



## Robust PCA and classification in biosciences

Mia Hubert\* and Sanne Engelen

Department of Mathematics, Katholieke Universiteit Leuven, W. De Croyleaan 54,  
B-3001 Leuven, Belgium

Received on October 9, 2003; revised on December 23, 2003; accepted on January 7, 2004  
Advance Access publication February 26, 2004

### ABSTRACT

**Motivation:** Principal components analysis (PCA) is a very popular dimension reduction technique that is widely used as a first step in the analysis of high-dimensional microarray data. However, the classical approach that is based on the mean and the sample covariance matrix of the data is very sensitive to outliers. Also, classification methods based on this covariance matrix do not give good results in the presence of outlying measurements.

**Results:** First, we propose a robust PCA (ROBPCA) method for high-dimensional data. It combines projection-pursuit ideas with robust estimation of low-dimensional data. We also propose a diagnostic plot to display and classify the outliers. This ROBPCA method is applied to several bio-chemical datasets. In one example, we also apply a robust discriminant method on the scores obtained with ROBPCA. We show that this combination of robust methods leads to better classifications than classical PCA and quadratic discriminant analysis.

**Availability:** All the programs are part of the Matlab Toolbox for Robust Calibration, available at <http://www.wis.kuleuven.ac.be/stat/robust.html>.

**Contact:** Mia.Hubert@wis.kuleuven.ac.be

### INTRODUCTION

Principal components analysis (PCA) is a very popular dimension reduction technique (Jolliffe, 1986) that is often applied in chemometrics, engineering, computer vision and many other applied sciences. It is of particular interest if the original data are high-dimensional, as PCA reduces the number of variables (features) to a more manageable size (often  $< 10$ ). High-dimensional data are frequently encountered in the form of spectral data, or gene expression levels, and so they are often subject to a PCA analysis. In the next stage of the data analysis, the reduced dataset can then further be analyzed with a cluster analysis, or some classification technique (see, e.g. Alter *et al.*, 2000).

Although PCA is a powerful statistical tool, the results are highly affected by anomalous observations in the data. In this paper, we will illustrate these effects on several real datasets. To avoid the sensitivity toward outliers, we have recently developed robust PCA methods. The first approach,

based on projection pursuit (Hubert *et al.*, 2002), has been successfully applied in Model *et al.* (2002). Asymptotic results of this method are presented in Cui *et al.* (2003). In this paper, we will concentrate on the second and more recent method ROBPCA, which combines projection-pursuit techniques with robust covariance estimation in lower dimensions (Hubert *et al.*, 2004).

Matrices will be denoted with capital letters. We assume that our data matrix  $X$  has dimensions  $(n \times p)$  where  $n$  denotes the number of observations and  $p$  the number of variables. A vector is always indicated in bold, e.g.  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  stands for the  $i$ -th observation. Classical estimates are denoted by means of a tilde. The following abbreviations will be used: MCD for minimum covariance determinant, CPCA for classical PCA, ROBPCA for robust PCA, CQDA for classical quadratic discriminant analysis and RQDA for robust quadratic discriminant analysis.

### SYSTEM AND METHODS

#### Classical PCA

Principal components analysis is a dimension reduction technique. From the original set of variables  $X_j$ , classical PCA (CPCA) constructs a new set of uncorrelated and orthogonal variables  $\tilde{P}_j$ . They are linear combinations of the mean-centered variables  $\tilde{X}_j = X_j - \bar{X}_j$ , and are often called the loadings or the principal components. It is well-known that these loadings correspond with the eigenvectors of the sample covariance matrix  $S = 1/(n-1) \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$  of the data. For each loading vector  $\tilde{P}_j$ , the corresponding eigenvalue  $\tilde{l}_j$  of  $S$  tells us how much of the variability of the data is explained by  $\tilde{P}_j$  through the relation  $\tilde{l}_j = \text{Var}(\tilde{P}_j)$ . Usually, these loading vectors are sorted in descending order of the eigenvalues. Hence, the first  $k$  principal components explain most of the variability of the data.

After selecting  $k$ , we can project the  $p$ -dimensional data points onto the subspace spanned by the  $k$  loading vectors and compute their coordinates with respect to these  $\tilde{P}_j$ . This yields the scores

$$\tilde{\mathbf{t}}_i = \tilde{P}'(\mathbf{x}_i - \bar{\mathbf{x}}), \quad (1)$$

for each  $i = 1, \dots, n$ , which have trivially zero mean. With respect to the original coordinate system, the projected data

\*To whom correspondence should be addressed.

point is computed as the fitted value

$$\hat{\mathbf{x}}_i = \bar{\mathbf{x}} + \tilde{P}\tilde{\mathbf{t}}_i. \quad (2)$$

Note that the  $(p \times k)$  loading matrix  $\tilde{P}$  contains the loadings column-wise. The  $(k \times k)$  diagonal matrix  $\tilde{L} = (\tilde{l}_j)_j$  will be used to denote the eigenvalues (in decreasing order).

To choose the appropriate number of loadings  $k$ , there exist many criteria. A very popular graphical one is based on the scree plot, which exposes the eigenvalues in decreasing order. The index of the last component before the plot flattens is then selected. A more formal criterion considers the total variation that is explained by the first  $k$  loadings, and requires, e.g. that

$$\left( \sum_{j=1}^k \tilde{l}_j \right) / \left( \sum_{j=1}^p \tilde{l}_j \right) \geq 80\%. \quad (3)$$

The classical mean and the sample covariance matrix  $S$  are, however, very sensitive to outliers, and consequently also CPCA is affected by anomalous observations. We will illustrate this on a small artificial dataset in  $p = 4$  dimensions. The Hawkins–Bradu–Kass dataset (Rousseeuw and Leroy, 1987) consist of  $n = 75$  observations in which two groups of outliers are created, labeled 1–10 and 11–14. The first two eigenvalues explain already 98% of the total variation, so we select  $k = 2$ . The CPCA scores plot is depicted in Figure 1a.

In this figure, we can clearly distinguish the two groups of outliers, but we see several other undesirable effects. In the plot we have superimposed the 97.5% tolerance ellipse, defined by the set of vectors whose squared Mahalanobis distance is equal to the 0.975 quantile of the  $\chi^2$ -distribution with  $k$  degrees of freedom:

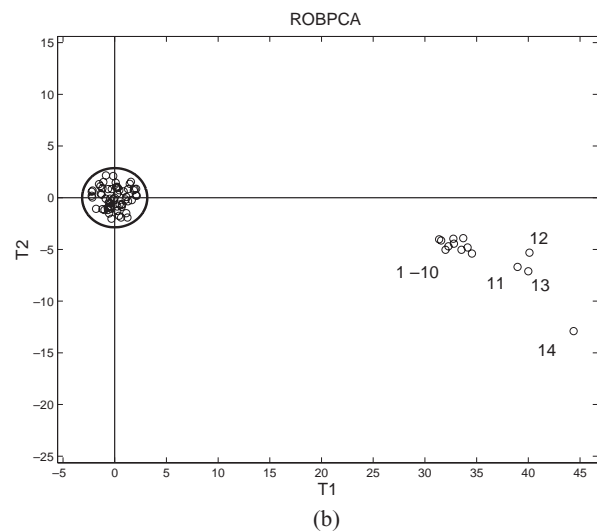
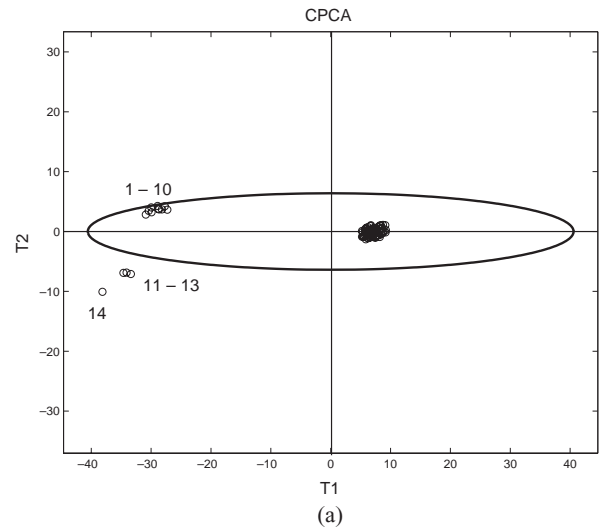
$$E_{0.975} = \left\{ \mathbf{t} \in \mathbb{R}^k; \text{MD}(\mathbf{t}) = \sqrt{\chi_{k,0.975}^2} \right\},$$

with

$$\text{MD}(\mathbf{t}) = \sqrt{\tilde{\mathbf{t}}' \tilde{L}^{-1} \tilde{\mathbf{t}}}. \quad (4)$$

In general, when we have  $k$ -dimensional data points that are normally distributed, the interior of the tolerance ellipse constitutes a 97.5% confidence region for the center of the data  $\boldsymbol{\mu}$ . The center of the ellipse is an estimate for  $\boldsymbol{\mu}$ , whereas the shape of the ellipse reflects the covariance structure between the variables. Here, the scores are uncorrelated, hence the principal axes of the ellipse are in the direction of the coordinate axes (the loadings). Data points that fall inside the ellipse can be classified as regular data points, because they are not too far from the estimated center taking into account the different variances of the loadings.

In Figure 1a, we see that, although the scores have zero mean, the regular data points are lying far from zero. This stems from the fact that the mean of the data points  $\bar{\mathbf{x}}$  is a bad estimate of  $\boldsymbol{\mu}$  in the presence of outliers. It is clearly shifted toward the outlying group and consequently the origin even



**Fig. 1.** Score plot and 97.5% tolerance ellipse of the Hawkins–Bradu–Kass dataset obtained with (a) CPCA and (b) ROBPCA.

falls outside the cloud of the regular data points. Moreover, the outliers 1–10 are within the tolerance ellipse, and thus are not recognized based on their Mahalanobis distance. The ellipse is highly inflated to accommodate these outliers.

### Robust PCA

Next, we have applied our ROBPCA method to the Hawkins–Bradu–Kass data. Here, we describe briefly the ROBPCA algorithm, all details of which can be found in (Hubert *et al.*, 2004). When the number of regressors  $p$  is smaller than the data size  $n$ , the Minimum Covariance Determinant (MCD) estimator is used (Rousseeuw, 1984, 1985). This location and covariance estimator is very popular because of its high resistance towards outliers and because a fast algorithm has been developed recently for its computation (Rousseeuw and Van Driessen, 1999). To define the MCD estimator, we consider

subsets of size  $h$  out of the whole dataset (of size  $n$ ). The number  $h$  determines the robustness of the estimator and should be at least  $\lceil (n + p + 1)/2 \rceil$ . The MCD estimator then seeks for that  $h$ -subset whose classical covariance matrix has a minimal determinant. The MCD location estimate  $\hat{\boldsymbol{\mu}}_{\text{MCD}}$  is given by the mean of that optimal  $h$ -subset, and the MCD scatter estimator  $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$  by its covariance matrix, multiplied by a consistency factor. Based on the raw MCD estimate, a reweighting step can be added which increases the finite-sample efficiency considerably. Each data point  $\mathbf{x}_i$  receives a weight 1 if its robust distance, defined as

$$\text{RD}(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})' \tilde{\mathbf{t}} \hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{-1} \tilde{\mathbf{t}} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})}$$

is smaller than  $c = \sqrt{\chi_{p,0.975}^2}$  and weight zero otherwise. The reweighted MCD estimator then equals the classical mean and covariance matrix of the data points with weight one. The first  $k$  eigenvectors of the MCD covariance matrix, sorted in descending order of the eigenvalues, then yield robust loadings.

It is intuitively clear that the MCD estimator can resist  $n - h$  outliers. More formally, it is said that the MCD estimator has a breakdown value of  $(n - h + 1)/n$ , which means that we need at least  $n - h + 1$  outliers to make the estimates worthless. The default choice for  $h$  is roughly  $0.75n$ , allowing for 25% outliers.

For high-dimensional regressors ( $p > n$ ), we cannot use the MCD anymore because the determinant of a covariance matrix of  $h < p$  observations will always be zero and thus cannot be minimized. ROBPCA then proceeds as follows. First, the  $x$ -data are preprocessed by reducing their data space to the affine subspace spanned by the  $n$  observations. This can be easily performed using a singular value decomposition of the data matrix  $X$ . As a result, the data are represented using at most  $n - 1 = \text{rank}(\tilde{X})$  variables without loss of information.

In the second step of the ROBPCA algorithm, a measure of outlyingness is computed for each data point. This is obtained by projecting the high-dimensional data points on many univariate directions  $\mathbf{v}$ . On every direction a robust center and scale of the projected data points  $\mathbf{x}_i' \mathbf{v}$  is computed, namely the univariate MCD estimator of location  $m_{\text{MCD}}$  and scale  $s_{\text{MCD}}$ . Next, for every data point its standardized distance to that center is measured. Finally, for each data point its largest distance over all the directions is considered. This yields the outlyingness

$$\text{outl}(\mathbf{x}_i) = \max_{\mathbf{v}} \frac{|\mathbf{x}_i' \mathbf{v} - m_{\text{MCD}}(\mathbf{x}_i' \mathbf{v})|}{s_{\text{MCD}}}$$

Note that the notion of outlyingness has been introduced by Stahel (1981) and Donoho (1982) and has recently been used by Zuo *et al.* (2004) to define projection depth as a tool to construct robust estimators of multivariate location.

The  $h$  data points with smallest outlyingness are then retained and from the covariance matrix  $\boldsymbol{\Sigma}_1$  of this  $h$ -subset,

the number of principal components to retain,  $k$ , is selected. The last stage of ROBPCA consists of projecting the data points onto the  $k$ -dimensional subspace spanned by the  $k$  largest eigenvectors of  $\boldsymbol{\Sigma}_1$  and of computing their center and shape by means of the reweighted MCD estimator. The eigenvectors of this scatter matrix then determine the robust principal components, and the MCD location estimate serves as a robust center.

Computation times for the ROBPCA method are reported in detail in Hubert *et al.* (2004). For the rat data ( $n = 112, p = 9$ ) and for the mice data ( $n = 30, p = 2050$ ), which we will analyze in the Discussion section, the computation times are, respectively, 3 and 4.7 s on a Pentium IV with 2.40 GHz.

The result of the ROBPCA analysis is thus a robust estimate of the center of the data  $\hat{\boldsymbol{\mu}}$ , a set of robust loadings  $P$  and eigenvalues  $l_j$  ( $j = 1, \dots, k$ ) and, similar to (1), robust scores

$$\mathbf{t}_i = P'(\mathbf{x}_i - \hat{\boldsymbol{\mu}}). \tag{5}$$

The score plot of the Hawkins–Bradu–Kass data obtained with ROBPCA is shown in Figure 1b. We now see that the center is correctly estimated in the middle of the regular observations. The 97.5% tolerance ellipse nicely encloses these points and excludes all the 14 outliers.

We can also represent the result of the PCA analysis by means of a diagnostic plot. This figure will highlight the outliers and classify them into several types. Hence, we can also call it an outlier map. An outlier is defined as an observation that does not follow the model followed by the majority of the data. In the context of PCA, this means that an outlier either lies far from the subspace spanned by the  $k$  eigenvectors, and/or that the projected observation lies far from the bulk of the data within this subspace. To measure this degree of outlyingness, we use two distances. The orthogonal distance measures the distance between an observation and its projection in the  $k$ -dimensional PCA subspace:

$$\text{OD}_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|. \tag{6}$$

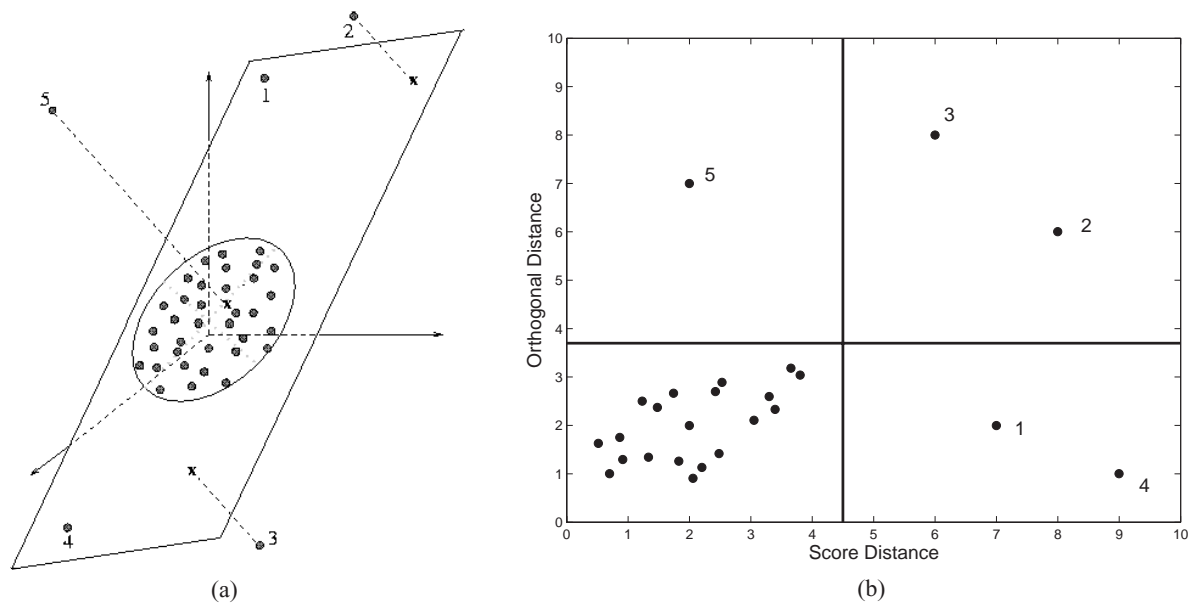
The score distance is measured within the PCA subspace, where, due to the knowledge of the eigenvalues, we have information about the covariance structure of the scores. Hence, we define the score distance as in (4)

$$\text{SD}_i = \sqrt{\mathbf{t}_i' L^{-1} \mathbf{t}_i} = \sqrt{\sum_{j=1}^k (t_{ij}^2 / l_j)}. \tag{7}$$

When using CPCA, this score distance thus corresponds exactly with the Mahalanobis distance.

The orthogonal and the score distances now define four types of observations, as illustrated in Figure 2a.

Regular observations have a small orthogonal and a small score distance. When samples have a large score distance, but a small orthogonal distance, we call them good leverage points. Observations 1 and 4 in Figure 2a can be classified



**Fig. 2.** (a) Different types of outliers when a three-dimensional dataset is projected on a robust two-dimensional PCA-subspace. (b) The corresponding outlier map.

into this category. These observations lie close to the space spanned by the principal components but far from the regular data. This implies that they are different from the majority, but there is only a little loss of information when we replace them by their fitted values in the PCA-subspace. Orthogonal outliers have a large orthogonal distance, but a small score distance, e.g. case 5. They cannot be distinguished from the regular observations once they are projected onto the PCA subspace, but they lie far from this subspace. Consequently, it would be dangerous to replace that sample with its projected value, as its outlyingness would not be visible anymore. Bad leverage points, such as observations 2 and 3, have a large orthogonal distance and a large score distance. They lie far outside the space spanned by the principal components, and after projection far from the regular data points. Their degree of outlyingness is high in both directions, and typically they have a large influence on CPCA, as the eigenvectors will be tilted toward them.

The outlier map displays the  $OD_i$  versus the  $SD_i$ , and, hence, classifies the observations according to Figure 2b. In this plot, we have drawn lines to distinguish the observations with a small and a large orthogonal distance, and with a small and a large score distance. For the latter distances, we use the property that normally distributed data have normally distributed scores, and consequently their squared Mahalanobis distances have a  $\chi_k^2$  distribution. Hence, we use as cut-off value  $c = \sqrt{\chi_{k,0.975}^2}$ . To obtain a cut-off for the orthogonal distances, we use the approximation proposed by Box (1954). The squared orthogonal distances can be approximated by a scaled  $\chi^2$  distribution with  $g_1$  degrees of freedom  $OD^2 \sim g_2 \chi_{g_1}^2$ .

Robust estimates for  $g_1$  and  $g_2$  are derived using the Wilson–Hilferty transformation to normality (see Hubert *et al.*, 2004, <http://www.wis.kuleuven.ac.be/stat/robust.html>).

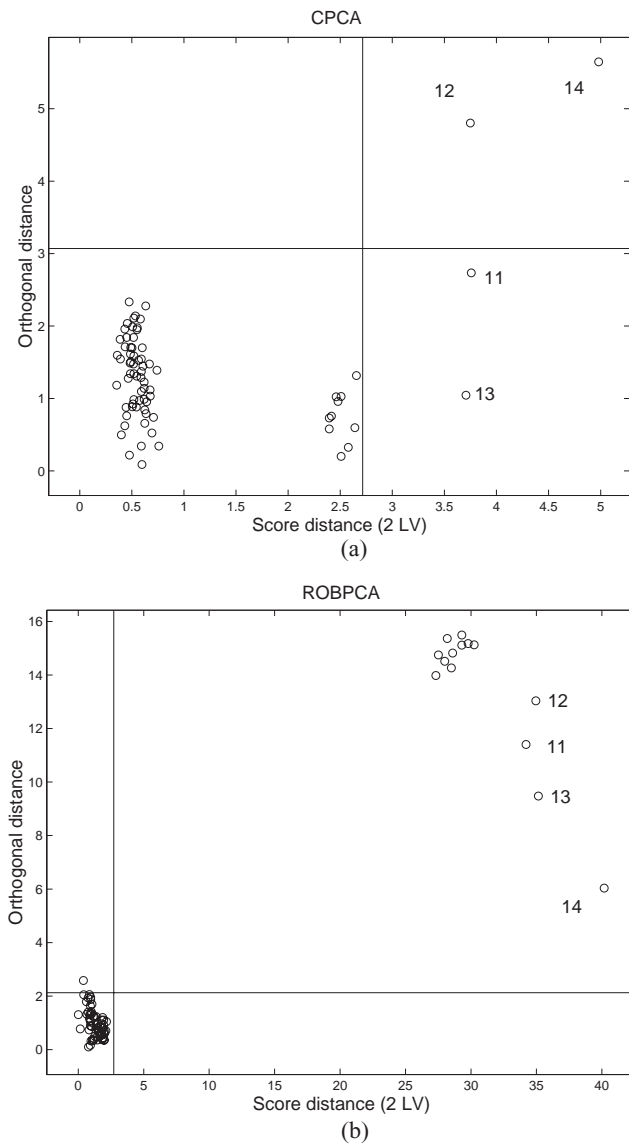
Let us return to the Hawkins–Bradu–Kass data and take a look at the outlier map based on CPCA, displayed in Figure 3a. In this plot, we can again identify the two groups of outliers, but in accordance with Figure 1a, observations 1–10 do not have an outlying score distance. So CPCA did not succeed in detecting all the outliers. If we would blindly apply CPCA on these data without looking at the resulting score plot or the diagnostic plot, we would thus conclude that only cases 11–14 are different from the other observations.

The robust outlier map of Figure 3b again gives a much better representation of the data. Both clusters of outliers are highly separated from the regular data, and they are both identified as bad leverage points.

From this artificial example, we already see that ROBPCA can handle data with outliers in a much better way than CPCA. In the Discussion section, we show that this also holds for real datasets. First, we analyze the rat dataset that contains gene expressions of 112 genes (of albino rats) at nine different time points. Next, we apply ROBPCA on the mice dataset, which consists of high-dimensional NMR spectra for two treatment groups. We will also show that ROBPCA, combined with a robust discriminant rule, yields lower probabilities of misclassification than a classical approach.

### Robust discriminant analysis

Let us briefly explain the classical and robust quadratic discriminant analysis. For all details we refer to Hubert and Van



**Fig. 3.** The outlier maps of the Hawkins–Kass–Bradru data, based on two principal components obtained with (a) and ROBPCA (b) ROBPCA.

Driessen (2004). Discriminant analysis (or supervised learning) construct discriminant rules from a dataset in which the group structure is known. These rules then allow to classify new observations into one of the groups. We denote the number of groups by  $l$  and assume that we can describe our experiment in each population  $\pi_j$  by a  $p$ -dimensional random variable  $X_j$  with distribution function (density)  $f_j$ . Moreover, we denote  $p_j$  as the membership probability, i.e. the probability for an observation to come from  $\pi_j$ . The maximum-likelihood rule then classifies an observation  $\mathbf{x} \in \mathbb{R}^p$  into  $\pi_k$  if  $\ln(p_k f_k(\mathbf{x}))$  is the maximum of the set  $\{\ln(p_j f_j(\mathbf{x})); j = 1, \dots, l\}$ . If we assume that the density  $f_j$

for each group is gaussian with mean  $\mu_j$  and covariance matrix  $\Sigma_j$ , then it can easily be derived that the maximum-likelihood rule is equivalent to maximizing the discriminant scores  $d_j(\mathbf{x})$  with

$$d_j(\mathbf{x}) = -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x} - \mu_j)' \Sigma_j^{-1} (\mathbf{x} - \mu_j) + \ln(p_j). \quad (8)$$

So,  $\mathbf{x}$  is allocated to  $\pi_k$  if  $d_k(\mathbf{x}) > d_j(\mathbf{x})$  for all  $j = 1, \dots, l$  (see e.g. Johnson and Wichern, 1998).

In practice,  $\mu_j$ ,  $\Sigma_j$  and  $p_j$  have to be estimated. Classical quadratic discriminant analysis (CQDA) uses the group mean  $\bar{\mathbf{x}}_j$  and empirical covariance matrix  $S_j$  as estimators of  $\mu_j$  and  $\Sigma_j$ . The membership probabilities are usually estimated by the relative frequencies of the observations in each group, hence  $\hat{p}_j^C = n_j/n$  with  $n_j$  the number of observations in group  $j$ . A robust quadratic discriminant analysis (RQDA) is derived by using robust estimators of  $\mu_j$ ,  $\Sigma_j$  and  $p_j$ . In particular, we can apply the reweighted MCD estimator of location and scatter in each group. As a byproduct of this robust procedure, outliers (within each group) can be distinguished from the regular observations. Finally, the membership probabilities can be robustly estimated as the relative frequency of regular observations in each group.

## IMPLEMENTATION

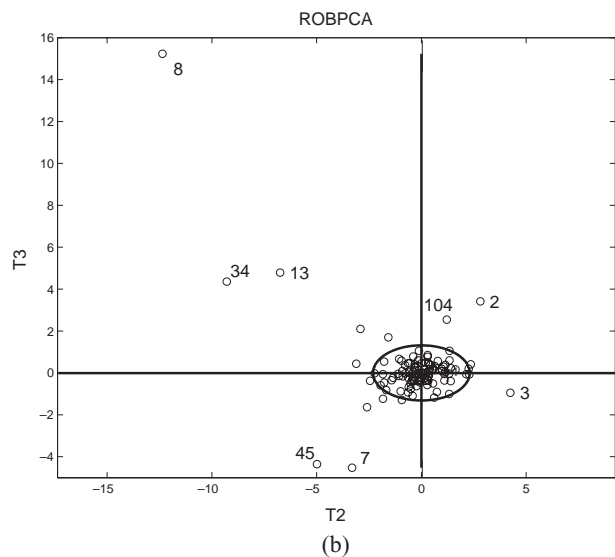
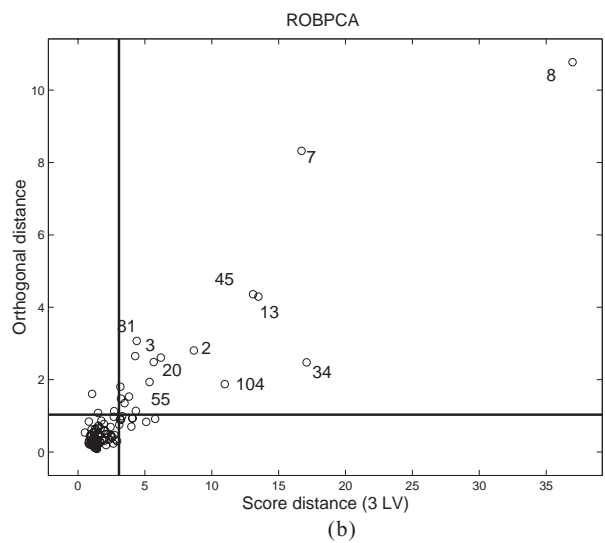
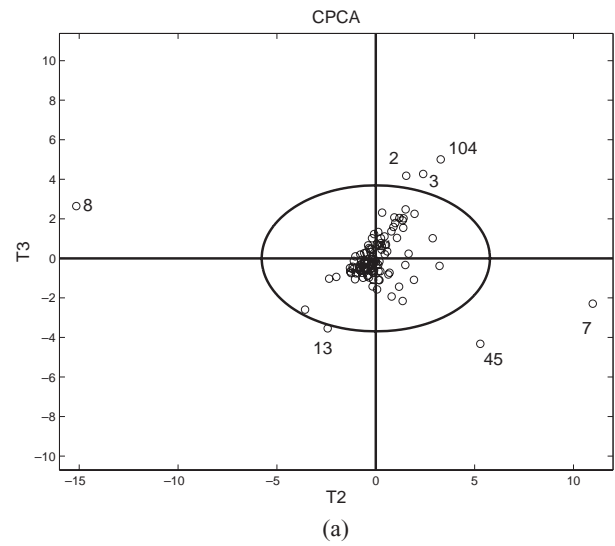
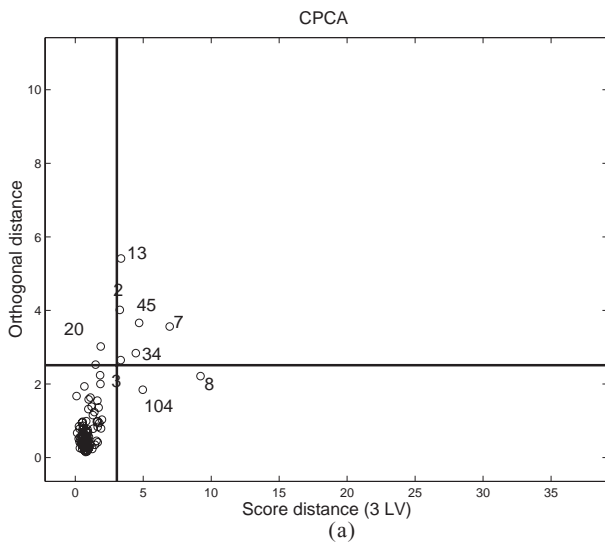
All the programs are part of the Matlab Toolbox for Robust Calibration, available at <http://www.wis.kuleuven.ac.be/stat/robust.html>

## DISCUSSION

### Analysis of the rat data

The rat data that are used in Wen *et al.* (1998) help investigate the development of the central nervous system (CNS) from Sprague–Dawley albino rats based on the temporal expression patterns in the cervical spinal cord. Another point of interest is whether it is possible to find clusters between genes by examining their patterns of expressions over the development period. The data contain the expressions of  $n = 112$  genes. Measurements were taken on  $p = 9$  different points in time. To be able to perform a clustering on the genes, they are put in the rows of the data matrix. The first five variables correspond with the embryonic days 11 until 21. The following three variables represent the first 14 postnatal days whereas the last variable corresponds with adulthood. Each row in the data matrix contains the average expression of a group of three animals.

In Wen *et al.* (1998), first a clustering was performed and then CPCA with  $k = 3$  was done to check whether the three main clusters could be recognized on the three-dimensional score plot. Note that the authors have normalized the data by dividing each row by its maximum value. This approach is, however, not very robust as it depends heavily on the largest



**Fig. 4.** Outlier maps of the rat data obtained with (a) CPCA and (b) ROBPCA.

**Fig. 5.** The score plot of the second and the third scores of the rat data obtained with (a) CPCA and (b) ROBPCA.

value measured for each gene. Hence, we did not apply this normalization and performed the analysis on the raw data.

From the scree plots obtained with CPCA and ROBPCA, we decided to retain  $k = 3$  components for both analysis. The ratio (3) then yields 95% for CPCA as well as for ROBPCA.

The corresponding diagnostic plots are drawn in Figure 4. CPCA finds as good leverage points the observations 8 and 104 and as bad leverage points 3, 34, 7, 45, 2 and 13. When we look at the results of the robust analysis in Figure 4b we see that the score distances of the outliers have increased a lot. Moreover, gene 8 is detected as a bad leverage point with a very unusual orthogonal distance. Also, cases 7, 13 and 45 are much more prominent outliers than could be seen from the classical outlier map.

To find out which result is the most reliable, we made several score plots as in Figure 1. Among others, we looked at the projections of the observations onto the plane formed by the second and the third principal components, together with the corresponding tolerance ellipse.

This yields Figure 5a and b for CPCA and ROBPCA, respectively. We notice that the second component of CPCA is attracted by cases 7 and 8. Consequently, the tolerance ellipse is inflated, and the variability in the second component is even smaller than in the third component. ROBPCA, on the other hand, estimates the eigenvectors correctly and its tolerance ellipse only encloses the regular observations.

When we looked at the raw data, we noticed that the measurements of the outlying genes were also very different from

the other observations, because they all have higher expression levels in most of the variables. So our method was able to detect these differences automatically, whereas CPCA tries to convert bad leverage points into good leverage points and masks the outliers.

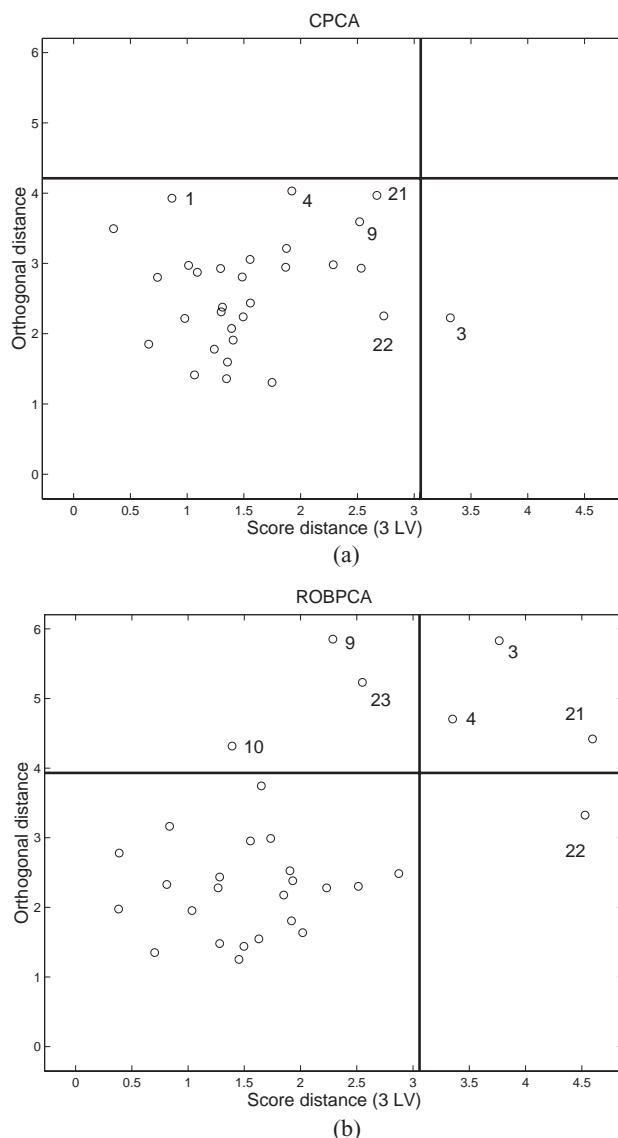
**Analysis of the mice data**

The mice data consists of NMR spectra, measured at  $p = 2050$  wavelengths, of  $n = 30$  mice with cancer. These data were kindly provided by Dr David Axelson (Kingston, Canada). There are two groups present in the data. The first class represents 10 mice in which a tumor was implanted and who then received no treatment. This is the control group. The second group contains the mice that received chemotherapy or radiation therapy.

We again started by performing a PCA analysis on these high-dimensional data. In order to avoid the curse of dimensionality, we decided to retain  $k = 3$  components. The ratio (3) is then 85% for ROBPCA and 81% for CPCA, which is also an indication that we retain a lot of information by selecting three components. Note that for ROBPCA we used  $h = 22$ , allowing slightly more outliers than with the default value of  $h = 25$ .

The resulting outlier maps are displayed in Figure 6. With CPCA we only find mouse 3 as a good leverage point and it is even a border case. The PCA subspace estimated with ROBPCA clearly deviates from the CPCA subspace as the distances of several observations change significantly. The third observation now also has a large orthogonal distance and becomes a bad leverage point. The score distance of cases 4, 21 and 22 increases, making them all leverage points as well. To be sure that the robust approach is more reliable, we looked at the mutual effect of the observations 3, 4, 21 and 22 on the first loading vector  $P_1$ . Therefore, we computed the angle between the first loading vector based on the full dataset and based on the reduced dataset in which observations 3, 4, 21 and 22 were removed. For CPCA this angle is  $36.09^\circ$ , whereas for ROBPCA it is only  $4.87^\circ$ . These four samples thus have a large influence on the classical analysis, hence they can indeed be considered as leverage points. This is correctly observed when applying ROBPCA.

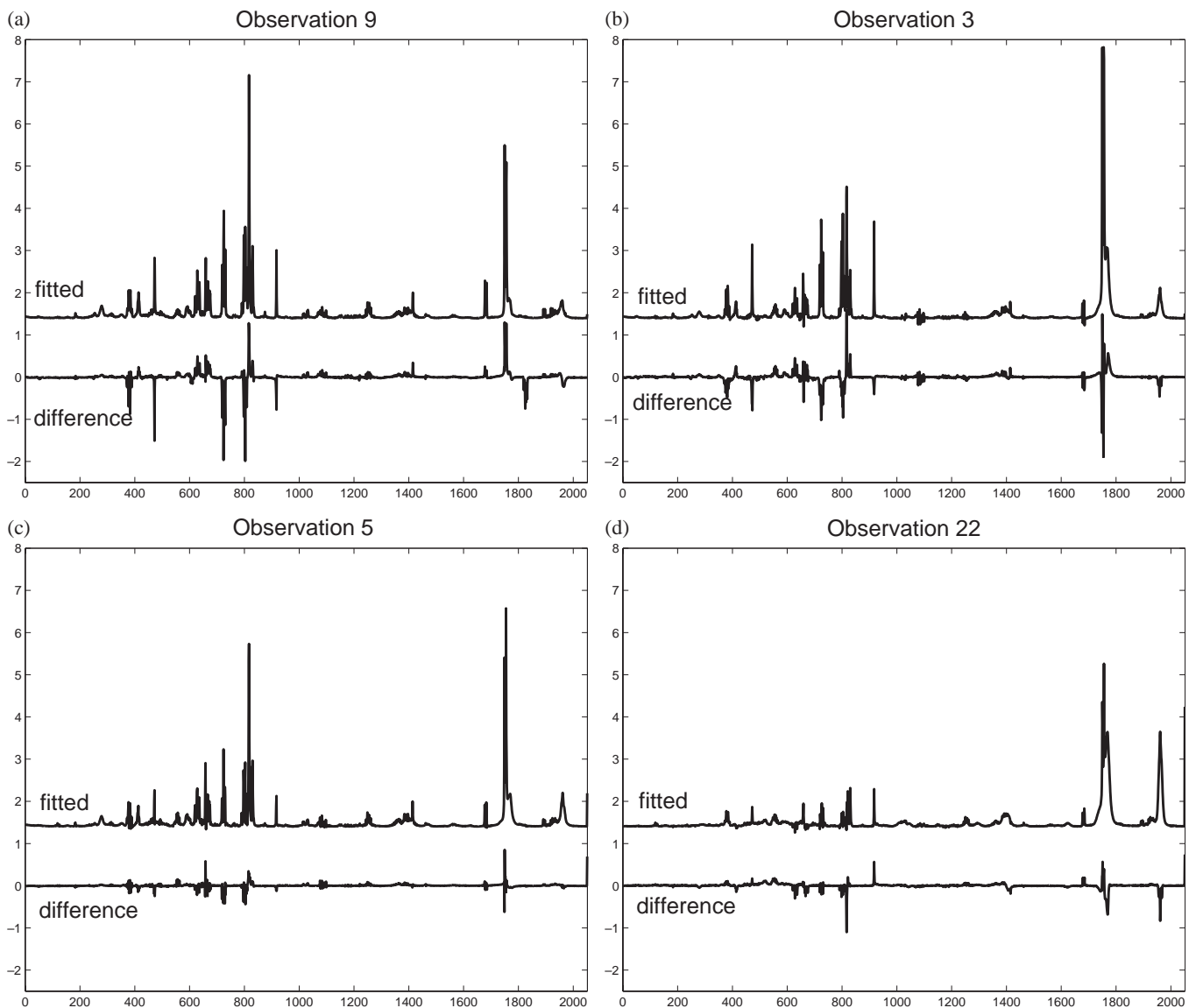
To study in more detail the classification of the outliers, we consider a regular mouse (sample 5), an orthogonal outlier (9), a good leverage point (22) and a bad leverage point (3). We calculated both the estimated spectrum of these four observations using Equation (2) as well as the difference between the estimated and the observed spectrum. Both are shown in Figure 7, where for clarity we have shifted the fitted curve with 1.5. For the orthogonal outlier 9 we notice a large difference between the fitted and the observed spectra, but the profile of the fitted spectrum is very similar to that of the regular mouse 5. For a good leverage point, the reverse effect can be recognized. The fitted spectrum is close to the observed one, but its profile is quite different from that of sample 5. The



**Fig. 6.** Outlier maps of the mice data obtained with (a) CPCA and (b) ROBPCA.

first high peak is missing, which probably indicates a very low concentration of choline. For the bad leverage point 3, we notice that the fitted spectrum does not correspond with that of a regular point. The larger second peak is probably due to a higher concentration of lactate. Moreover, the observed spectrum differs strongly from the estimated one.

Next, we performed a quadratic discriminant analysis on the three-dimensional scores. The goal of this analysis is to find out whether it is possible to classify a mouse, which does not belong to the actual dataset of 30 animals and whose NMR spectrum is measured, into one of the known groups. Remember that the first group of 10 mice has cancer but receives no treatment, whereas the second group (observations



**Fig. 7.** Estimated spectra and the differences between the estimated and the observed spectra for (a) an orthogonal outlier 9, (b) a bad leverage point 3, (c) a regular observation 5 and (d) a good leverage point 22 of the mice data.

11–30) represents the mice treated with chemotherapy. Some of these even got radiation therapy.

To evaluate a classification rule, we can count the number of misclassified observations. However, when outliers are badly classified we should not blame the method. So, in the sequel we will only evaluate the different rules by means of the set of regular observations. We consider as outliers the ones found by computing the MCD on the robust scores in each group. For the mice data, they are samples 3, 4, 21, 22, 23 and 29. The different classification rules that we consider are CQDA and RQDA applied to the CPCA and the ROBPCA scores.

First, we looked at the number of misclassified samples, summarized in Table 1. We see that preprocessing our data

with ROBPCA instead of CPCA yields less misclassifications. Note that the combination ROBPCA–CQDA misclassifies samples 2 and 16, whereas ROBPCA–RQDA assigns cases 2 and 19 to the wrong group.

This result is still too optimistic as it evaluates the classification rules on the training set. As we do not have a test set available, and since the dataset is rather small to be split into a training set and a test set, we next compute leave-one-out cross-validated misclassifications. For this, we remove a sample from the dataset, and then distract a discriminant rule on the reduced dataset. Next, we see whether the removed observation is well classified or not. This yields the misclassifications listed in Table 2.

**Table 1.** Number of misclassified observations from the mice dataset

		Group 1	Group 2	Total
CPCA	CQDA	2	4	6
	RQDA	2	3	5
ROBPCA	CQDA	1	1	2
	RQDA	1	1	2

**Table 2.** Leave-one-out misclassifications for the mice dataset

		Group 1	Group 2	Total
CPCA	CQDA	8	9	17
	RQDA	8	5	13
ROBPCA	CQDA	4	3	7
	RQDA	2	3	5

We see, e.g. that ROBPCA–CQDA wrongly assigns four observations from the control group (cases 2, 5, 8 and 10) into the treatment group, and three from the treatment (12, 14 and 16) in the control group. The overall error rate is thus  $7/24 = 0.29$ . Remember that we do not perform the cross-validation on the six outliers. ROBPCA–RQDA yields better results. Only two cases from the control group (2, 5) are not correctly classified, and three cases from the treatment group (11, 19 and 20), resulting in an overall error rate of  $5/24 = 0.21$ . These rates are very low in comparison with the classification results based on the classical scores. With CPCA–CQDA, an error rate of  $17/24 = 0.71$  is attained and  $13/24 = 0.54$  with CPCA–RQDA. For this dataset, we may thus conclude that combining ROBPCA with robust discriminant analysis yields the lowest estimates of misclassification.

Finally, we graphically checked the different methods by means of three-dimensional score plots. These figures confirmed the numerical results. The ellipsoids based on the classical covariance matrices for each group were inflated, and had much more overlap than those based on RQDA.

## ACKNOWLEDGEMENTS

We thank Patrick Glenisson (ESAT, KULeuven) for motivating us to study robustness on micro-array data, and David Axelson for providing us the mice data. The term ‘outlier map’

for the diagnostic plot is due to Salvador Garcia (McMaster University, Canada).

## REFERENCES

- Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci., USA*, **97**, 10101–10106.
- Box, G.E.P. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification. *Ann. Math. Statist.*, **25**, 33–51.
- Cui, H., He, X. and Ng, K.W. (2003) Asymptotic distributions of principal components based on robust dispersions. *Biometrika*, **90**, 953–966.
- Donoho, D.L. (1982) *Breakdown Properties of Multivariate Location Estimators*. PhD Qualifying paper, Harvard University.
- Hubert, M., Rousseeuw, P.J. and Vanden Branden, K. (2004) ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **45**, 301–320.
- Hubert, M., Rousseeuw, P.J. and Verboven, S. (2002) A fast robust method for principal components with applications to chemometrics. *Chemometrics Intell. Lab. Syst.*, **60**, 101–111.
- Hubert M. and Van Driessen, K. (2004) Fast and robust discriminant analysis. *Comput. Stat. Data Anal.*, **45**, 301–320.
- Johnson, R.A. and Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River.
- Jolliffe, I.T. (1986) *Principal Component Analysis*. Springer-Verlag, New York.
- Model, F., König, T., Piepenbrock, C. and Adorjan, P. (2002) Statistical process control for large scale microarray experiments. *Bioinformatics*, **1**, 1–9.
- Rousseeuw, P.J. (1984) Least median of squares regression. *J. Am. Stat. Assoc.*, **79**, 871–880.
- Rousseeuw, P.J. (1985) Multivariate estimation with high breakdown point. In Grossmann, W. et al. (ed.), *Mathematical Statistics and Applications*, Vol. B. Reidel, Dordrecht, pp. 283–297.
- Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*. Wiley, New York.
- Rousseeuw, P.J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
- Stahel, W.A. (1981) *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators*. Ph.D. thesis, ETH, Zürich.
- Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R. (1998) Large-scale temporal gene expression mapping of the central nervous system development. *Proc. Natl Acad. Sci., USA*, **95**, 334–339.
- Zuo, Y., Cui, H. and He, X. (2004) On the Stahel–Donoho estimator and depth-weighted means of multivariate data. *Ann. Stat.*, **32**, 189–218.