Computational Statistics (2007) 22:307–321
DOI 10.1007/s00180-007-0034-3

ORIGINAL PAPER

An empirical study of PLAD regression using the bootstrap

Athanassios Kondylis · Joe Whittaker

Published online: 24 February 2007

Abstract Partial LAD regression uses the L_1 norm associated with least absolute deviations (LAD) regression while retaining the same algorithmic structure of univariate partial least squares (PLS) regression. We use the bootstrap in order to assess the partial LAD regression model performance and to make comparisons to PLS regression. We use a variety of examples coming from NIR experiments as well as two sets of experimental data.

Keywords Resampling · Partial least squares · LAD regression · NIR experiments

1 Introduction

Resampling methods are used in several related but distinct areas of statistics, for instance: randomization tests, cross-validation, the jackknife, and the bootstrap. These techniques are used to assess the performance of a given statistic under study, as well as in model selection and validation. For instance, see (Diaconis and Efron 1983; Efron and Gong 1983; Shao 1993; Edgington 1995).

Partial least absolute deviations (PLAD) regression (Dodge et al. 2004), is a variation of partial least squares (PLS) regression that estimates the median response rather than the mean response. However, it seems impossible to extract theoretically the sampling properties of PLS and consequently of PLAD.

A. Kondylis (✉)
Statistical Institute, University of Neuchâtel, Neuchâtel, Switzerland
e-mail: atanassios.kondylis@unine.ch

J. Whittaker
Department of Mathematics and Statistics, Lancaster University, Lancaster, England
e-mail: Joe.whittaker@lancaster.ac.uk

We use the bootstrap here in order to evaluate the repeated sampling properties of PLAD and to compare these with PLS.

The variation of PLS regression that leads to PLAD is described in Sect. 3. Its relation to principal components and PLS regression methods are given therein. In Sect. 4 a brief description of the bootstrap is given, while Sects. 4.1, 4.2, and 4.3 give the necessary details on the use of the bootstrap in PLAD regression. Section 5 uses near infra-red (NIR) data sets, commonly used in PLS regression applications, in order to assess the PLAD regression and to make comparisons with PLS regression. Experimental data analysis follows in Sect. 6 and finally, in Sect. 7 some conclusions are drawn.

2 Notation and preliminaries

The following notation is used throughout this article: bold faced lower case symbols are vectors, upper case are matrices. The superscript T is used to denote the transpose of a matrix. We use the subscript n to denote that the expectations are taken on sample quantities, so that $\mathsf{E}_n(\cdot)$ is a sample mean, and so $\mathsf{var}_n(\cdot)$, $\mathsf{sd}_n(\cdot)$, and $\mathsf{cov}_n(\cdot|\cdot)$ denote the sample variance, standard deviation, and covariance, respectively. We also use med_n to denote the sample median, while the median absolute deviation is denoted as mad_n. Later, we introduce the use of the subscript R in the above notation to indicate estimates obtained by R bootstrap replicates; the latter are flagged by the use of the superscript *.

The subscript k indicates the number of dimensions extracted in PLS and PLAD regression, and k_{max} is the largest model dimension envisaged. Since both PLS and PLAD regression use components as regressors instead of original predictors it is sensible to use the notation $\boldsymbol{t}_k$ for the components while $\boldsymbol{x}_j$ denotes the jth original predictor. The loadings and weights of a component $\boldsymbol{t}_k$ are denoted by $\boldsymbol{p}_k$ and $\boldsymbol{w}_k$, respectively. The vectors $\boldsymbol{t}_k, \boldsymbol{p}_k$, and $\boldsymbol{w}_k$ are collected together in the matrices $\boldsymbol{T}_k, \boldsymbol{P}_k$, and $\boldsymbol{W}_k$, of appropriate dimension. We consider univariate regression problems with the original predictors being the columns of the $n \times p$ matrix $\boldsymbol{X}$, and response vector $\boldsymbol{y}$. We use $i = 1, \ldots, n$ to denote observations and and $j = 1, \ldots, p$ for predictors. The fitted value for the ith observation in a regression model based on k components is denoted by $\widehat{y}_{ik}$ and the fitted response vector by $\widehat{\boldsymbol{y}}_k$. Further notation is introduced as needed.

3 Partial LAD regression

Partial LAD regression, introduced by Dodge et al. (2004), uses the L_1 norm associated with least absolute deviations regression. It takes the structure of the PLS algorithm for univariate partial least squares regression (Martens and Naes 1989; Tenenhaus 1998), and similarly extracts components $\boldsymbol{t}$, in directions that depend upon the response variable. In PLAD these directions are determined by a Gnanadesikan–Ketterning (GK) covariance estimate (Gnanadesikan and Kettenring 1972), that replaces the usual variance based on the L_2 norm with

mad, the median absolute deviation based on L_1;

$$\mathsf{mad}_n(\boldsymbol{x}) = \mathsf{med}_n\,|\boldsymbol{x} - \mathsf{med}_n(\boldsymbol{x})|.$$

We use the superscript *mad* to emphasize that covariance is calculated from mad instead of from the more common variance.

Algorithm 1 Partial LAD regression

(1) Center (standardise) the data $\mathcal{D} = (\boldsymbol{X}_0, \boldsymbol{y})$

(2) For $k = 1, \ldots, k_{max}$

(2a) compute $\boldsymbol{w}_k^{mad}$ from:

$$w_{jk}^{mad} = \frac{1}{4}\left(\mathsf{mad}_n^2(\boldsymbol{x}_{jk-1} + \boldsymbol{y}) - \mathsf{mad}_n^2(\boldsymbol{x}_{jk-1} - \boldsymbol{y})\right); \tag{1}$$

(2b) scale $\boldsymbol{w}_k^{mad}$ to 1;

(2c) build the component

$$\boldsymbol{t}_k = \boldsymbol{X}_{k-1}\boldsymbol{w}_k^{mad}; \tag{2}$$

(2d) orthogonalise each $\boldsymbol{x}_{jk-1}$, $j = 1, \ldots, p$ with respect to $\boldsymbol{t}_k$, to give $\boldsymbol{X}_k$.

(3) Compute the LAD regression to give the fitted vectors $\widehat{\boldsymbol{y}}_k = \boldsymbol{T}_k\,\widehat{\boldsymbol{q}}_k{}^{lad}$, where $\boldsymbol{T}_k = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k)$ is the score matrix, and $\widehat{\boldsymbol{q}}_k^{lad} = (\widehat{q}_1^{lad}, \ldots, \widehat{q}_k^{lad})^T$ is the LAD regression coefficient.

(4) Recover the implied partial LAD regression coefficients from $\widehat{\boldsymbol{\beta}}_k = \widetilde{\boldsymbol{W}}_k\,\widehat{\boldsymbol{q}}_k^{lad}$, where the matrix $\widetilde{\boldsymbol{W}}_k$ includes as columns the $\boldsymbol{w}_k^{mad}$ expressed in terms of the original $\boldsymbol{x}_j$.

With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, denoting observation units and predictors, the PLAD regression algorithm is given in Algorithm 1. In contrast to principal components regression, which extracts the same components whatever the response, PLS and PLAD regression share similar properties, we note

(1) $\quad \boldsymbol{y} = q_1\,\boldsymbol{t}_1 + \cdots + q_k\,\boldsymbol{t}_k + \boldsymbol{\epsilon},$

(2) $\quad \boldsymbol{X} = \boldsymbol{p}_1\,\boldsymbol{t}_1 + \cdots + \boldsymbol{p}_k\,\boldsymbol{t}_k + \boldsymbol{f},$

(3) $\quad \mathsf{cov}_n(\boldsymbol{t}_i, \boldsymbol{t}_j) = 0 \quad \text{for} \quad i \neq j,$

where $\boldsymbol{e}$ and $\boldsymbol{f}$ correspond to residual terms, and $\boldsymbol{p}_k$ are the $\boldsymbol{X}$-loadings. PLAD builds a regression model at each iteration k that relates the predictors to the response according to

$$\widehat{\boldsymbol{y}}_k = \sum_{j=1}^{p} \widehat{\beta}_{jk}\,\boldsymbol{x}_j. \tag{3}$$

The implied regression coefficients $\widehat{\beta}_{jk}$ are determined by the derived components retained in the final regression model. In fact,

$$\widehat{\boldsymbol{\beta}}_k^{plad} = \widetilde{\boldsymbol{W}}_k \widehat{\boldsymbol{q}}_k^{lad}, \tag{4}$$

with $\widetilde{\boldsymbol{W}}_k = \boldsymbol{W}_k(\boldsymbol{P}_k^T \boldsymbol{W}_k)^{-1}$ being the X-weights expressed in terms of the original predictors. Expression (4) demonstrates the similarity of PLAD to PLS regression, and displays its main difference. PLAD regression deflates the X-data as in PLS regression. It retains therefore orthogonal components, while it recovers the weight vectors $\boldsymbol{w}_k$ contained in the matrix $\boldsymbol{W}_k$ in terms of the original predictors (instead of in terms of deflated data) according to $\widetilde{\boldsymbol{W}}_k = \boldsymbol{W}_k(\boldsymbol{P}_k^T \boldsymbol{W}_k)^{-1}$. The PLAD coefficients $\boldsymbol{q}_k^{lad}$ are obtained using LAD regression of the scores $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k$ on the response $\boldsymbol{y}$. While for PLS regression at dimension k the coefficients $q_1, \ldots, q_{k-1}$ remain the same, for PLAD regression this is not the case. In the same sense we use the suffix *plad* in $\boldsymbol{w}_k^{plad}$ to emphasize that the direction vectors depend upon the GK-type covariance used in expression (1).

We choose to assess the overall prediction error incurred by the partial LAD regression model in (3) for the data at hand $\mathcal{D}$ by the standard root mean squared error (RMSE) loss function given by

$$\mathcal{L}(\mathcal{D}) = \sqrt{\mathsf{E}_n[loss(\boldsymbol{y}, \widehat{\boldsymbol{y}}_k)]} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{ik})^2}. \tag{5}$$

Recall the notation $\mathsf{E}_n[\cdot]$ indicates averaging over the n observations in $\mathcal{D}$.

The essential motivation for PLAD regression is to model the median of the response, instead of the mean as with PLS, and so it employs the L_1 instead of the L_2 norm. Additionally, as LAD regression is less sensitive to outliers than least squares regression PLAD may profit in the same way. However, PLAD is not a robust alternative to PLS for two main reasons: firstly, because the partial LAD algorithm does not bound the influence arising from high leverages in the predictor space. This is verified by the scores expression in (2). Secondly, because of the GK type of covariance, with the mad replacing the variance, has unstable robust properties depending on the scales of $\boldsymbol{X}$ and $\boldsymbol{y}$ (Huber 1981). A simple alternative is to set

$$w_{jk} = \frac{1}{4\,\alpha_j \beta}(\,\mathsf{mad}_n^2(\alpha_j \boldsymbol{x}_{jk-1} + \beta \boldsymbol{y}) - \,\mathsf{mad}_n^2(\alpha_j \boldsymbol{x}_{jk-1} - \beta \boldsymbol{y})), \tag{6}$$

with $\alpha_j = 1/\,\mathsf{mad}_n(\mathbf{x}_j)$ and $\beta = 1/\,\mathsf{mad}_n(\boldsymbol{y})$. This is the definition we use in this paper.

The rationale for replacing (1) by (6) is that the latter is less sensitive to the choice of scale for the predictor variables. Recall that PLS regression is not scale invariant.

Our goal is to further examine the properties and, in particular, the stability of PLAD regression under resampling. This is in line with most PLS procedures where prediction performance and model selection are determined by means of cross validation or some other resampling method. In Sect. 4, we consider how

to use the bootstrap to evaluate sampling variability. In particular, we choose to bootstrap data, rather than residuals, for building empirical distributions of regression coefficients and of prediction errors.

4 PLS, PLAD and the bootstrap

The bootstrap (Efron and Tibshirani 1993) is a resampling method which provides an assessment of uncertainty when theoretical solutions are not available, as is the case with PLS and PLAD regression. The general bootstrap algorithm is sketched in Algorithm 2.

Algorithm 2 Bootstrap

Let $(y_1, \ldots, y_n)$ be an observed sample of size n, and θ the statistic of interest. For $b = 1, \ldots, R$ the number of repetitions:

(1) Generate a random sample $(y_1^b, \ldots, y_n^b) \sim F$, with replacement, from the distribution F which is either given or estimated by the empirical distribution $\widehat{F}$.

(2) Compute $\widehat{\theta}_b$ from the simulated data $(y_1^b, \ldots, y_n^b)$.

Use $(\widehat{\theta}_1, \widehat{\theta}_2, \ldots \widehat{\theta}_R)$ to estimate the sampling distribution of $\widehat{\theta}$ and summaries of interest such as $\mathsf{E}_R(\widehat{\theta})$, $\mathsf{var}_R(\widehat{\theta})$, and $\mathsf{sd}_R(\widehat{\theta})$.

The subscript R in $\mathsf{E}_R(\cdot)$, $\mathsf{var}_R(\cdot)$, and $\mathsf{sd}_R(\cdot)$ indicates averaging over the R bootstrap replicates.

4.1 Bootstrapping data or residuals?

In principle may base our analysis on bootstrapping the raw data $\mathcal{D}$ or on bootstrapping the residuals having fitted a regression. For several reasons we employ the former here. Bootstrapping residuals requires fixing the number of components k in order to define the fitted values and so determine the residuals. This may conflict with our interest in comparison of how different regression methods might choose k. Furthermore our interest is to compare the manner in which PLS and PLAD regression generalise performance to new observations rather than to new samples of residuals. More generally, bootstrapping data is less sensitive to assumptions made on the regression model [see (Efron and Tibshirani 1993), p 113].

We generate R bootstrap samples $\mathcal{D}^{\star\, b}$, $b = 1, 2, \ldots, R$, each consisting of $(\boldsymbol{X}^{\star\, b}, \boldsymbol{y}^{\star\, b})$ obtained by resampling the rows of $(\boldsymbol{X}, \boldsymbol{y})$. We compare the bootstrap results from PLS and PLAD regression. When necessary we use a suffix *pls* or *plad* to denote quantities computed from the corresponding methods.

Let $s = s(\mathcal{D})$ be the statistic of interest calculated from the original data, and assumed to be invariant to permutation of the rows of the data $\mathcal{D}$. Its mean and

standard deviation calculated over the bootstrap samples are given as

$$\mathsf{E}_R(s^\star) = \frac{1}{R}\sum_{b=1}^{R} s(\mathcal{D}^{\star b}) \text{ and } \mathsf{sd}_R(s^\star) = \sqrt{\frac{1}{R-1}\sum_{b=1}^{R}[s(\mathcal{D}^{\star b}) - \mathsf{E}_R(s^\star)]^2}. \quad (7)$$

Constructing confidence intervals and hypothesis testing based on the statistic s is then straightforward [see (Efron and Tibshirani 1993), chapters 12, 13, 14, 16].

4.2 Bootstrapping the regression coefficients

Consider the implied coefficients $\widehat{\beta}_{jk},\ j = 1, 2, \ldots, p$ using PLAD from (4) and a similar expression for PLS. These statistics are invariant to permuting the rows of $\mathcal{D}$. From the bootstrap we compute $\mathsf{E}_R(\widehat{\beta}^\star_{jk})$ and $\mathsf{sd}_R(\widehat{\beta}^\star_{jk})$ as at (7) above, that is

$$\mathsf{E}_R(\widehat{\beta}^\star_{jk}) = \frac{1}{R}\sum_{b=1}^{R} \widehat{\beta}^{\star b}_{jk} \text{ and } \mathsf{sd}_R(\widehat{\beta}^\star_{jk}) = \sqrt{\frac{1}{R-1}\sum_{b=1}^{R}[\widehat{\beta}^{\star b}_{jk} - \mathsf{E}_R(\widehat{\beta}^\star_{jk})]^2}. \quad (8)$$

The $(1-\alpha)\%$ percentile bootstrap confidence limits for $\widehat{\beta}_{jk}$ are

$$[\mathrm{q}^\star_{\alpha/2},\ \mathrm{q}^\star_{1-\alpha/2}] \quad (9)$$

corresponding to the $\alpha/2$ and $1-\alpha/2$ empirical quantiles of the distribution of the bootstrap replicates for $\widehat{\beta}^{\star b}_{jk}$. We set α equal to 0.05. Bootstrap confidence intervals may be also obtained by using the ABC or the BCa methods (Efron 1987). We use the percentile approach here because it is simpler to interpret and it is not more computationally expensive.

4.3 Bootstrapping the prediction error

For any value of k the RMSE, or *apparent* prediction error, $\mathcal{L}$ at (5) is invariant to permutation, and its bootstrap standard deviation is

$$\mathsf{sd}_R(\mathcal{L}^\star) = \sqrt{\frac{1}{R-1}\sum_{b=1}^{R}[\mathcal{L}(\mathcal{D}^{\star b}) - \mathsf{E}_R(\mathcal{L}^\star)]^2}, \quad (10)$$

from (7). These quantities are computed for PLS and PLAD, giving $\mathsf{sd}_R(\mathcal{L}^{\star pls})$ and $\mathsf{sd}_R(\mathcal{L}^{\star plad})$.

The mean bootstrapped RMSE

$$\mathsf{E}_R(\mathcal{L}^\star) = \frac{1}{R}\sum_{b=1}^{R} \mathcal{L}(\mathcal{D}^{\star b}), \quad (11)$$

is the *resampling* prediction error. It is commonly used for model selection (Denham 2000) as the magnitude of the errors decrease with the number of the components k retained in the final regression model.

The bootstrap estimate of the apparent prediction error can be improved by subtracting the bias induced in using E_R instead of E_n. That is, by using $\widehat{F}$ instead of F (see Algorithm 2, step 1). This is the *optimism*. The bootstrap estimate of the expected optimism $\widehat{\omega}(\widehat{F})$ is obtained by

$$\widehat{\omega}(\widehat{F}) = \mathsf{E}_n[\mathcal{L}(\mathcal{D}^{\star,b}, F) - \mathcal{L}(\mathcal{D}^{\star,b}, \widehat{F})], \tag{12}$$

with $\mathcal{L}(\mathcal{D}^{\star,b}, \widehat{F})$ equally denoting the resampling error $\mathcal{L}(\mathcal{D}^{\star\,b})$. This is often called type II error. The first term in the right hand side of expression (12) corresponds to the loss induced by testing predictions for models constructed on the bootstrap sample on the original response, and it is often called type I error.

The averaged optimism above is approximated for the R bootstrap samples by

$$\widehat{\omega}(\widehat{F}) = \frac{1}{\sqrt{n}\,R} \left\{ \sum_{b=1}^{R} \sqrt{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k^{\star\,b})^2} - \sum_{b=1}^{R} \sqrt{\sum_{i=1}^{n} (y_i^{\star\,b} - \boldsymbol{x}_i^{\star\,b\,T} \widehat{\boldsymbol{\beta}}_k^{\star\,b})^2} \right\}. \tag{13}$$

The final estimate for the prediction error (FPE) of the partial LAD regression model including k components is given by

$$FPE_k^{plad} = \mathcal{L}(\mathcal{D}, \widehat{F}) + \widehat{\omega}(\widehat{F}), \tag{14}$$

with a similar expression for the PLS regression model.

Further improvement in assessing the prediction error can be obtained by the use of the 0.632 bootstrap estimator (Efron and Tibshirani 1997). In that case the optimism is estimated in the apparent prediction error for observation i by using bootstrap samples that do not include the latter observation. The 0.632 estimate is computationally more expensive. We therefore base our estimation for prediction error on the expression (14).

5 Experience based on NIR data sets

5.1 NIR data sets

The field of near infra-red (NIR) experimentations is a principal application of PLS methods (Martens and Naes 1989). We use here three NIR data sets: the Wheat data (Fearn 1983), the Fish data (Naes 1985), and the Octane data (Tenenhaus 1998).

The predictors in each data set are spectra at different number of wavelengths, and are highly collinear. The reflectance of the NIR radiation by the

sample units at different wavelengths are used to model chemical concentrations. The Wheat and the Fish data have a relatively small number of regressors, while the Octane data count more than 200 regressors for 39 observations. Further details on these data sets are given in the references.

5.2 Results

The analyzed data are initially centered but not rescaled as they are measured on similar physical scales. The number k_{max} is known for each data set due to previous analysis. Nevertheless, in our results we obtain and we present estimates for $k = 1, \ldots, 4$. The number of the bootstrap replicates is set to $R = 500$ for all data sets. Table 1 gives the values of the apparent and the resampling prediction errors for the NIR data sets together with the bootstrap estimates for the optimism.

The prediction errors from PLS and PLAD are close for both the Wheat and the Fish data sets. The PLAD estimates of the resampling prediction error are slightly more variable in comparison to the PLS. For the Octane data, PLAD regression reduces the apparent error on the second component rather more than PLS does (though including the third component brings PLS and PLAD together). However this reduction is accompanied by a relatively large variation in the second PLAD component, which is illustrated in Fig. 1, where the solid line (green) represents the apparent prediction error and the boxplots are constructed according to the prediction error values on the bootstrap samples. Finally, the horizontal dashed lines (red) correspond to the final prediction error estimate given by equation (14).

Table 1 leads to the conclusion that PLS generally reaches slightly lower levels of prediction error than PLAD regression. PLAD regression in certain cases has lower apparent prediction error for the data at hand in comparison to PLS. For the Octane data set, the apparent PLAD error provides some evidence that two components (instead of three for PLS) could have been retained in the final model. This is probably due to six outlying observations in the Octane data (Dodge et al. 2004). Outlying observations and high leverages are also found in the Fish data set, where observations 43, 44, 45 are high leverages while observations 1 and 43 are outliers. For PLAD regression the difference in the prediction loss for $k = 2$ and $k = 3$ is much less than for PLS. This is in line with our findings for the Octane data. However for both data sets the resampling procedure does not indicate that this performance generalizes (note the right hand panel of Fig. 1 for $k = 2$).

Figures 2 and 3 display the percentile bootstrap confidence limits for the implied regression coefficients from the Octane and the Wheat data. The PLAD implied coefficients in the right panel of Fig. 2 are more variable than PLS coefficients given in the left panel.

This is also apparent for the Fish and Wheat data sets as well. We illustrate the latter in Fig. 3 where the estimated regression coefficients for PLAD and PLS

Table 1 NIR data: RMSE: apparent error (AE), resampling error (RE), with standard errors (in parentheses), and the expected optimism $\widehat{\omega}(\widehat{F})$ for the three NIR data sets, and for $k = 1, \ldots, 4$

$k=1$	Data	Method	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$
	Wheat	PLS	1.2326	1.1690 (0.1282)	0.0882
		PLAD	1.2181	1.1919 (0.1528)	0.1101
	Fish	PLS	3.0161	2.8758 (0.3484)	0.2445
		PLAD	3.0220	2.9429 (0.3601)	0.2612
	Octane	PLS	1.7395	1.6041 (0.2691)	0.0901
		PLAD	1.8308	1.8325 (0.2443)	0.1448
$k=2$	Data	Method	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$
	Wheat	PLS	0.9790	0.8050 (0.1979)	0.1114
		PLAD	0.9688	0.9697 (0.1957)	0.1352
	Fish	PLS	1.7914	1.6728 (0.1998)	0.1748
		PLAD	1.7234	1.6881 (0.2530)	0.1570
	Octane	PLS	0.6973	0.7364 (0.1152)	0.0408
		PLAD	0.5878	0.5783 (0.1740)	0.0472
$k=3$	Data	Method	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$
	Wheat	PLS	0.2413	0.2127 (0.0546)	0.0545
		PLAD	0.3352	0.2901 (0.1200)	0.0792
	Fish	PLS	1.2777	1.1170 (0.2094)	0.3396
		PLAD	1.2900	1.2796 (0.2771)	0.3063
	Octane	PLS	0.2574	0.3801 (0.1469)	0.0260
		PLAD	0.3664	0.3878 (0.1515)	0.0434
$k=4$	Data	Method	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$
	Wheat	PLS	0.1968	0.1673 (0.1968)	0.0569
		PLAD	0.2112	0.2049 (0.0557)	0.0709
	Fish	PLS	1.2266	1.0616 (0.2042)	0.3457
		PLAD	1.2713	1.1907 (0.2566)	0.3102
	Octane	PLS	0.2409	0.3555 (0.1533)	0.0482
		PLAD	0.2740	0.2511 (0.0478)	0.0505

are displayed together. The larger intervals for PLAD regression coefficients show them to be more variable than the PLS coefficients.

6 Experimental data

We consider two sets of data constructed from simulation. The first from a switching regression model is an example where the results of PLS and PLAD differ because they are estimating different features of the data. The second illustrates the effect of outlier contamination on a standard factor model often employed to illustrate PLS regression (Martens and Naes 1989).

6.1 Data from a switching regression model

There are underlying latent variables here with distributions $M \sim$ Uniform $(0,1)$, $Z \sim$ N$(0,1)$, and, to induce a switch, $S \sim$ Bernoulli(0.65) on $\{-1,1\}$.

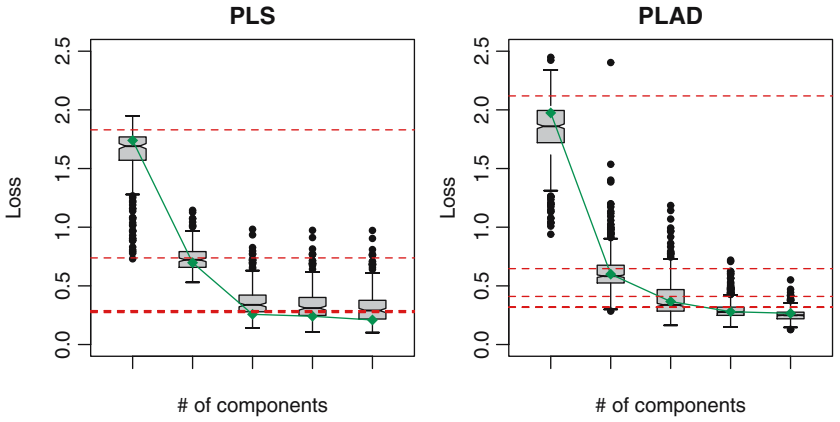

Fig. 1 Octane data. *Apparent* and *resampling prediction error* for PLS (*left panel*) and PLAD (*right panel*) regression models. The *apparent* error follows the *solid line* (green) while the *resampling* prediction error is summarized by *boxplots*. The *horizontal dashed lines* (red) correspond to the final prediction error

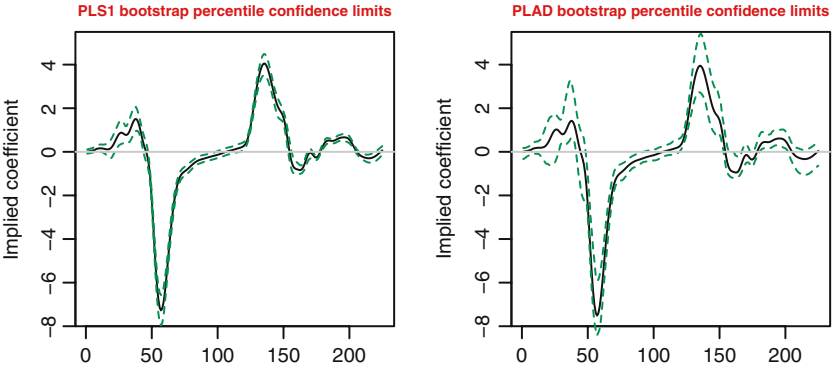

Fig. 2 Octane data. Regression coefficients for PLS (*left panel*) and PLAD (*right panel*) regressions using on three components. Percentile bootstrap confidence limits for percentiles 0.025 and 0.975 are displayed as *dashed lines* (green)

The covariates are partitioned into $X = (X_1, X_2)$, and are related to the latent variables by

$$X_1|M \sim \mathrm{N}(m1, I), \quad X_2|Z \sim \mathrm{N}(z1, I), \quad \text{and } Y|(S, M) \sim \mathrm{N}(sm + z, 1).$$

The non linear model generated shows that scatter plots of elements of X_1 with Y are somewhat triangular with the lower quantile decreasing with X_1 but the upper quartile increasing.

A sample of $n = 100$ observations based on $p = 20$ explanatory variables is generated, and the results are displayed in Fig. 4. The fitted median and the fitted mean resulting from the PLAD and PLS regression models are illustrated

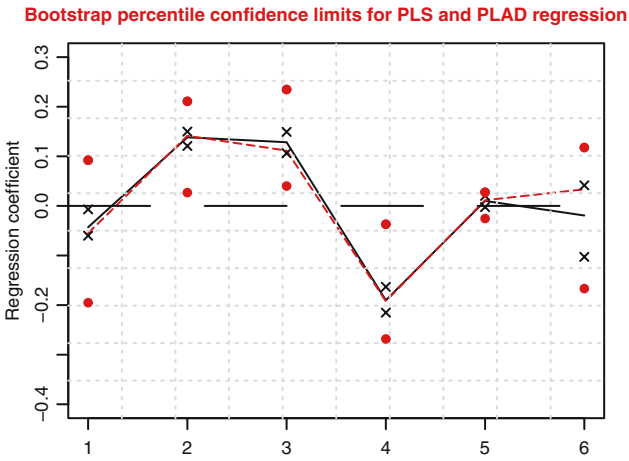

Fig. 3 Fearn's Wheat data. Regression coefficients for PLS (*black solid line*) and PLAD (*red dashed line*) regression models on three components. The percentile bootstrap confidence limits for percentiles 0.025 and 0.975 are displayed as points, with bullets for PLAD and crosses for PLS

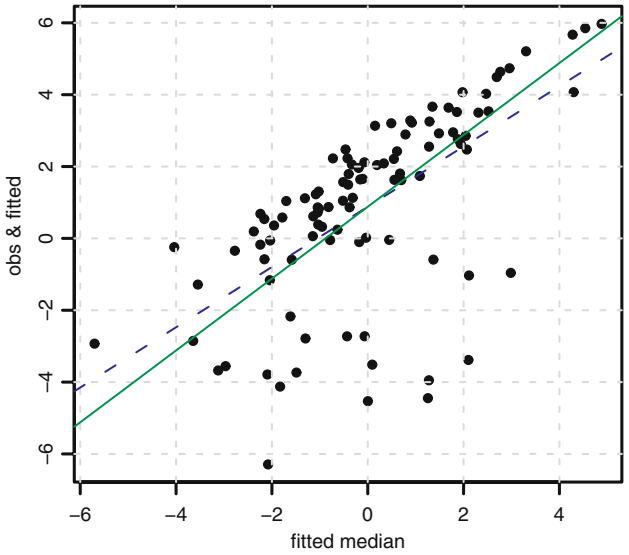

Fig. 4 Data from a switching regression model. The PLS and PLAD fitted values plotted versus the fitted median. The fitted median and the fitted mean are illustrated by the *dashed* and *solid line*, respectively

in Fig. 4 by the dashed and solid line, respectively. In this figure the fitted PLS and PLAD lines are plotted versus the fitted median. Differences between the two methods are here detected on the two fits and the PLS and PLAD lines are now well distinguished.

6.2 Data from the bilinear factor model

Both PLS and PLAD methods are based on a bilinear model (see Sect. 3) of the form

$$\boldsymbol{y} = q_1\,\boldsymbol{t}_1 + \cdots + q_k\,\boldsymbol{t}_k + \boldsymbol{\epsilon}, \quad \text{and} \quad \boldsymbol{X} = \boldsymbol{p}_1\,\boldsymbol{t}_1 + \cdots + \boldsymbol{p}_k\,\boldsymbol{t}_k + \boldsymbol{f}. \tag{15}$$

The bilinear model is described in (Martens and Naes 1989), and here forms the basis for our experimental data. In particular, we simulate the components $\boldsymbol{T}$ by taking n independent realisations of

$$\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{0}_k,\ \Sigma_{kk}).$$

The k-variate normal distribution has parameters $\boldsymbol{0}_k$ (as the data are centered) and the variance–covariance matrix Σ_{kk} is diagonal with specified variances.

$$\Sigma_{\boldsymbol{T}} = \begin{pmatrix} \mathrm{var}(\boldsymbol{t}_1) & 0 & \ldots & 0 \\ 0 & \mathrm{var}(\boldsymbol{t}_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathrm{var}(\boldsymbol{t}_k) \end{pmatrix}.$$

The data set $\boldsymbol{X}$ and $\boldsymbol{y}$ is obtained according to Eq. (15) for a specified choice of residual structure in $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$. We fix $k_{max} = 3$, $\boldsymbol{P} = \boldsymbol{I}_{3,p}$, and $\boldsymbol{q} = (1,1,1)^T$. Finally the matrix Σ_{kk} is set to $\mathrm{diag}\,(10,5,1)$, so that

$$\Sigma_{\boldsymbol{T}} = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We generate the artificial data set as follows: the elements of the error term $\boldsymbol{\epsilon}$ in expression (15) are independent normals, that is $\epsilon_i \sim \mathcal{N}(0, 0.01)$ where $i = 1, \ldots, n$. We then contaminate by replacing a small fraction, 10%, of the data set with outliers. The contaminated error vector is denoted as $\boldsymbol{\epsilon}_{cont}$ and its elements are generated according to $\epsilon_{i',cont} \sim \mathcal{N}(\mu, 1)$ with $\mu = 5$ and $i' = 1, \ldots, \ell$ for $\ell = 0.10 \cdot n$. The residual term $\boldsymbol{f}$ in expression (15) remains a random normal variate centered to zero, and no contamination on the X-space has been used. The dimensions of the data sets are set to $n = 100$ and $p = 50$. For the simulated data we apply bootstrap methods with R equal to 500.

Results

Table 2 gives the *apparent* and the *resampling* prediction error together with the *optimism* and the *FPE* estimate for both PLS and PLAD methods. For both regression models the simulation setting is verified since three components are

Table 2 Data from the bilinear factor model: the *apparent* and the *resampling* prediction error together with the *optimism* and the *FPE* estimate for PLS and PLAD methods

k	PLS			
	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$	*FPE*
1	1.5066	1.48 (0.0906)	0.0203	1.5269
2	1.2406	1.22 (0.0508)	0.0258	1.2665
3	1.0479	1.01 (0.0465)	0.0478	1.0958
4	0.9557	0.87 (0.0467)	0.1755	1.1312
5	0.9397	0.83 (0.0497)	0.2300	1.1697
k	PLAD			
	AE	RE (sd)	$\widehat{\omega}(\widehat{F})$	*FPE*
1	1.3667	1.42 (0.1265)	0.0392	1.4060
2	1.1951	1.26 (0.1160)	0.0401	1.2353
3	1.0902	1.15 (0.1000)	0.0476	1.1379
4	1.0873	1.11 (0.0719)	0.0462	1.1335
5	1.0810	1.09 (0.0630)	0.0499	1.1309

The standard deviations for the *resampling* error given in the parenthesis

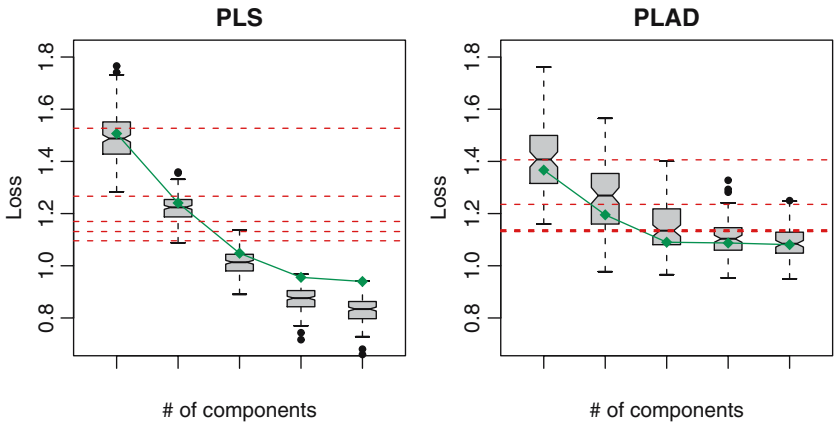

Fig. 5 Data from the bilinear factor model. *Apparent* and *resampling prediction error* for PLS (*left panel*) and PLAD (*right panel*) regression models. The *apparent* error is displayed by *solid lines* (green) while the *resampling* prediction error is summarized in *boxplots*. The horizontal *dashed lines* (red) correspond to the final prediction error

retained. The PLAD resampling prediction error is slightly more variable than the PLS resampling prediction error for all the components.

The available information on prediction error assessment for the contaminated case is displayed in Fig. 5. Figure 5 is similar to Fig. 1 with the horizontal dashed lines (red) indicating the final prediction error estimate (*FPE*) given by Eq. (14).

Both regression methods retain three components. In this sense contamination has no effect on model dimension neither for PLS nor for PLAD. Yet PLAD is accompanied by a larger variability of the resampling prediction

error. This is seen on the first three boxplots (that correspond to $k = 1, 2, 3$) in the right panel of Fig. 5. Note also that PLAD seems to be much less vulnerable to overfitting in comparison to PLS. Especially for $k \geq 3$ the PLAD *apparent error* is much closer to the *FPE* in comparison to PLS regression for which the *apparent error* decreases constantly with the number of the components.

The study of the regression coefficient vectors for the PLAD and the PLS regression methods for the two experimental data sets have led us to the following two conclusions. Firstly, the PLAD estimates are nearly twice more variable than PLS estimates on both data sets. Secondly, both PLS and PLAD regression vectors are subject to nearly the same loss compared to the regression vector $\boldsymbol{\beta}$ computed according to (4), for $\boldsymbol{q} = (1, 1, 1)$. The loss was taken as the vector norm for $(\widehat{\boldsymbol{\beta}}_k^{plad} - \boldsymbol{\beta})$ and $(\widehat{\boldsymbol{\beta}}_k^{pls} - \boldsymbol{\beta})$ for $k = 1, 2, 3$.

7 Conclusions

The PLAD regression has been tested and compared to PLS regression using the bootstrap. In the limited examples and experiments considered, we have established that PLAD and PLS estimate different features, but PLS is superior to PLAD in the sense that the implied regression coefficient estimates have smaller bootstrap confidence intervals; so that when the features are the same PLS is to be preferred. The magnitude of the difference is in line with the well known ratio of the standard deviation of the sample mean compared to that of the sample median when sampling from a Normal distribution. This gives some confidence in the basic structure of the PLAD algorithm when the response distribution is far from Normal. Furthermore, we have established that PLAD performs as well as PLS in model selection and prediction error assessment with final prediction error estimates nearly equal.

Using the bootstrap reveals two main drawbacks for PLAD regression. Firstly, PLAD estimates of prediction error are more variable than PLS estimates, so that achieving a small number of retained components might be more hazardous. Secondly, the PLAD regression method is computationally more expensive than PLS regression. This is due to the LAD algorithm as well as the GK-type covariance which demands the computation of $\boldsymbol{w}_k^{mad}$ for all the columns of matrix $\boldsymbol{X}$. A nice feature observed in the experimental studies is that PLAD is more resistant to overfitting in comparison to PLS. The apparent error from PLAD is always closer to the FPE, especially for $k \geq k_{max}$. In these cases the variability of PLAD resampling error reaches smaller levels which are comparable to those obtained from PLS.

This research has provided a further illustration of using the bootstrap to provide numerical solutions for analytically intractable problems. Further research on PLAD include (i) experiments with more distinctly non-Gaussian features, and (ii) computing error ellipses for component plots of the sample members based on the bootstrap.

References

Denham M (2000) Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. J Chemom 14:351–361
Diaconis P, Efron B (1983) Computer-intensive methods in statistics. Sci Am 248:116–130
Dodge Y, Kondylis A, Whittaker J (2004) Extending PLS to PLAD regression and the use of the L1 norm in soft modelling. In: Antoch J (ed) Proceedings in computational statistics, COMPSTAT'04. Physica-Verlag/Springer, Heidelberg, pp 935–942
Edgington ES (1995) Randomization tests. Marcel Dekker, New York
Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37:36–48
Efron B (1987) Better bootstrap confidence intervals. J Am Stat Assoc 82:171–185
Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York
Efron B, Tibshirani R (1997) Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 92:548–560
Fearn T (1983) A Missue of ridge regression in the calibration of a near infrared reflectance instrument. Appl Stat 32:73–79
Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81–124
Huber PJ (1981) Robust statistics. Wiley, New York
Kondylis A, Whittaker J (2005) Using the bootstrap on PLAD regression. In: Aluja T, Casanovas J, Vinzi VE, Morineau A, Tenehaus M (eds) PLS and related methods. Proceedings of the PLS'05 international symposium, Barcelona, pp 395–402
Martens H, Naes T (1989) Multivariate calibration. Wiley, UK
Naes T (1985) Multivariate calibration when the error covariance matrix is structured. Technometrics 27:301–311
Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494
Tenenhaus M (1998) La régression PLS. Théorie et pratique. Technip, Paris