Spatial Sign Preprocessing: A Simple Way To Impart Moderate Robustness to Multivariate Estimators

Sven Serneels,* Evert De Nolf, and Pierre J. Van Espen

Department of Chemistry, University of Antwerp, Universiteitsplein 1, 2610 Antwerpen, Belgium

Received November 14, 2005

The spatial sign is a multivariate extension of the concept of sign. Recently multivariate estimators of covariance structures based on spatial signs have been examined by various authors. These new estimators are found to be robust to outlying observations. From a computational point of view, estimators based on spatial sign are very easy to implement as they boil down to a transformation of the data to their spatial signs, from which the classical estimator is then computed. Hence, one can also consider the transformation to spatial signs to be a preprocessing technique, which ensures that the calibration procedure as a whole is robust. In this paper, we examine the special case of spatial sign preprocessing in combination with partial least squares regression as the latter technique is frequently applied in the context of chemical data analysis. In a simulation study, we compare the performance of the spatial sign transformation to nontransformed data as well as to two robust counterparts of partial least squares regression. It turns out that the spatial sign transform is fairly efficient but has some undesirable bias properties. The method is applied to a recently published data set in the field of quantitative structure−activity relationships, where it is seen to perform equally well as the previously described *best linear* model for these data.

1. INTRODUCTION

Increasing attention has recently been accorded to the development of multivariate nonparametric estimators. Parametric estimators have optimality properties at a considered model, generally (but not necessarily) the normal model. If these estimators are applied to data which follow a different model than the specified one, they are known to lose power rapidly in terms of being biased. Nonparametric estimators, however, do not require model specification. They are less biased than parametric estimators at a wide range of models deviating from the underlying model of the parametric estimator, whereas they only suffer from a slight increase in variance at that model. This wide range of models includes contaminated models, such as a normal model contaminated with outliers. As nonparametric estimators should perform well at these models, they are also robust estimators. Univariate and bivariate nonparametric estimators are easy to define and construct; their use is widespread (for comprehensive textbooks introducing univariate nonparametric statistics as well as some extensions to higher dimensions, see e.g. refs 1 or 2). Well-known examples of bivariate nonparametric estimators are the Spearman and Kendall correlation coefficients. At models which deviate from normality, they outperform the Pearson (classical) correlation coefficient. Recently, the robustness properties of both nonparametric estimators have been investigated. It has been shown[3] that they readily outperform other robust estimators such as the correlation coefficient based on the Minimum Covariance Determinant estimator:[4] they have a smooth, bounded influence function and do not lose much efficiency at the normal model. The Spearman correlation coefficient is based on ranks, whereas the Kendall correlation coefficient is based on signs. Many nonparametric statistical estimators are based either on these two concepts or on L1 distances. Most of these concepts are not straightforward to generalize to the multivariate setting, although several successful nonparametric multivariate estimators have been constructed.

In any multivariate statistical procedure, the covariance matrix assumes a key position. Hence, several attempts have been made to obtain nonparametric estimators for the covariance and correlation matrices. Visuri et al. have generalized the concepts of nonparametric estimation based on signs or ranks to covariance and correlation matrices.[5] They propose several generalizations of signs and ranks to multivariate statistics, which finally leads to six different sign and rank covariance estimators. They conclude that of these six estimators, it is preferrable to use the so-called Oja signs and ranks, because these lead to affine equivariant covariance matrix estimators. The properties of these affine equivariant sign and rank covariance matrices have since been investigated.[6,7] Various multivariate analysis techniques based on these covariance matrices, such as principal component analysis, have been proposed.[8] These methods, however, still do not resist extreme outliers (they have a zero break-down point) and are hard to compute in high dimensions.

In chemistry frequently very high dimensional data have to be modeled. Often a relation has to be modeled between these high-dimensional data and a dependent variable. For example, in the field of quantitative structure−activity relationships, a dependent variable of (biological) activities of molecules has to be predicted from a high-dimensional predictor matrix which consists of values of molecular descriptors. As there are normally less molecules in such a

* Corresponding author phone: +32/3/8202378; fax: +32/3/8202376; e-mail: sven.serneels@ua.ac.be. Corresponding author address: Departement Scheikunde, Universiteit Antwerpen, Universiteitsplein 1, 2610 Antwerpen, Belgium.

10.1021/ci050498u CCC: $33.50 © 2006 American Chemical Society
Published on Web 02/23/2006

data set than predictors, a multiple least squares model cannot be built. Solutions one can resort to are variable selection, such that only a few variables are left which optimally carry enough relevant information such that the predictand can be predicted with satisfactory accuracy. Another way to proceed is to summarize the information in the predictors into a smaller set of uncorrelated components, such as principal components, upon which regression can then be performed. Regression methods of this type which are frequently applied to chemical data sets are, among others, principal component regression (PCR) and partial least squares regression (PLS).

A seminonparametric alternative to these methods can be constructed using the eigenvectors of a sign correlation matrix. Due to computational complexity, the affine equivariant sign covariance matrix is not a good choice for these purposes. As an alternative, the spatial sign covariance matrix[5] (one of the remaining four estimators described in this paper) can be used. The spatial sign covariance matrix has received little attention in the statistical literature because it lacks the property of affine equivariance. Howbeit, as methods such as PCR and PLS do not have this property either, there is no objection to using a seminonparametric PLS estimator based on the spatial sign covariance matrix. From theoretical results concerning that estimator,[9] we can expect the estimators based on the spatial sign covariance matrix to be entirely robust (in contrast to those based on the affine equivariant sign covariance matrix) and very easy to compute. In fact, computing the spatial sign covariance matrix comes down to transforming the data to their spatial signs, followed by the computation of a normal covariance matrix. The same holds for PCR, PLS, etc. based on this covariance matrix. Hence, these methods can indeed be seen as being normal PCR, PLS, etc., applied to data which have been preprocessed to spatial signs. In what follows, we will refer to this transform as *spatial sign preprocessing (SS-PP)* and to the combined methods simply by *SS-PP+PLS,* etc.

A special case of the combination of spatial sign preprocessing with multivariate methods is SS-PP+PCA, which has been proposed in the literature as a method in its own right, remarkably 2 years earlier than the spatial sign covariance matrix.[10] As partial least squares is very frequently applied to chemical data, we will investigate in this article the properties of SS-PP applied to PLS. Similar results can be expected for SS-PP in combination with other linear modeling techniques such as PCR or RRR (*reduced rank regression*). It will turn out that SS-PP+PLS is very easy to compute, is acceptably efficient at the normal model, performs well at several non-normal models such as the Cauchy and Laplace models, and is moderately robust with respect to bad contamination in the data. However, its robustness properties cannot be tuned, which leads to several drawbacks.

The article is organized as follows. In the following section, the spatial sign transform is introduced. In section 3, SS-PP+PLS is shown to be a robust alternative to PLS. In section 4 the robustness properties of SS-PP+PLS are investigated, and in section 5 the method's applicability to quantitative structure−activity relationships is illustrated.

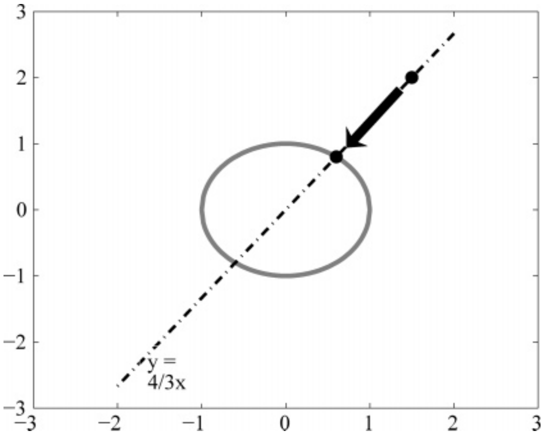

Figure 1. The spatial sign transform in two dimensions. The point (1.5,2) is transformed into the point (0.6,0.8).

2. THE SPATIAL SIGN AND THE SPATIAL SIGN COVARIANCE MATRIX

In the univariate setting the sign function, denoted by sgn(·), is defined as

$$\mathrm{sgn}(w) = \begin{cases} -1 & \text{if } w < 0 \\ 0 & \text{if } w = 0 \\ 1 & \text{if } w > 0 \end{cases} \tag{1}$$

Another way to write this equation is

$$\mathrm{sgn}(w) = \begin{cases} w/|w| & \text{if } w \neq 0 \\ 0 & \text{if } w = 0 \end{cases} \tag{2}$$

where |·| denotes the absolute value. This version of the sign function's definition can easily be generalized to the multivariate setting

$$\mathrm{sgn}(\mathbf{w}) = \begin{cases} \mathbf{w}/||\mathbf{w}|| & \text{if } \mathbf{w} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{w} = \mathbf{0} \end{cases} \tag{3}$$

where || · || denotes the Euclidian norm of its argument.

Geometrically, the spatial sign function is a projection of any point **w** in the direction of the origin onto a unit sphere. In two dimensions, this is illustrated in Figure 1.

As an example, the line $y = 4/3x$ is plotted. Except for the origin, all points on the line are transformed to their projections onto the unit circle.

For stochastical variables, it is clear that these have to be centered before the signs can carry a sizable amount of information. In what follows, if we refer to the spatial sign transform, we will implicitly assume that the data have been mean centered.

If we have a finite sample consisting of n observations of a p-variate stochastical variable $\mathbf{x}$, the spatial sign covariance matrix is defined as

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} \mathrm{sgn}(\mathbf{x}_i - \hat{\mu})\mathrm{sgn}(\mathbf{x}_i - \hat{\mu})^{\mathrm{T}} \tag{4}$$

To accomplish mean centering, several location estimators can be plugged into this equation. We propose to use the L1 median for these purposes, as the L1 median is a spatial median which can be computed very fast.[11] Indeed, the spatial sign covariance matrix is the classical covariance matrix applied to data which have been preprocessed with the spatial sign transform.

3. ROBUSTIFYING PLS BY THE SPATIAL SIGN TRANSFORM

Partial least squares regression is a regression technique which is able to model a linear relation to the predictand even if the predictors are multicollinear and in case $p > n$, given that the predictor data matrix is of dimensions $n \times p$. These properties of PLS are due to the fact that it is a *latent variables* regression technique: it first estimates a set of uncorrelated latent variables, whereafter it models the linear relationship of these latent variables to the predictand. Several algorithms for partial least squares regression exist, all of which follow an inductive scheme where at each phase a set of vectors is estimated (these vectors are not the same in each of the algorithms). For a univariate predictand, all algorithms yield identical predictions. We will not discuss further into detail the properties of the different algorithms. Instead we shortly describe the SIMPLS[12] algorithm.

Let $\mathbf{X} \in \mathscr{R}^{n\times p}$ be the predictor data matrix; its covariance matrix is estimated by $\mathbf{S} = (n-1)^{-1}\, \mathbf{X}^{\mathrm{T}}\mathbf{X}$, and its covariance to the predictand is estimated by $\mathbf{s} = (n-1)^{-1}\, \mathbf{X}^{\mathrm{T}}\mathbf{y}$. As starting values for the algorithm, we need $\mathbf{v}_0 = \mathbf{0}_p$ and $\mathbf{b}_0 = \mathbf{0}_p$. The algorithm goes as follows

$$\mathbf{a}_h = \begin{cases} \mathbf{s} & \text{for } h = 1 \\ \left(\mathbf{I}_p - \dfrac{\mathbf{v}_{h-1}\,\mathbf{v}_{h-1}^{\mathrm{T}}}{\mathbf{v}_{h-1}^{\mathrm{T}}\,\mathbf{v}_{h-1}}\right)\mathbf{a}_{h-1} & \text{for } h > 1 \end{cases} \tag{5a}$$

$$\mathbf{r}_h = \frac{\mathbf{a}_h}{\sqrt{\mathbf{a}_h^{\mathrm{T}}\mathbf{S}\mathbf{a}_h}} \tag{5b}$$

$$\mathbf{p}_h = \mathbf{S}\mathbf{r}_h \tag{5c}$$

$$\mathbf{v}_h = \left(\mathbf{I}_p - \frac{\mathbf{v}_{h-1}\,\mathbf{v}_{h-1}^{\mathrm{T}}}{\mathbf{v}_{h-1}^{\mathrm{T}}\,\mathbf{v}_{h-1}}\right)\mathbf{p}_h \tag{5d}$$

$$\mathbf{b}_h = \mathbf{b}_{h-1} + \mathbf{r}_h\,\mathbf{r}_h^{\mathrm{T}}\,\mathbf{s} \tag{5e}$$

This algorithm shows that all PLS vectors derive from the covariance parameters **S** and **s**. It is well-known that the covariance matrix of the augmented data $\mathbf{Z} = (\mathbf{X}|\mathbf{y})$ takes on the partitioned form:

$$\Sigma_{\mathbf{Z}} = \begin{pmatrix} \Sigma_{\mathbf{x}} & \sigma_{\mathbf{xy}} \\ \sigma_{\mathbf{xy}}^{\mathrm{T}} & \sigma_{\mathbf{y}} \end{pmatrix} \tag{6}$$

Since **S** and **s** are classical estimators for $\Sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{xy}}$, in fact all PLS vectors are based on the covariance matrix of augmented data. To construct a robust variant of PLS, one can simply plug in a robust variant of the covariance matrix such as the spatial sign covariance matrix.

It is easy to see that plugging a spatial sign covariance matrix into algorithm 5 is equivalent to applying the spatial sign transform to the data (*spatial sign preprocessing (SS-PP)*) and ensuingly using any standard PLS algorithm. This makes the routine very fast in terms of computation.

4. ROBUSTNESS PROPERTIES

4.1. Influence Function. The influence function[13] is a tool which leads to the theoretical assessment of (qualitative) robustness of an estimator. Loosely, the influence function measures the influence a small fraction of outliers, placed at any point in space, has on the estimator. If an estimator's influence function is bounded, the estimator is said to be robust.

The influence functions of all functionals corresponding to the PLS vectors can be defined sequentially, following the different steps of the SIMPLS algorithm.[14] The algorithm to compute the SIMPLS influence functions uses as a starting point the influence function of **s** and **S**. Since the SS-PP+SIMPLS algorithm is identical to the classical SIMPLS algorithm but uses a different plug-in for the covariance entities **S** and **s**, it can be expected that the same holds for the algorithms for the influence function. Indeed, the influence functions for the functionals corresponding to the SS-PP+SIMPLS vectors can be computed using the algorithm given in ref 14 but using as a starting value the influence functions of the spatial sign covariance matrix. It has been shown[9] that the influence functions for the spatial sign transformed entities $\Sigma_{\mathbf{x}}^{SS}$ and $\sigma_{\mathbf{x}}^{SS}$ equal

$$\mathrm{IF}(\Sigma_{\mathbf{x}}^{SS}) = \mathrm{sgn}(\mathbf{x})\mathrm{sgn}(\mathbf{x})^{\mathrm{T}} - \Sigma_{\mathbf{x}}^{SS} \tag{7}$$

and

$$\mathrm{IF}(\sigma_{\mathbf{Xy}}^{SS}) = \mathrm{sgn}(\mathbf{x})\mathrm{sgn}(y) - \sigma_{\mathbf{Xy}}^{SS} \tag{8}$$

In these equations, $(\mathbf{x}, y)$ is the point in space at which the influence function is being evaluated. Usually this is one of the measured data points (one of the rows of **X**). Since the sign function is bounded, the influence functions of the spatial sign covariance entities are bounded as well. Because all further SS-PP+PLS influence functions are derived from (7) and (8), the SS-PP+PLS estimates are robust.

4.2. Statistical Efficiency. As explained in the Introduction, at each model a parametric estimate can be found which has some optimality properties. For example, at the normal model, the least squares estimator for regression is known to be the minimum variance unbiased estimator. This implies that any robust or nonparametric estimator of regression will have a higher variance than the least squares estimator at this model. If the data considered do not follow the normal model, the least squares estimator loses power. Especially in the presence of outliers (both vertical outliers and leverage points) the least squares estimator is known to perform badly. In these contexts, a robust or nonparametric estimator will yield better results.

In analyzing a real data set one never knows if the data perfectly follow a model. If they do not, it is interesting to know if the deviation from the specified model (usually normality) is big enough to compensate for the increase in variance which can be expected if applying a robust estimator. One can also look at this problem from the opposite point of view and think of constructing robust estimators such that they are still efficient at the normal model but perform better than the nonrobust estimator at

Table 1. Simulated Mean Squared Error for the PLS, PRM, SS-PLS, and RSIMPLS Regression Estimators at Several Error Distributions (Standard Normal, Laplace, Student's *t* with 2 and 5 df, Cauchy and Slash) and under 5 Different Sampling Schemes of Sample Size *n* and Predictor Dimension *p*: 24 × 6, 15 × 60 (at Two Different Complexities), 10 × 100 and 15 × 1500

error distribution		N(0,1)	Laplace	t_5	t_2	Cauchy	Slash
$n/p = 4$, $h = 2$	PLS	0.0199	0.0425	0.0337	0.3432	48.011	37195
	PRM	0.0240	0.0315	0.0295	0.0435	0.070	0.1666
	SS-PLS	0.0326	0.0375	0.0362	0.0467	0.0544	0.0873
	RSIMPLS	0.0462	0.0521	0.0520	0.0672	0.1026	0.2105
$n/p = 1/4$, $h = 1$	PLS	0.0011	0.0021	0.0020	0.0112	30.742	10.923
	PRM	0.0013	0.0019	0.0018	0.0026	0.0047	0.0099
	SS-PLS	0.0020	0.0025	0.0023	0.0035	0.0060	0.0077
	RSIMPLS	0.0024	0.0027	0.0029	0.0036	0.0067	0.0135
$n/p = 1/4$, $h = 3$	PLS	0.0047	0.0098	0.0073	0.0737	41.85	113.08
	PRM	0.0055	0.0084	0.0072	0.0113	0.0258	0.0522
	SS-PLS	0.0071	0.0133	0.0105	0.0209	0.0434	0.0600
	RSIMPLS	0.0217	0.0326	0.0304	0.0540	0.1509	0.1643
$n/p = 1/10$, $h = 3$	PLS	0.0040	0.0078	0.0064	0.0582	124.62	188.32
	PRM	0.0047	0.0069	0.0061	0.0099	0.0283	0.0469
	SS-PLS	0.0138	0.0243	0.0209	0.0473	0.0992	0.1344
	RSIMPLS	0.0155	0.0247	0.0211	0.0346	0.0969	0.1517
$n/p = 1/100$, $h = 3$	PLS	0.0018	0.0019	0.0019	0.0030	1.4778	2.8315
	PRM	0.0018	0.0019	0.0018	0.0020	0.0024	0.0032
	SS-PLS	0.0019	0.0022	0.0021	0.0032	0.0102	0.0123
	RSIMPLS	0.0023	0.0028	0.0025	0.0033	0.0064	0.0075

deviating models. Recently, a highly efficient robust partial least squares estimator has been proposed,[15] the partial M-regression (PRM) estimator. The robustness properties of the estimator can be tuned as it uses an iterative algorithm which evaluates in each step a tuning function of which the tuning constant can be chosen by the user. As the SS-PP+PLS is a very "crude" robust estimator (no iterative optimization, no tuning functions, no tuning constants), its properties cannot be adapted by the user. It can also be expected that the spatial sign transform is not specific enough to equal the good efficiency properties of the partial robust M-estimator.

We carried out a simulation study to assess the efficiency properties of SS-PP+PLS. A predictor matrix was simulated with known model complexity. A thousand corresponding response vectors were simulated as

$$\mathbf{y}_m = \mathbf{X}\beta + \epsilon_m \tag{9}$$

where the vector of regression coefficients β is known, and ϵ is a random error term taken from one of the following distributions: the standard normal ($N(0, 1)$), Cauchy, Laplace, Slash, and Student's *t*, the last one at 2 and 5 degrees of freedom. For each of the $m \in (1{,}1000)$ response vectors, the vector of regression coefficients was estimated by PLS, SS-PP+PLS, PRM, and RSIMPLS,[16] a robust PLS method which is part of a recently distributed MATLAB toolbox for robust regression methods.[17] For PRM we used the settings of tuning function and tuning constant as in the original article,[15] and for RSIMPLS we used the standard settings from the LIBRA[17] toolbox. To summarize the results, mean squared errors were computed from the known and estimated regression coefficients ($\mathrm{MSE} = 1/m(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)$). The whole simulation was done for five different settings of the size of the predictor data matrix and the model complexity. The results of the simulation are shown in Table 1.

The mean-squared errors in Table 1 are affected by two contributing terms: bias and variance. Generally, for each case the trends seen vertically are due to variance, whereas the horizontal trends are caused by bias. Partial least squares can be seen as a "partial" version of the least squares estimator (in the sense that it is a least squares in the score space), such that it can be expected to perform best for normal error terms. This is indeed observed in each single situation. Of the robust estimates, the partial robust M-regression estimator is the most efficient at the normal model. This implies that the partial robust M-regression estimator is "safest" to apply in practice: it performs better than PLS at non-normal models and barely loses efficiency at the normal model. It is surprising to see that spatial sign preprocessed PLS outperforms RSIMPLS in most situations, except in the last simulation setting. This setting is, however, common in chemistry. This observation can to some extent be explained by the fact that the RSIMPLS is reminiscent of the minimum covariance determinant (MCD)[4] estimator for the covariance matrix, which is known to have a poor statistical efficiency as well.[18]

At models deviating from normality, it is observed that the robust estimators yield better results than PLS. In nearly all of the simulation settings, the partial robust M-estimator comes out best. For the Cauchy and Slash distributions (which are the distributions with the widest tails considered here), surprisingly SS-PP+PLS is best for low-dimensional data (atypical in a chemical context). We also think from these simulations a trend can be observed for spatial sign preprocessed PLS: it performs worse as the data dimensionality increases, an assumption corroborated in the following section. This trend is not observed for PRM nor for RSIMPLS. Finally, note that a higher model complexity leads to higher MSEs, regardless of the method used, due to an increase in variance. This is a well-known result which holds for multivariate calibration in general[19] and is not specific to PLS; hence, also the robust estimators follow this trend. Summarizing, from this simulation study we conclude that spatial sign preprocessed PLS performs well, as it is a very simple method which needs few computations.

4.3. Bias Curves. Apart from the efficiency properties, it is interesting to know how the estimator behaves if some percentage of badly outlying points are present in the data. The way we assess this question is by means of the *bias*

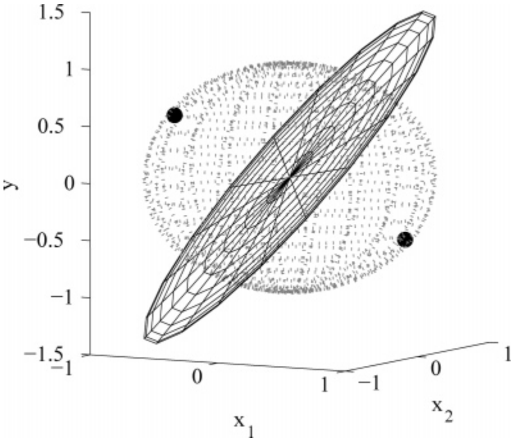

Figure 2. The plane $y = x_1 + x_2$ and two of the corresponding bad contamination points in the direction perpendicular to this plane.

curve: a curve which plots the bias an estimator has versus the percentage of contaminated data. Bias curves are often derived analytically; for PLS and its robust counterparts it is not clear how such analytical results can be obtained. Therefore, we propose a numerical alternative.

The bias curves presented in this section are computed as follows. At first predictor data are simulated with a fixed model complexity, as $\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}$. The error term, necessary to stabilize computations, is composed of standard normal random numbers divided by 100. Predictand data are computed such that the points $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ are elements of the hyperplane defined by the equation $\mathbf{y} = \mathbf{X}\beta$. Subsequently, 2, 4, ..., 100 percent of the data are replaced by bad outliers. From these contaminated data sets the vector of regression coefficients is estimated ($\hat{\beta}$). The angle between the hyperplanes $\mathbf{y} = \mathbf{X}\beta$ and $\mathbf{y} = \mathbf{X}\hat{\beta}$ is taken as a measure for the bias due to the outliers.

The outliers with which to replace data points should ideally be of the worst possible type of contamination. For spatial sign preprocessed PLS, bad contamination is readily found as these points which lie on a line through the origin and perpendicular to the hyperplane $\mathbf{y} = \mathbf{X}\beta$ (except for the origin itself). As an example, in Figure 2 we illustrate bad contamination for spatial sign preprocessed PLS.

The situation where $\beta = \mathbf{1}_2$ is considered. Bad contamination is found on the line perpendicular to this plane, i.e., the line with equations

$$y = -\frac{x_1 + x_2}{2}$$

$$x_1 = x_2$$

As an example, we plot both points on the intersection of this line with the unit sphere, as these two points are the image of the spatial sign projection of all points on this line. These points are $\pm (\sqrt{3}/3, \sqrt{3}/3, -\sqrt{3}/3)$. For PLS and the two remaining robust methods, bad contamination consists of the point at infinity of the same line. In the computations, the point at infinity was replaced by $\pm 1000(\sqrt{3}/3, \sqrt{3}/3, -\sqrt{3}/3)$ (choosing a bigger multiplication factor did not seem to alter the results significantly).

The bias curves for PLS, PRM, and RSIMPLS and spatial sign preprocessed PLS for these bivariate data are plotted in Figure 3. From these curves it can be seen that PLS is a nonrobust method: already a single outlier severely distorts the estimates. The robust methods are far less affected by the outliers. PRM and RSIMPLS show bias behavior which can be expected from a robust estimator: up to a certain amount of contamination, they are virtually not affected by the outliers, whereas at this percentage the bias curve increases drastically (the estimator breaks down). Let the estimate of the breakdown point, i.e., the percentage of contamination at which the estimator is unreliable, be given by the argument at which the bias curve corresponds to a angle bigger than 45°. It can be seen that in terms of breakdown properties, PRM performs best. Spatial sign preprocessed PLS has a somewhat a-typical bias curve. The SS-PP+PLS estimate does not break down at a low percentage of outliers. However, even a small amount of outliers may have a significant effect on the outcome.

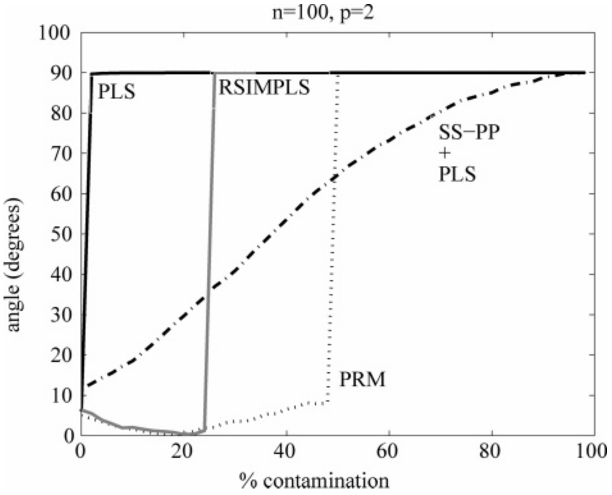

Figure 3. Bias curves for bivariate predictor data.

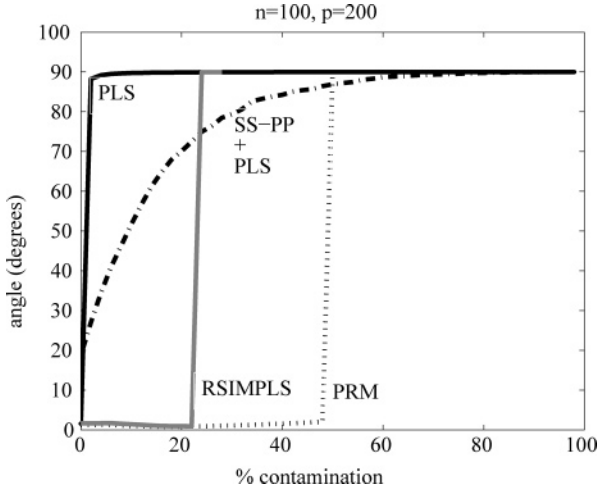

Figure 4. Bias curves for 200-variate predictor data.

We repeated the computation of the breakdown point for matrices of higher dimensionality. As an example, we show the results for $p = 200$. Analogously to the bivariate setting, bad contamination can be found in the direction perpendicular in the origin to the model hyperplane, such that the points of bad contamination with which the data were successively replaced, are $\pm (\{\sqrt{201}/201\}\mathbf{1}_{200}, -\sqrt{201}/201)$, where $\mathbf{1}_n$ denotes an n vector of which all entries equal 1. The thus obtained bias curves are shown in Figure 4.

Several things can be concluded from the picture shown both in Figures 3 and 4. At first, both PRM and RSIMPLS behave as expected for robust methods, i.e., up to a certain amount of contaminated data, the results are almost unbiased. Both methods do not seem to suffer from increasing predictor dimensionality. For both methods, the “standard” simulation settings were used (as in the simulations for efficiency in the previous section). For these settings, it is observed that PRM has better breakdown properties than RSIMPLS. PRM has a 50% breakdown point, whereas RSIMPLS breaks down at 25% of contamination. The breakdown properties of RSIMPLS can be improved by choosing different settings than those from the LIBRA[17] toolbox but inevitably at a cost of efficiency loss. Notably RSIMPLS using the current settings is already outperformed by PRM in terms of efficiency.

Spatial sign preprocessed PLS has peculiar robustness properties. Whereas it needs a very high amount of contamination to estimate a hyperplane perpendicular to the one which truly describes the data, it shows a non-negligible bias at low degrees of contamination. Moreover, when comparing both figures, it can be seen that the bias at low degrees of contamination even increases when the number of predictors increases. We investigated this behavior by computing the bias in function of the number of predictor variables at several fixed percentages of bad outliers. A trend is observed showing that indeed spatial sign preprocessed PLS systematically performs worse as the dimensionality increases, although the increase in bias per added predictor is most significant when comparing low predictor dimensions (up to 20).

From these plots we should thus conclude that SS-PP+PLS leads to moderately robust estimates which can be, however, obtained much faster than their robust competitors. Especially when not only the estimation of the regression parameters is desired but also the model complexity has to be investigated (e.g. by cross-validation), SS-PP+PLS does not require excessive computational efforts, whereas cross-validation on PRM may be more time-consuming.

5. APPLICATION TO QSAR DATA

In this section we will show the results of SS-PP+PLS to a data set recently described in this journal.[20] The data set consists of entries of 210 molecular descriptors for 79 piperazyinylquinazoline analogues which exhibit properties in inhibiting cell growth, which is important in anticancer research. The dependent variable considered is inhibition of the platelet derived growth factor receptor (PDGFR), which has been measured in the presence of human plasma.[20] It is reported in IC_{50} units of biological activity.

The molecular descriptors often describe analogous molecular properties, such that they are highly collinear variables in the data matrix. Other descriptors are not sensitive enough to the compounds included such that they hardly vary with respect to the different compounds. Both types of redundant predictors were eliminated. E.g., of each pair of predictors whose Pearson correlation exceeds 0.75, one predictor was eliminated. In addition to “crude” elimination of redundant variables, the uninformative variables elimination algorithm of Centner et al.[21] was applied to the data, such that the final data matrix was of size 79×38.

The data were split into a model set and a test set by the Kennard and Stone[22] algorithm. The model set contained 66 of the cases available, whereas the test set contained the remaining 13 cases. These sizes are equal to those chosen in the original paper;[20] howbeit, we cannot know whether the model used here and the model set used in ref 20 contained the same cases.

In the aforementioned paper, the authors proceed by applying different calibration techniques to these data. They either use a complex genetic algorithm to find the three most informative variables, upon which they use ordinary least squares regression (after detection and removal of an outlier). This model is referred to as the *best linear model*. As an alternative, the authors investigate the applicability of nonlinear calibration techniques. They show that application of a nonlinear artificial neural network (ANN) outperforms the best linear model.

Partial least squares regression, as has been explained in the Introduction, is also a full **X** block method which does not require more cases than explicative variables. We applied PLS regression to the aforementioned model set, using all 38 predictor variables. The PLS model was validated by dint of Monte Carlo cross-validation, with 33 objects being left out randomly in each run and 2*n* runs being computed. These settings were shown to be the most reliable to estimate model complexity[23,24] (the optimal number of latent variables). Monte Carlo cross-validation showed that the optimal number of latent variables to be used for PLS was equal to seven. This calibration we refer to as *PLS-full*. Consecutively we computed the squared influence diagnostic,[14] based on which we decided to delete an outlier from the model set. PLS calibration was again performed on the model set without outlier; seven latent variables were estimated to be the optimal number. This run we denominate *PLS-no out*.

For quantitative structure−activity relationships, often support vector machines (SVM) are applied to model the type of data described here. It is thus straightforward to provide a limited comparison of the obtained results to support vector machines. As all methods proposed in this article belong to the class of linear modeling techniques, the comparison to support vector machines will be limited to so-called *least squares support vector machines (LS-SVM)*, with linear and robust kernels[25] (denoted as *L-LS-SVM* and *R-LS-SVM*, respectively). The computations concerning support vector machines were done by means of the LS-SVMLab 1.5 toolbox[26,27] for MATLAB, using its standard parameter settings and for the full data set.

Although only one evident outlier was present in the model set, the squared influence diagnostic unveiled that some cases were clearly more influential to calibration than others, which advocates for the application of robust statistics. Hence, both partial robust M regression and spatial sign preprocessed PLS were applied to the full model set (all 79 cases, all 38 variables). Again the optimal complexity was estimated to be seven.

Based on all models, the response for the cases from the test set were computed. The results of these computations are shown in Table 2. The results show that the best linear model of Guha and Jurs and the PRM and SS-PP+PLS models perform equally well for this data set. The least squares support vector machine is clearly outperformed by PLS-based methods: the linear LS-SVM yields worse results

Table 2. Root Mean Squared Errors of Prediction for an Independent Test Set, Based on Different Models

method	RMSEP	method	RMSEP
L-LS-SVM	1.10	PRM	0.47
PLS-full	0.72	SS-PP+PLS	0.46
PLS-no out	0.65	best linear[a]	0.47
R-LS-SVM	0.55	ANN[a]	0.32

[a] Results are taken from ref 18.

than PLS for the full data set; the same holds for the robust LS-SVM compared to the robust PLS variants. PLS on the full data set is significantly worse than any of the robust methods due to the outlier. Removal of the outlier does not fully counter the effects of non-normality, as PLS applied to the data set without the outlier is still significantly inferior to the robust methods. Remarkably, the effect of nonlinearity can in this case alternatively be corrected for by variable selection, as the best linear model shows. A possible explanation thereto can be that after outlier removal, the data follow a marginally normal trend in the three selected variables, whereas on the whole the trend deviates from normality.

The robust methods are robust linear methods. If the data intrinsically follow a nonlinear trend, the effect of nonlinearity will not be countered by applying a robust technique. Hence, the artificial neural network outperforms all linear estimation techniques.

The goal of the present study is not to outperform the results obtained by Guha and Jurs for this data set but to show the effect of robust estimation techniques to QSAR data. We can conclude that the robust estimation techniques applied to the full data matrix, without further variable selection nor outlier removal, perform as well as a quite complex amalgam of a genetic algorithm, outlier detection, and removal and ordinary least squares regression. It is clear that the robust methods yield these results in a much shorter time span and are conceptually more straightforward.

6. SUMMARY AND OUTLOOK

In this paper, we have proposed a new method of data preprocessing for multivariate calibration based on the transformation of the data to their spatial signs. Preprocessing the data by the spatial sign transform before applying some multivariate calibration technique ensures moderate robustness to non-normality.

The technique of spatial sign preprocessing is conceptually simple and requires limited computational efforts. However, owing thereto, the technique must be used as it is, because it does not contain any parameter or function which can be tuned. In this respect, it is inferior to robust estimators specifically designed to be a robust counterpart of multivariate calibration techniques.

In the context of partial least squares regression, the properties of the spatial sign transform have been investigated in simulations and by examining a data set from quantitative structure−activity relationships. The major conclusion to draw is that the spatial sign transform does indeed lead to a robustification of partial least squares which is easy to compute. However, it should only be applied in these cases where computation times are an issue, as the bias properties of the spatial sign transform are far less desirable than the bias properties of the robust alternatives to PLS.

In the data set examined, the spatial sign transform performs as well as partial robust M-regression, probably due to the low predictor dimensionality (38). Howbeit, judging from the simulation studies, one cannot expect this to be a general trend. One can expect PRM to outperform the spatial sign transform although at a higher computational cost. Remarkably, direct application of robust methods to the data yields results which are as good as those previously obtained by successive application of outlier detection, outlier rejection, variable selection by means of a genetic algorithm, and least squares regression. One can expect such results to be found as well for other QSAR data sets as these data sets carry a big natural variation which may not fit into the normal model.

Finally, application of the robust methods does not compensate for deviations from linearity. In future work, specific robust nonlinear PLS estimators shall be designed, and it should be investigated whether it is also meaningful to apply the spatial sign transform prior to performing nonlinear calibration.

ACKNOWLEDGMENT

Research financed by a Ph.D. Grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

REFERENCES AND NOTES

(1) Maritz, J. S. *Distribution-free statistical methods*; Chapman and Hall: London, 1995.

(2) Hollander, M.; Wolfe, D. A. *Nonparametric statistical methods*; Wiley: New York, 1999.

(3) Croux, C.; Filzmoser, P. Projection pursuit based measures of association. In press.

(4) Rousseeuw, P. J. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol B*; Grossmann, W., Pflug, G., Vincze, I., Wertz, W., Eds.; Reidel: Dordrecht, 1985; pp 283−297.

(5) Visuri, S.; Koivunen, V.; Oja, H. Sign and rank covariance matrices. *J. Statist. Plan. Infer.* **2000**, *91*, 557−575.

(6) Ollila, E.; Croux, C.; Oja, H. Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. *Statist. Sin.* **2004**, *14*, 297−316.

(7) Ollila, E.; Oja, H.; Croux, C. The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. *J. Multivariate Anal.* **2003**, *87*, 328−355.

(8) Visuri, S.; Ollila, E.; Koivunen, V.; Mottonen, J.; Oja, H. Affine equivariant multivariate rank methods. *J. Statist. Plan. Infer.* **2003**, *114*, 161−185.

(9) Croux, C.; Ollila, E.; Oja, H. Sign and rank covariance matrices: statistical properties and applications to principal component analysis. In *Statistical data analysis based on the L1-norm and related methods*; Dodge, Y., Ed.; Birkhauser: Basel, 2002; pp 257−271.

(10) Locantore, N.; Marron, J. S.; Simpson, D. G.; Tripoli, N.; Zhang, J. T.; Cohen, K. L. Principal component analysis for functional data. *Test* **1998**, *8*, 1−73.

(11) Hössjer, O.; Croux, C. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *J. Nonparametr. Stat.* **1995**, *4*, 293−308.

(12) de Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.***1993**, *42*, 251−263.

(13) Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; Stahel, W. A. *Robust Statistics: the approach based on influence functions*; Wiley: New York, 1986.

(14) Serneels, S.; Croux, C.; Van Espen, P. J. Influence properties of partial least squares regression. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 13−20.

(15) Serneels, S.; Croux, C.; Filzmoser, P.; Van Espen, P. J. Partial robust M-regression. *Chemom. Intell. Lab. Syst.* **2005**, *79*, 55−64.

(16) Hubert, M.; Vanden Branden, K. Robust methods for partial least squares regression. *J. Chemom.* **2003**, *17*, 537−549.

(17) Verboven, S.; Hubert, M. LIBRA: a MATLAB library for robust analysis. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 127−136.

(18) Croux, C.; Haesbroeck, G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Anal.* **1999**, *71*, 161−190.

(19) Faber, N. M. A closer look at the bias-variance trade-off in multivariate calibration. *J. Chemom.* **1999**, *13*, 185−192.

(20) Guha, R.; Jurs, P. C. Development of linear, ensemble and non-linear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.

(21) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* **1996**, *68*, 3851−3858.

(22) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.

(23) Baumann, K.; Albert, H.; von Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *J. Chemom.* **2002**, *16*, 339−350.

(24) Baumann, K.; von Korff, M.; Albert, H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications. *J. Chemom.* **2002**, *16*, 351−360.

(25) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific Pub. Co.: Singapore, 2002.

(26) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *LS-SVMlab: a Matlab/C toolbox for Least Squares Support Vector Machines*; Internal Report 02-44, ESAT-SISTA, K.U. Leuven (Leuven, Belgium), (presented at NIPS2002 Vancouver in the demo track), 2002.

(27) The LSSVMLab Toolbox can be downloaded at http://www.esat.kuleuven.ac.be/sista/lssvmlab/toolbox.html.

CI050498U