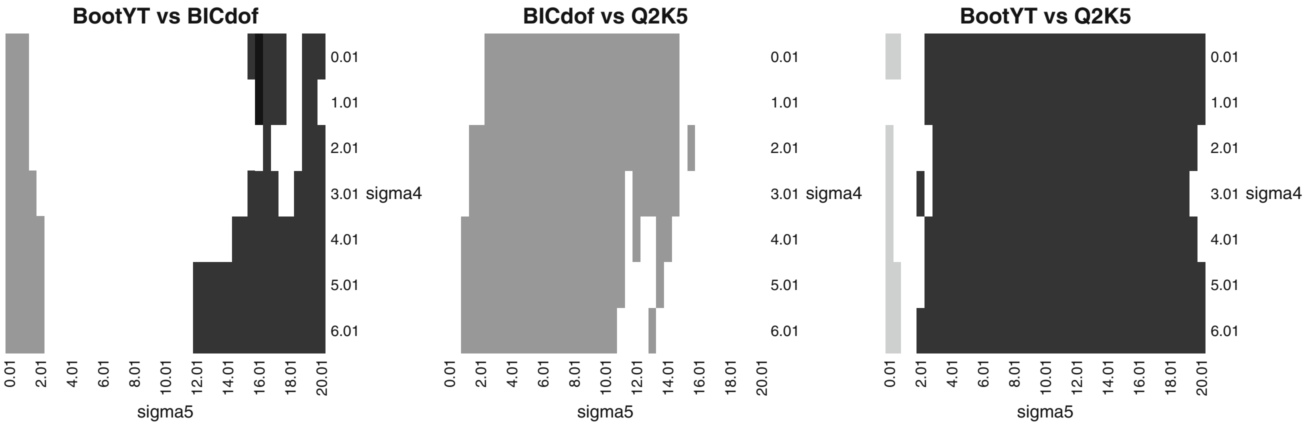


Fig. 7 PLSR, $n < p$, sets of NSD values from (C); graphical representation of t-test results for PNMSE averages comparison; color code: BootYT better (*black*), BICdof better (*dark gray*), Q2K5 better (*light gray*), no significant difference (*white*)

4.4 PLSR: conclusion

As a global conclusion, and in light of the results shown, the BootYT criterion can be seen as an interesting compromise between the other two in the $n > p$ framework, retaining their advantages but without their drawbacks. Indeed, this criterion offers both better robustness than Q2K5 against noise variability in $\mathbf{y}$, and better robustness than BICdof against variability of the uninformative fourth component in $\mathbf{X}$. It also features appealing stability compared to that of BICdof, especially for high σ_4 values. Concerning the $n < p$ case, extreme numbers of components selected by the BICdof criterion means that it is not pertinent to compare it to the other methods. While it returns 19338 (67.380 %) results between two and four components, the over-determination issue cannot be ignored, while our criterion returns all its results below five. The BootYT criterion is also more robust against noise variability in $\mathbf{y}$ than Q2k5. Lastly, concerning predictive abilities, our new criterion has comparable if not better performance than the other two, with the exception of the case of negligible noise variability in $\mathbf{y}$, for which Q2K5 is advised. Recommendations are summarized in Table 1.

Table 1 Recommended criteria for PLSR

	$n > p$		$n < p$	
	Low σ_4 values	High σ_4 values	Low σ_4 values	High σ_4 values
Negligible σ_5 values	BootYT/BICdof	BootYT	Q2K5	Q2K5
Non-negligible σ_5 values	BootYT/BICdof	BootYT	BootYT	BootYT

5 PLSGLR results

In this section, results related to the comparison between the bootstrap-based criterion and four other criteria (AIC, BIC, CV-MClassed and p_val, see Sect. 3.1) are presented. Note that, in this framework, due to specific distributions linked to **y** and the link-functions g used, an increase in σ_5 does not lead to a linear increase of noise variance in **y** as it does for PLSR simulated datasets. However, the bijectivity of these link functions ensures that a spanned subspace of dimension three is extracted.

5.1 PLS-LR results

5.1.1 PLS-LR: the $n > p$ case

The following sets of values for NSD are considered for criteria comparison:

$$(D):\begin{cases}\sigma_4 \in \{0.01, 0.51, \ldots, 9.51\}\\ \sigma_5 \in \{0.01, 0.51, \ldots, 15.51\}.\end{cases}$$

Both the means and variances of numbers of components over the 100 datasets per paire (σ_4, σ_5) are displayed in Fig. 8.

Based on these, CV-MClassed performs well in estimating the optimal number of components, on average. However, the downside is the higher variances related to its results than those of the others. Therefore, this criterion should be used with caution. The BootYT and p_val criteria return similar results in the $n > p$ case. Both of them slightly underestimate the optimal number of components, but show stability in their results.

The uncorrected DoF lead the AIC and BIC criteria to globally overestimate the number of components (Supplementary Material 4). Thus, these criteria should be avoided until the development of a DoF correction in this PLSGLR framework, and will not be considered in the $n < p$ case.

5.1.2 PLS-LR: the $n < p$ case

Here, the following sets of values for NSD are considered for criteria comparison:

$$(E):\begin{cases}\sigma_4 \in \{0.01, 0.51, \ldots, 9.51\}\\ \sigma_5 \in \{0.01, 0.51, \ldots, 9.51\}.\end{cases}$$

Both the averages and variances of numbers of components over the 100 datasets per pair (σ_4, σ_5) are displayed in Fig. 9.

The CV-MClassed criterion retains both the same property of well estimating, on average, the number of components, and still has the variability issue. Concerning the two other criteria, we observe a greater underestimation issue linked to the p_val criterion than for BootYT. Furthermore, they both feature low variability.

5.1.3 PLS-LR: PNMSE and misclassified values analysis

Since the binary response obtained by the model is equal to 1 if the estimated response is over 0.5, 0 if not, returning

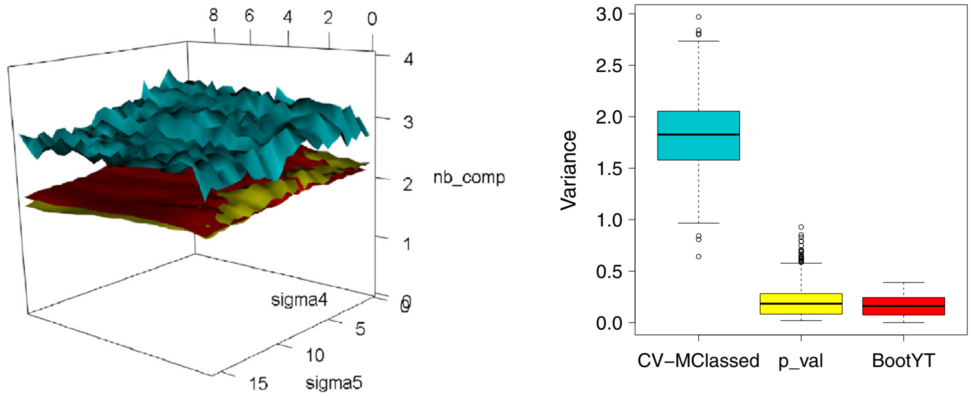


Fig. 8 PLS-LR, $n > p$, sets of NSD values from (D); *left* evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); *right* boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5)

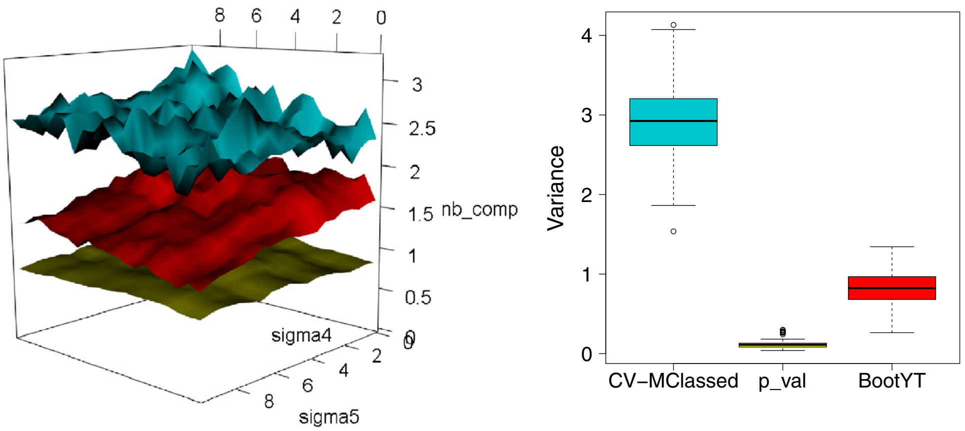


Fig. 9 PLS-LR, $n < p$, sets of NSD values from (E); *left* evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5), *right* boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5)

higher PNMSE does not necessarily lead to a higher number of misclassified values. Thus, we also computed the number of misclassified predicted values (*M_classed*) for each of the three criteria. The results of *t*-tests are shown in Fig.10.

The bootstrap-based criterion is never less efficient than the others. If there is globally no significant difference between bootstrapping pairs and the p_val criterion related to the PNMSE, BootYT performs better than it in terms of number of misclassified predictions. Next, there are only a few cases where bootstrapping pairs are significantly better than CV-MClassed for the number of misclassified predictions. But, in terms of the PNMSE, the BootYT criterion is better than the latter as it returns significantly smaller PNMSE values, especially for high values of σ_5.

5.1.4 PLS-LR: conclusion

From these simulations, it reasonable to assume that the bootstrap-based criterion is globally more efficient than the others. In the $n > p$ case, it has similar stability to p_val. However, it globally underestimates the optimal number of components, while CV-MClassed does not, but with high variability. As for the $n < p$ case, BootYT has better predictive performance than the two other criteria in terms of both PNMSE and predictive misclassified values. It also has low variability, important for any future implementation. Lastly, AIC and BIC are clearly not useful, since corrected DoF have not yet been established (Supplementary Material 4). These conclusions are summarized in Table 2.

5.2 PLS-PR results

5.2.1 PLS-PR: Row mean analysis

In the PLS-PR case, the following sets of values for NSD are considered for respectively the $n > p$ (F) and $n < p$ (G) cases:

$$(F):\begin{cases}\sigma_4 \in \{0.01, 0.51, \ldots, 9.51\}\\ \sigma_5 \in \{0.01, 0.21, \ldots, 2.21\} \cup \{2.51, 3.01, \ldots, 7.01\}\end{cases}$$

$$(G):\begin{cases}\sigma_4 \in \{0.01, 0.51, \ldots, 9.51\}\\ \sigma_5 \in \{0.01, 0.21, \ldots, 2.21\} \cup \{2.51, 3.01, \ldots, 5.01\}.\end{cases}$$

Averages of number of components over the 100 datasets per pair (σ_4, σ_5), related to the four criteria considered (AIC, BIC, p_val and BootYT), are shown in Fig. 11 for both the $n > p$ and $n < p$ frameworks.

Apart from the bootstrap-based criterion, all criteria return an increasing number of components as σ_5 increases. These results lead us to conclude that our new bootstrap-based stopping criterion is the only one which is relevant for Poisson distributions, in that it selects, on average, a decreasing number of components as σ_5 increases. Based on these plots, no additional analyses of the numbers of components was done. Only the two criteria that give results, on average, closest to the expected result, are retained for further comparisons related to MSE, namely the p_val and BootYT criteria.

5.2.2 PLS-PR: MSE analysis

First, training log(MSE) were computed using the $n < p$ framework, and their means over all datasets related to each value of σ_5 are shown in Fig. 12.

The global decrease in log(MSE) linked to p_val confirms that, as expected by the increasing number of extracted components observed in Sect. 5.2.1, this criterion models the random noise in $\mathbf{y}$. In contrast, the bootstrap-based criterion shows a systematic increase in log(MSE), which empirically suggests that it better succeeds in separating the real information from the noise.

Variances of PNMSE results over datasets related to each pair (σ_4, σ_5) were computed. Means of these variances, related to fixed values of σ_5, are displayed in Fig. 13.

While results obtained by the bootstrap-based criterion are linked to acceptable variances when $\sigma_5 \leqslant 1.61$, the out-of-range variances linked to the p_val results, due to the

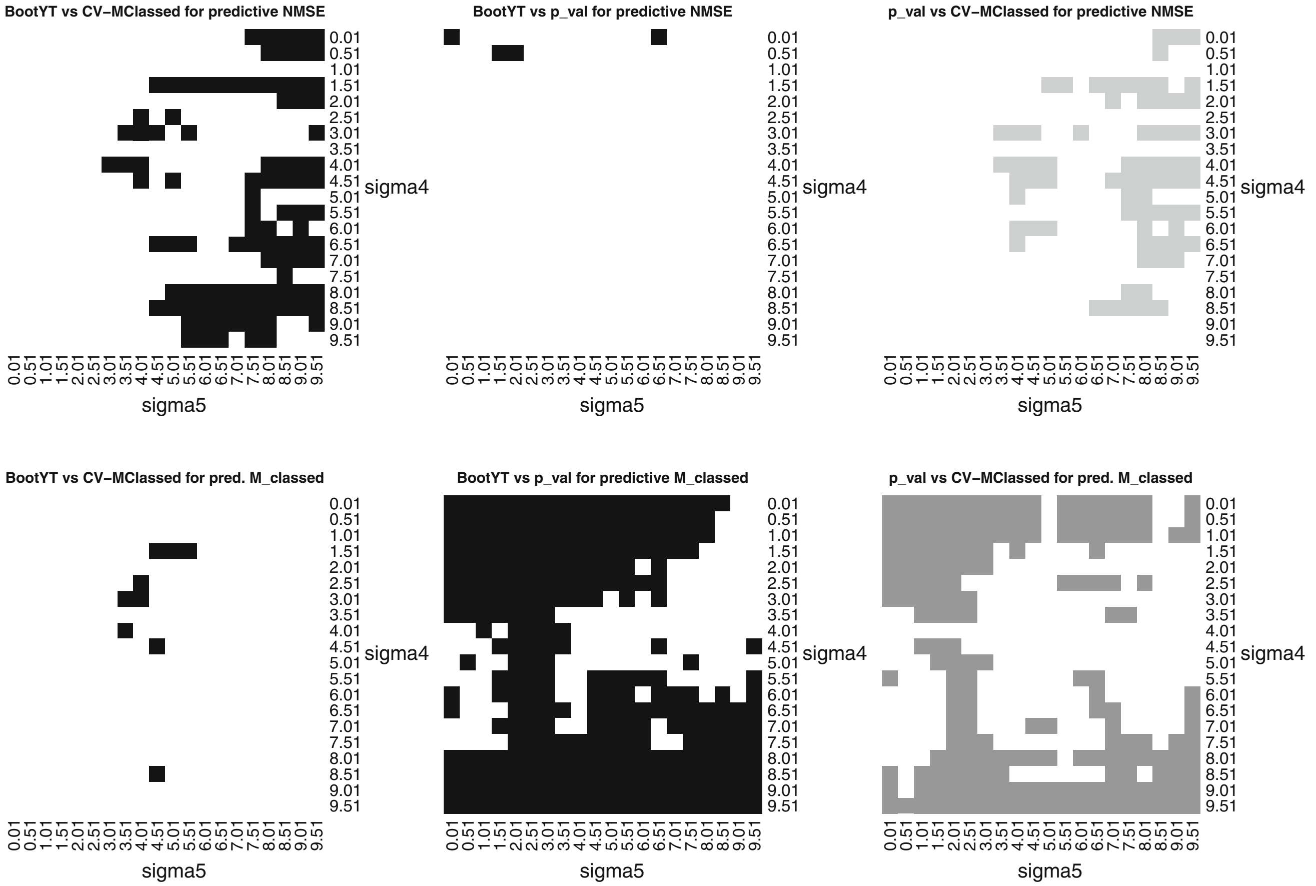


Fig. 10 PLS-LR, $n < p$, sets of NSD values from (E); graphical representation of t-test results for both PNMSE and misclassified values averages comparison; color code: BootYT better (*black*), CV-MClassed better (*dark grey*), p_val better (*light grey*), no significant difference (*white*)

Table 2 Recommended criteria for PLS-LR

Aim	Optimal number of components	Stability	Predictive abilities
	CV-MClassed	BootYT / p_val	BootYT

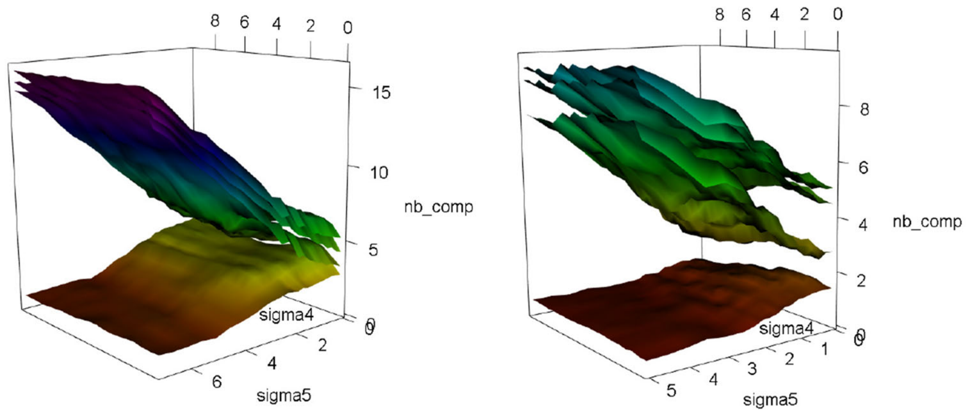


Fig. 11 PLS-PR, evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); *left* $n > p$, sets of NSD values from (F), *right* $n < p$, sets of NSD values from (G); from *top* to *bottom* AIC, BIC, p_val and BootYT results

models' over-complexity observed in Sect. 5.2.1, lead to non-significant differences in mean while using t-tests on these datasets.

To obtain consistent t-test outcomes, models obtained in the $n > p$ framework were used. One hundred additional samples were simulated for each dataset to build test sets. Both means and means of variances of PNMSE over datasets for fixed values of σ_5 were computed, and are shown in Fig. 14.

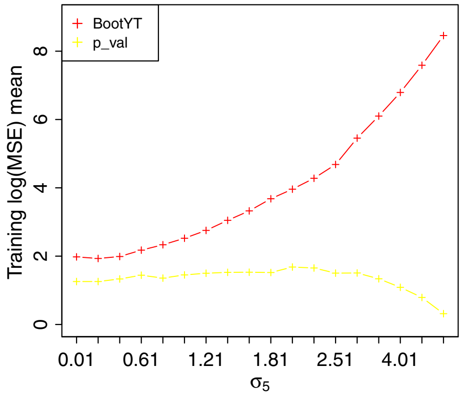


Fig. 12 PLS-PR, $n < p$, sets of NSD values from (G); evolution of training log(MSE) means

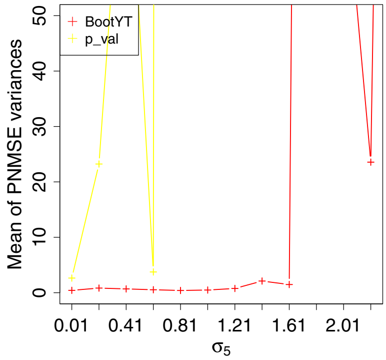


Fig. 13 PLS-PR, $n < p$, sets of NSD values from (G); evolution of means of PNMSE variances for each σ_5

Based on these plots, it is clear that models built with the bootstrap-based criterion are on average better than the trivial ones when $\sigma_5 \leqslant 2.51$, while the p_val criterion fails to build better models than the trivial ones when the NSD in $\mathbf{y}$ is higher than 1.81. Both criteria return low variances in PNMSE for $\sigma_5 \leqslant 3.01$, so t-tests return consistent outcomes in this range of values. Results of these t-tests are displayed in Fig. 15.

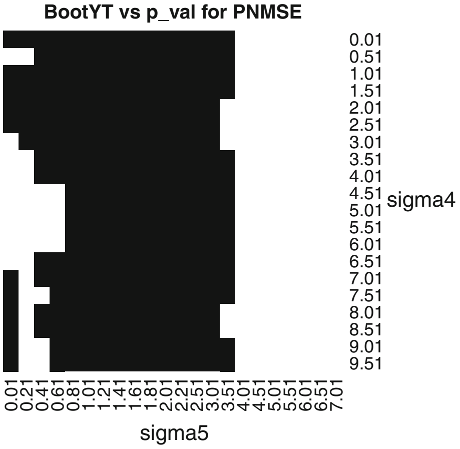


Fig. 15 PLS-PR, $n > p$, sets of NSD values from (F); plot of t-test results for PNMSE means comparison; color code: BootYT better (*black*), no significant difference (*white*)

Based on these t-tests results, our new criterion is to be recommended when setting up a predictive model. Note that non-significant differences for $\sigma_5 \geqslant 3.51$ are due to the high increase in variances linked to the p_val results (see Fig. 14).

5.2.3 PLS-PR: conclusion

In the case of response vector $\mathbf{y}$ linked to a Poisson distribution, the bootstrap-based criterion stands out as the only one which should be used. Indeed, the others can be interpreted as increasing functions of σ_5, so they model the random noise in $\mathbf{y}$, leading to over-fitting issues. As a direct consequence, they return models with poor predictive abilities compared to the new criterion.

6 Applications on real datasets

6.1 Illustration of CV issues: first applications on real datasets

As mentioned by Boulesteix (2014), important issues concerning the stability of the q-fold CV procedure for the choice

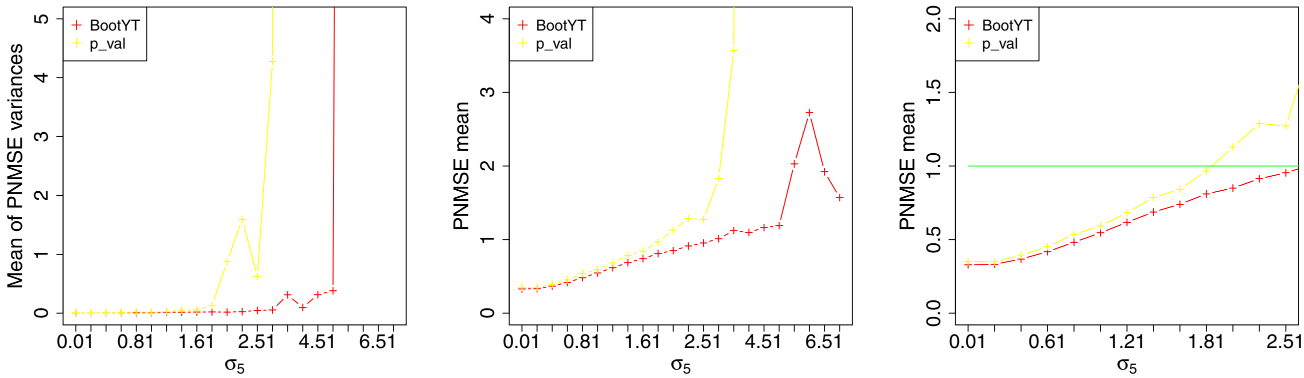


Fig. 14 PLS-PR, $n > p$, sets of NSD values from (F); *left* evolution of means of PNMSE variances for each σ_5; *center* evolution of PNMSE means for each σ_5, *right* evolution of PNMSE means for $\sigma_5 \leqslant 2.51$

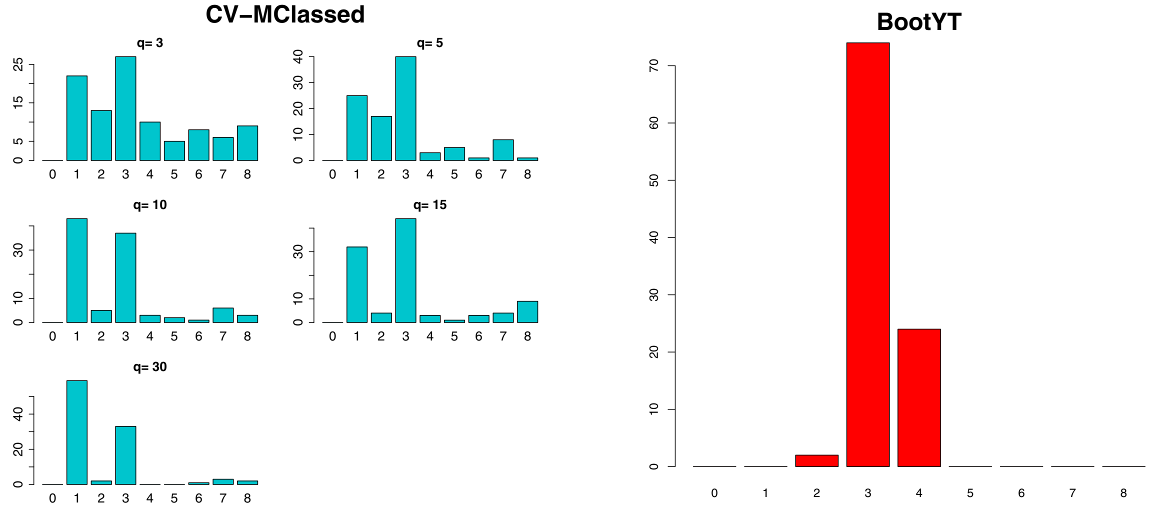


Fig. 16 Extracted number of components using q-fold CV-MClassed (*left*) and BootYT (*right*) criteria

of tuning parameters, here the number of components, have been observed. These issues are directly induced by the value of q and by the random character of this resampling-based procedure while splitting the original dataset into two distinct sets, a training one and a test one. To illustrate consequences on the tuning parameter, we treated two real datasets.

The first dataset was collected from patients with colon adenocarcinoma. It has 104 observations on 33 binary qualitative explanatory variables, and one binary response variable representing the cancer stage according to the Astler-Coller (AB vs. CD) classification (Astler and Coller 1954). This binary response leads us to perform PLS-logistic regressions. This dataset, named *aze_compl*, is available in the *R* package *plsRglm* (Bertrand et al. 2014).

We ran 100 times the selection process for the number of components using the CV-MClassed criterion, with $q \in \{3, 5, 10, 15, 30\}$. Then, we performed the same process using our new criterion. Results are shown in Fig. 16.

Results obtained through q-fold CV, with $q \neq n$, are displayed in Fig. 16, and typical examples for these types of issue. Depending on the choice of q and the way the different folds are split, the extracted number of components can be dramatically different. In addition, obtaining a complete distribution of the number of components is essentially impossible, due to the high number of different possibilities for splitting the original datasets into q groups.

Proposition 2 *Let $n = pq + r$, $0 \leqslant r \leqslant q - 1$ be the Euclidean division of n by q. Then, the number of distinct partitions of the original dataset into r $(p + 1)$-elements subsets and $(q - r)$ p-elements subsets for a CV does not depend on the order of their placement, and is equal to:*

$$f(n, q) = \frac{n!}{r!\,(q - r)!} \times \left(\frac{1}{(p + 1)!}\right)^{r} \times \left(\frac{1}{p!}\right)^{q-r}. \quad (11)$$

Leave-one-out CV, which is the only complete CV ($f(n, n) = 1$, i.e., there is only one way to choose n folds out of n observations), selects one component. However, it suffers from variance issues concerning the bias-variance tradeoff on the estimation of the prediction error (Hastie et al. 2009; Kohavi 1995). Our new criterion is more stable on this dataset and leads the user to choose the number of components, in this case three, via a more accurate process.

The second example is a benchmark dataset, called "Processionnaire du Pin", which is treated in depth by Tenenhaus (1998). It has 33 observations each with 10 explanatory variables, and is also available in the *R* package *plsRglm* under the name *pine*. More details on this dataset are available in Tenenhaus (1998).

The same process was applied to this second example, with the usual PLS regressions. Thus, we can compare the Q^2 criterion obtained through q-fold CV, $q \in \{2, 3, 5, 10, 15\}$, and our new criterion. The Q^2 criterion obtained through leave-one-out CV chooses one significant component. All results are shown in Fig. 17.

Here, q-fold CV does not suffer from stability issues as seen before, since the Q^2 criterion is much more stable than the minimization of the number of misclassified values. However, extracting one component is not recommended. Tenenhaus (1998), after a complete analysis of this dataset, showed that four components is the best decision. This under-estimating issue linked to the Q^2 criterion confirms the simulation results obtained in Sect. 4.2. Thus, while the Q^2 criterion under-estimates this optimal number of

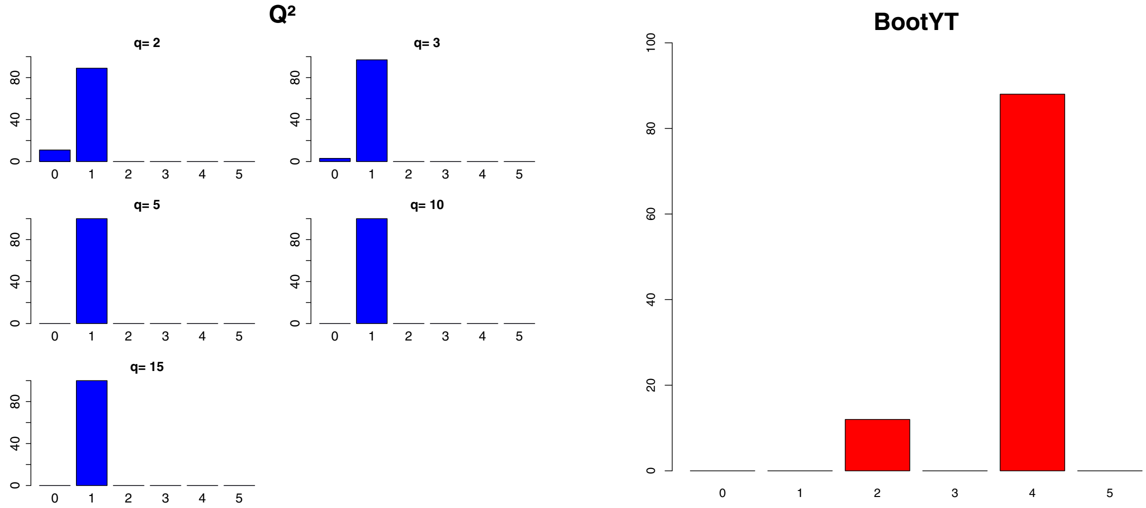
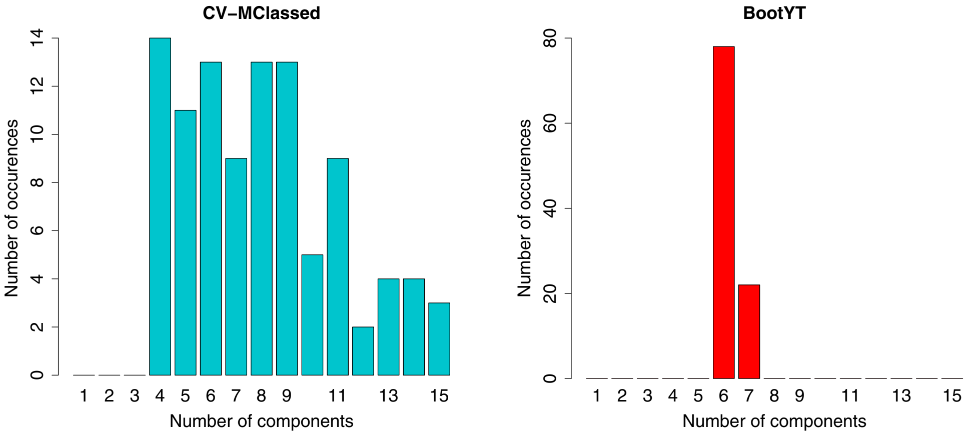


Fig. 17 Extracted numbers of components using q-folds Q^2 (*left*) and BootYT (*right*)

Fig. 18 Extracted number of components using $q = 5$ CV-MClassed (*left*) and BootYT (*right*)



components, our new bootstrap-based criterion selects four components more than 80 % of the time.

6.2 Application on an allelotyping dataset

In this section, we focus on an allelotyping study. Our method is applied to a dataset that concerns 267 subjects with colon cancer. Measures were made on 33 microsatellites, in search of an allelic imbalance that indicates an abnormal number of allele copies of a nearby gene of interest. The aim of the study was to find the microsatellite subsets that would best discriminate left and right colon tumors. Thus, the univariate response corresponds to the original location of a colon tumor, leading to a binary response **y**, taking the value 0 (resp. 1) if it was on the right colon (resp. left). This dataset is available in Supplementary Material 6 and more details are available in Weber et al. (2007).

This dataset contains missing values, so a preprocessing step was performed in order to complete it, using the *R* package *mice*. As **y** is a 0–1 response, we used the three following stopping criteria in component construction: our new bootstrap-based criterion, CV-MClassed, and p_val. We performed 100 times the selection of the number of components using both the $q = 5$ CV-MClassed criterion and our new one, leading to the distribution of the extracted number of components shown in Fig. 18. Then, we computed the mean of the 100 values of extracted numbers of components related to the $q = 5$ CV-MClassed criterion, obtaining 7.99, which is higher than that obtained for BootYT. These results match the simulation conclusions (Sect. 5.1.1).

Based on the distributions in Fig. 18, the major default of the CV-MClassed criterion is clear, namely the dependence of the extracted number of components on the way the group has been randomly formed. Thus, performing a single CV to find the number of components using this criterion, must be avoided. As expected, the BootYT criterion returns stable results and selects, in almost 80 % of cases, 6 components.

We also tested the robustness of these three criteria through a bootstrap re-sampling process with 100 bootstrap iterations, as well as a jackknife method. These two resam-

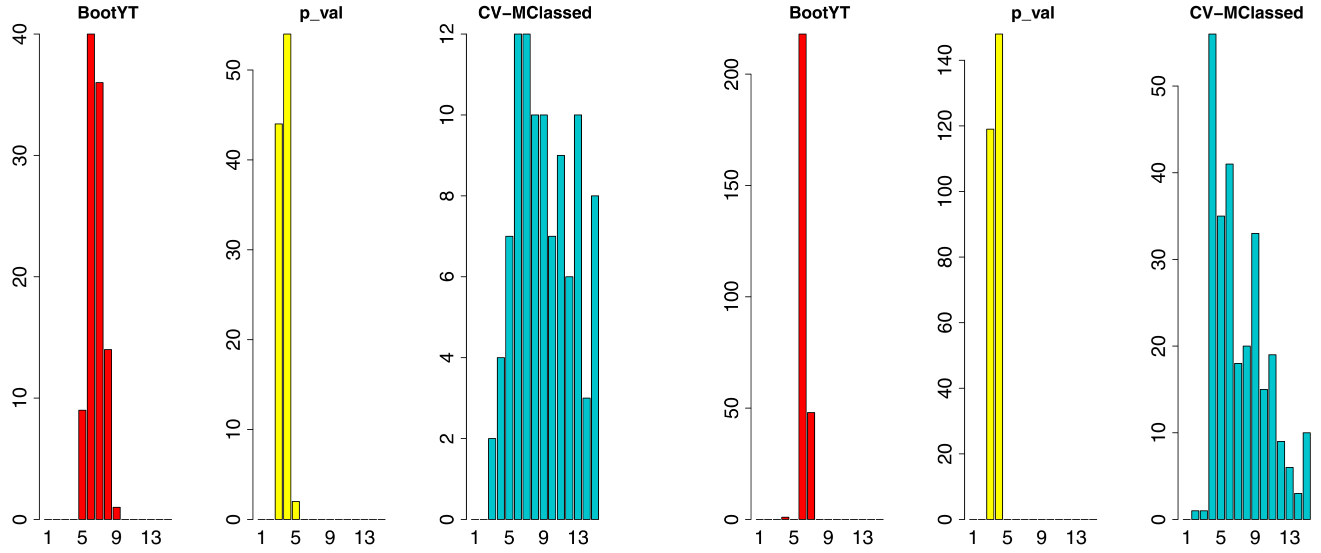


Fig. 19 Distribution of extracted number of components through bootstrap (*left*) and jackknife (*right*) re-sampling

pling methods lead to distributions of the extracted number of components linked to each of the three criteria. Results are shown in Fig. 19.

These results confirm the high resampling robustness of our new criterion compared to CV-MClassed. The p_val criterion has comparable robustness but, based on our simulation results, higher bias. Based on these results and our simulations, we can reasonably conclude that for this dataset, the optimal number of components is 6.

7 Discussion

A new bootstrap-based stopping criterion for PLS component construction was developed, characterized by a high level of stability and robustness with respect to noise level compared to other well-known criteria.

Its implementation requires a function integrating the bootstrap process while building the PLS components. Though not yet available online, an equivalent form can be coded using existing R functions, notably from the *plsRglm* and *boot* packages. It requires recomputing, for each increment of k, the entire set of PLS components. While this style of implementation performs more operations than needed, it does not take much more time, since the main part of the computational cost comes from the bootstrap process.

Indeed, our bootstrap-based criterion has a shortcoming with respect to computational time, which is greater than the other methods since it requires, in the PLSR framework, $[(k_{\max}+1)\times p_l + \ (K+1)]\times R$ least squares regressions per dataset. An initial improvement has already been made by developing parallel processing for this. We note also that the development of a corrected DoF in the PLSGLR framework would also allow the development of an adapted corrected BIC formulation. This could provide an interesting alternative to the bootstrap-based criterion since it might save a great amount of computational time.

Nevertheless, our new bootstrap-based criterion represents a reliable, consistent and robust stopping criterion for selecting the optimal number of PLS components. It avoids the use of CV techniques and, to the best of our knowledge, is the first to directly focus on the different loadings involved. Thus, it can be performed both in the PLSR and PLSGLR frameworks, and allows users to test the significance of a new component with a preset risk level α.

In the $n > p$ PLSR framework, our simulations confirm that both BICdof and BootYT are appropriate and well-designed criteria. Our new bootstrap-based criterion is also an appropriate alternative in the $n < p$ case, since the BICdof criterion suffers from high variance and overestimation issues, especially for models with low random noise levels in $\mathbf{y}$. Furthermore, both the BICdof and Q2K5 criteria are more sensitive than the bootstrap-based criterion to increasing noise levels in $\mathbf{y}$ in this case.

As for the PLSGLR framework, our simulation results, based on two specific distributions (binomial and Poisson), lead us to recommend this new bootstrap-based criterion. Indeed, in the PLS-LR case, we show that, depending on the statistic used (testing NMSE or number of misclassified predictions) to test predictive ability, the bootstrap-based criterion is never significantly worse than either the CV-MClassed or p_val criteria. Concerning results obtained for a response vector following a Poisson distribution, the bootstrap-based criterion is the only one which returns consistent results, by retaining a decreasing number of components while the random noise level in $\mathbf{y}$ increases. Adding this to the MSE analysis and the obtained t-test results, it is reasonable to advise using the new criterion in this framework.

Acknowledgements The authors gratefully acknowledge the LabEx IRMIA for J. Magnanensi's PhD grant.

References

Akaike, H.: A new look at the statistical model identification. Autom. Control IEEE Trans. **19**(6), 716–723 (1974)

Allen, D.M.: The prediction sum of squares as a criterion for selecting predictor variables. University of Kentucky, (1971). Technical report 23 - Department of Statistics

Amato, S., Vinzi, V.E.: Bootstrap-based q kh 2 for the selection of components and variables in pls regression. Chemom. Intell. Lab. Syst **68**(1), 5–16 (2003)

Astler, V.B., Coller, F.A.: The prognostic significance of direct extension of carcinoma of the colon and rectum. Annal. Surg **139**(6), 846 (1954)

Bastien, P., Vinzi, V.E., Tenenhaus, M.: PLS generalised linear regression. Comput. Stat. Data Anal. **48**(1), 17–46 (2005)

Bertrand, F., Magnanensi, J., Maumy-Bertrand, M., Meyer, N.: *Partial least squares regression for generalized linear models*. http://www-irma.u-strasbg.fr/fbertran/, User2014!, Los Angeles, p 150

Boulesteix, A.L.: PLS dimension reduction for classification with microarray data. Stat. Appl. Genet. Mol. Biol. **3**(1), 1–30 (2004)

Boulesteix, A.L.: Accuracy estimation for PLS and related methods via resampling-based procedures. In: *PLS'14 Book of Abstracts*, pp. 13–14, (2014)

Boulesteix, A.L., Strimmer, K.: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinform. **8**(1), 32–44 (2007)

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC press, Boca Raton (1984)

Denham, M.C.: Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. J. Chemom. **14**(4), 351–361 (2000)

Ding, B., Gentleman, R.: Classification using generalized partial least squares. J. Comput. Graph. Stat. **14**(2), 280–298 (2005)

Efron, B.: Bootstrap methods: another look at the jackknife. Annal. Stat. **7**(1), 1–26 (1979)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap, vol. 57. Chapman & Hall/CRC, Boca Raton (1993)

Freedman, D.A.: Bootstrapping regression models. Annal. Stat. **9**(6), 1218–1228 (1981)

Gómez-Carracedo, M., Andrade, J., Rutledge, D., Faber, N.: Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. Anal. Chim. Acta **585**(2), 253–265 (2007)

Gourvénec, S., Pierna, J.F., Massart, D., Rutledge, D.: An evaluation of the PoLiSh smoothed regression and the Monte Carlo cross-validation for the determination of the complexity of a PLS model. Chemom. Intell. Lab. Syst. **68**(1), 41–51 (2003)

Green, R.L., Kalivas, J.H.: Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off. Chemom. Intell. Lab. Syst. **60**(1), 173–188 (2002)

Haaland, D.M., Thomas, E.V.: Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. **60**(11), 1193–1202 (1988)

Hastie, T., Tibshirani, R., Friedman, J.J.H.: The Elements of Statistical Learning, vol. 1, 2nd edn. Springer, New York (2009)

Höskuldsson, A.: PLS regression methods. J. Chemom. **2**(3), 211–228 (1988)

Höskuldsson, A.: Dimension of linear models. Chemom. Intell. Lab. Syst. **32**(1), 37–55 (1996)

Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence, vol. 2, pp. 1137–1143. Morgan Kaufmann Publishers Inc., (1995)

Krämer, N., Sugiyama, M.: The degrees of freedom of partial least squares regression. J. Am. Stat. Assoc. **106**(494), 697–705 (2011)

Krzanowski, W.: Cross-validation in principal component analysis. Biometrics **43**(3), 575–584 (1987)

Li, B., Morris, J., Martin, E.B.: Model selection for partial least squares regression. Chemom. Intell. Lab. Syst. **64**(1), 79–89 (2002)

Manne, R.: Analysis of two partial-least-squares algorithms for multivariate calibration. Chemom. Intell. Lab. Syst. **2**(1), 187–197 (1987)

Marx, B.D.: Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics **38**(4), 374–381 (1996)

Mevik, B.H., Cederkvist, H.R.: Mean squared error of prediction (msep) estimates for principal component regression (pcr) and partial least squares regression (plsr). J. Chemom. **18**(9), 422–429 (2004)

Meyer, N., Maumy-Bertrand, M., Bertrand, F.: Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives: application aux données d'allélotypage. J. de la Société Française de Stat. **151**(2), 1–18 (2010)

Nguyen, D.V., Rocke, D.M.: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics **18**(1), 39–50 (2002)

Osten, D.W.: Selection of optimal regression models via cross-validation. J. Chemom. **2**(1), 39–48 (1988)

Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: Saunders, C.J., Gunn, S.R., Grobelnik, M., Shawe-Taylor, J. (eds.) Subspace, Latent Structure and Feature Selection, pp. 34–51. Springer, Berlin (2006)

Schwarz, G.: Estimating the dimension of a model. Annal. Stat. **6**(2), 461–464 (1978)

Tenenhaus, M.: La régression PLS, Théorie et pratique. Editions Technip, Paris (1998)

Umetrics, A.: User's guide to simca-p, simca-p+. version 10. 0. *Umeå, Sweden (2005). Umetrics AB* (2005)

Van der Voet, H.: Comparing the predictive accuracy of models using a simple randomization test. Chemom. Intell. Lab. Syst. **25**(2), 313–323 (1994)

Wakeling, I.N., Morris, J.J.: A test of significance for partial least squares regression. J. Chemom. **7**(4), 291–304 (1993)

Weber, J.C., Meyer, N., Pencreach, E., Schneider, A., Guérin, E., Neuville, A., Stemmer, C., Brigand, C., Bachellier, P., Rohr, S., et al.: Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. Int. J. Cancer **120**(3), 524–532 (2007)

Wehrens, R., Linden, Wvd: Bootstrapping principal component regression models. J. Chemom. **11**(2), 157–171 (1997)

Welch, B.L.: The generalization of student's problem when several different population variances are involved. Biometrika **34**(1–2), 28–35 (1947)

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., Faber, K.: A randomization test for PLS component selection. J. Chemom. **21**(10–11), 427–439 (2007)

Wold, S., Martens, H., Wold, H.: The multivariate calibration problem in chemistry solved by the PLS method. Matrix Pencils, pp. 286–293. Springer, Berlin (1983)

Wold, S., Ruhe, A., Wold, H., Dunn III, W.: The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. **5**(3), 735–743 (1984)

Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**(2), 109–130 (2001)